Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Veli Vural Uslu, B.Sc.
Born in Muğla, Turkey

Date of oral examination:
07/02/2014

Functions of the Distant Non-Coding Regions in Controlling *c-Myc* Expression

Referees: Dr. Anne Ephrussi

Prof.Dr. Jochen Wittbrodt

“Kahrolsun bağzı şeyler”*

Taksim Direniş, 2013

*”When you resist, you look gorgeous”. Taksim Resistance,2013.

Table of Contents

1. SUMMARY:

1.1 Summary in English

In mammalians, the protein-coding sequences, which make up less than two percent of the genome, are separated from each other with large non-coding intervals. Genomic rearrangements in the developmental gene loci indicate that the genes are regulated by long-range enhancers from these non-coding regions. Recently, Genome Wide Association Studies (GWAS) have indicated that forty percent of all human genomic variations, which are associated with a phenotype, are exclusively in the non-coding regions. This suggests that the long-range control of gene expression is a widespread phenomenon in the mammalian genome.

There are two main challenges to understand the long-range gene regulation. The first one is to discover the regulatory elements in the vast non-coding regions and to characterize their function. The second challenge is to identify the molecular mechanisms that enable and control the communication between the regulatory elements and their target genes. The genome may appear as a collection of elements but as shown by genomic rearrangements, its organization is important for the spatiotemporal regulation of gene expression. The processes, which convert the regulatory function of individual elements into collective spatiotemporal regulatory information, are poorly understood.

In this study, I used mouse *c-Myc/Pvt1* flanking locus as a model to understand the contribution of genome organization to endogenous gene expression. This locus is an evolutionarily conserved three megabase-long gene-poor region, with only one protein-coding gene: *c-Myc.* Retroviral insertions, chromosomal translocations and duplications, with breakpoints up to hundreds of kilobases far from *c-Myc* lead to various cancers both in mouse and in humans. Furthermore, GWAS showed that genetic variations in humans all along this locus are associated tissue and stage specific tumorigenic or developmental phenotypes. Moreover, a large-scale genome profile revealed by ENCODE project identified elements carrying signatures of enhancers in different cell types in this locus. These studies suggested that long-range regulatory activity is prominent in this locus.

In this project, I generated tens mouse lines with a regulatory sensor at different positions to monitor the regulatory activity in the *c-Myc/Pvt-1* flanking locus. I have revealed a long-range embryonic face enhancer, which overlaps with the linkage disequilibrium block in the orthologous human 8q24 locus associated with non-syndromic cleft lip and palate risk. In order to get insight into the biological role of this regulatory region, I have generated a series of genomic rearrangements and restricted the embryonic face specific regulatory elements to 250kb long interval. I have shown that this regulatory region acts on the *c-Myc* gene in a tissue specific manner over a megabase distance. Further analysis of *c-Myc* downregulation indicated deregulation of gene regulatory networks and metabolic pathways upon the deletion of face enhancer. These pathways may implicate the etiology of 8q24 dependent non-syndromic cleft lip and palate.

In addition, in collaboration with Andreas Trumpp's lab, we have investigated the effects of the deletions in the telomeric end of *c-Myc/Pvt1* flanking locus on the hematopoietic system. We have identified that the most telomeric 350kb long region in this locus is critical for different stages of hematopoiesis. We have shown that the regulatory region at this locus acts on *c-Myc* gene despite being more than 1.4 megabase far.

Finally, I investigated the elements that allow communication of distant regulatory regions with the promoter of *c-Myc.* I have shown that the regulatory landscape is confined in a Topologically Associated Domain (TAD) and the telomeric end of this TAD has dual functions for insulator and tethering activity.

1.2 Deutsche Zusammenfassung

In Säugetieren machen Sequenzen, die für Proteine kodieren weniger als 2% des Genoms aus und sind durch lange nicht-kodierende Bereiche getrennt. Genomische Umordnungen in entwicklungsgenetischen Loci deuten darauf hin, dass Gene von Enhancern mit großer Reichweite reguliert werden. Kürzlich wiesen Genom-weite Assoziationsstudien (GWAS) darauf hin, dass 40 Prozent aller humanen genomischen Variationen, die mit einem Phänotyp assoziiert sind, sich ausschließlich auf nicht-kodierende Bereiche beschränken. Das legt den Schluss nahe, dass Gen-Regulation über weite genomische Intervalle ein verbreitetes Phänomen in den Genomen von Säugetieren ist.

Um Gen-Regulation über weite Distanzen zu verstehen, existieren zwei Herausforderungen. Erstens müssen regulatorische Elemente in den ausgedehnten nicht-kodierenden Bereichen charakterisiert werden. Und zweitens müssen die molekularen Mechanismen, die die Kommunikation zwischen Enhancern und ihren Zielgenen ermöglichen, identifiziert werden. Obwohl das Genom auch als Ansammlung von Elementen angesehen werden kann, ist die Anordnung dieser doch von großer Bedeutung für räumliche und zeitliche Gen-Regulation. Die Prozesse, die die regulatorischen Funktionen von einzelnen Elementen in gemeinsame räumliche und zeitliche Expressionsmuster integrieren sind, wenn überhaupt, nur spärlich verstanden.

In dieser Arbeit nutzte ich den *c-Myc/Pvt1* flankierenden Locus als Model, um den Beitrag der Genom-Organisation auf Gen-Expression zu untersuchen. Dieser Locus ist ein evolutionär konservierter 3 Megabasen langer, Gen armer Intervall mit nur einem für ein Protein kodierenden Gen: *c-Myc*. Retrovirale Integrationen, chromosomale Translocationen und Duplikationen mit Bruchstellen bis zu mehreren hundert Kilobasen entfernt von c-Myc führen in Maus und Mensch zu unterschiedlichen Krebsformen. Darüber hinaus zeigten GWAS, dass genetische Variationen in diesem Locus mit Gewebe- und Entwicklungsstadiums-spezifischen kanzerogenen und entwicklungsbiologischen Phänotypen korrelieren. Zusätzlich wurden im großangelegten ENCODE Projekt in diesem Locus Elemente identifiziert, die in mehreren Zelltypen die Signaturen von Enhancern tragen. Diese Studien

deuten somit darauf hin, dass in diesem Locus regulatorische Aktivität mit großer Reichweite vorhanden ist.

In diesem Projekt erstellte ich zehn Mauslinien her, die einen regulatorischen Sensor an unterschiedlichen Positionen im *c-Myc/Pvt1* Locus tragen. Ich identifizierte so einen in der Gesichtsentwicklung involvierten Enhancer mit großer Reichweite, der mit dem Linkage Disequilibrium Block auf dem orthologen humanen 8q24 Locus überlappt, der mit nicht-syndromischen Lippen-Kiefer-Gaumenspalten Risiko assoziiert ist. Um einen Einblick in die biologische Funktion dieser regulatorischen Region zu erhalten, generierte ich eine Reihe von genomischen Umordnungen und konnte so die Region des embryonalen Gesichts-Enhancers auf ein 250kb langes Intervall eingrenzen. Ich konnte zeigen, dass diese regulatorische Region gewebsspezifisch *c-Myc* Expression über einen Megabasen Intervall hinweg beeinflusst. Weitere Analyse der c-Myc Herunterregulierung deutete auf eine Misregulation von Gen-regulatorischen Netzwerken und Stoffwechselwegen hin. Diese Signalwege konnten die Krankheitsursache des 8q24-abhängigen nicht-syndromischen Lippen-Kiefer-Gaumenspalten Syndroms sein.
Darüber hinaus erforschte ich, in Kollaboration mit dem Labor von Andreas Trumpp, den Effekt von Deletionen am telomeren Ende des *c-Myc/Pvt1* flankierenden Locus auf das hematopoietische System. Wir identifizierten das telomere 350kb-Intervall als kritisch für unterschiedliche Stadien der Blutbildung. Des Weiteren konnten wir zeigen, dass diese regulatorische Region c-Myc über einen Distanz von mehr als 1.4 Megabasen hinweg beeinflusst.

Abschließend, analysierte ich die Elemente, die die Kommunikation von entfernten regulatorischen Regionen mit dem Promoter von c-Myc ermöglichen. Ich zeigte, dass die regulatorische Landschaft von Topologisch Assoziierten Domänen (TADs) begrenzt wird und dass das telomere Ende dieses TADs eine duale Funktion als Isolator und Anbindungselement besitzt.

2. INTRODUCTION:

2.1. Gene Regulation by Long-Range Enhancers and Genome Organization

2.1.1. Gene Expression

Continuity of life is simply based on successful transmission of genetic material through generations. To ensure transmission of genetic material to their following progeny organisms carry out many biochemical reactions with the help of the resources available in the environment and the proteins that are encoded in its DNA. Due to the instable nature of the environmental conditions genetic data has to be modular and adaptable for organismal growth and division or reproduction. Francois Jacob, Jacques Monod and their colleagues revealed fundamental mechanisms, in which the metabolic genes are transcriptionally regulated via the nutrients and metabolites available to the bacteria. For example, the availability of lactose induces lactose-metabolizing enzymes (lactose and ß-galactosidase and ß-galactoside transacetylase) by inactivating the transcriptional repressor of these metabolic genes (*lacI*) (Jacob F and Monod J, 1961). This well-known concept of the lac operon postulated that the response to environmental stimuli is not only due to the modularity of a single protein ("enzyme adaptation") but also due to the change in the expression level of a set of proteins in *E.coli (*"enzyme induction") *(*Monod J, 1966)

Mechanisms similar to the prokaryotic gene expression were discovered in simple eukaryotic systems like yeast. One of the most well known mechanisms is the Gal4-UAS system in yeast, in which the galactose uptake and the metabolism genes are regulated according to the availability of the galactose itself (Traven A, 2006). In addition, temperature fluctuations lead to cellular and organismal response in fruit flies at the transcriptional level by the heat shock proteins (Ashburner M, 1970). It's been shown that changes in the osmotic pressure of the environment modulate transcription of the ion transport and the osmolyte regulation related genes (Wu MH, 2004). Besides, eukaryotic cells adapt to the oxygen levels by altering expression of metabolic genes (Chi J, 2006). Apart from the chemical (eg. oxygen, nutrients) and physical (eg. temperature, osmolality) state of the environment, pathogens also induce cellular response in the transcriptional level of immune system related genes in eukaryotes (Fujihara M, 1994). These findings suggest that altering gene expression contributes to the robustness of an organism in unpredictable environmental conditions.

Single cell eukaryotes (and prokaryotes) adapt themselves to extreme environmental conditions. Nevertheless, colonization, subsequent specialization, and cooperation among single cells drove the evolution of multicellular organisms by providing them more adaptive power than any of its components. Specialization of the cells in a multi-cellular organism was driven by usage of the same genetic material in a different way among the different compartments of the organism to establish cooperation between these compartments (Alberts B *et al* 1994). Unlike single cells, the genetic material in multicellular organisms does not only encode information for pure environmental stimuli but also information for cooperation and specialization within an organism. Therefore, different tissues have different gene expression programs and the differentiation is mediated by inducing gene expression changes. In brief, gene expression is one of the most fundamental biological phenomena that contributes to the flexibility and adaptability of a cell to environmental conditions along with differentiation of a cell and development of an organism.

2.1.2 Transcriptional Machinery and Principles of Eukaryotic Transcription

Gene expression is subject to regulation both at transcriptional and post-transcriptional stage. For example, siRNA pathway, 5'capping, and 3'polyadenylation control gene expression at post-transcriptional level in various eukaryotes (Hamilton AJ *et al*, 1999; Elbashir SM *et al*, 2004; Fire A *et al* 1991; Gu M *et al* 2001). In the introduction I will focus on transcription itself, which is a multi-step process and it is regulated at pre-initiation, initiation, promoter clearance, elongation and termination processes.

2.1.2.1 Transcriptional Machinery

Molecular players vary for transcription of different RNA types: For example, the main driver of tRNA transcription is RNA polymerase III, whereas RNA polymerase II (RNAP II) produces mRNA (Vannini A *et al* 2012). RNAPII synthesizes mRNA but cannot bind to the promoter by itself. Therefore, general transcription factors (GTFs) like TFIID, TFIIB, and TFIIF are required for the initiation of transcription. RNAP II, GTFs and promoter DNA sequence form the pre-initiation complex (PIC), which is sufficient for basal transcription both *in vivo* and *in vitro*. A universal cofactor complex called Mediator binds to the PIC *in vivo* to regulate transcription particularly via its interaction with the critical C-terminal domain (CTD) of RNAP II. This huge protein complex at the promoter is termed as transcription machinery (reviewed in (Woychik N, 2002); Myers LC *et* al 2000). Although, the transcription machinery is highly conserved in terms of protein structure from yeast to human, GTFs and the mediator composition may change from one promoter to another one in an organism for differential regulation of gene expression (reviewed in (Woychik N, 2002)).

The first step of RNAP II transcription is binding of GTFs to the site(s) close to transcription initiation (Hahn S, 2004). PIC is recruited to the core promoter sequence by binding of a very common GTF, called TBP (Goodrich JA *et al*, 1994). Preformed PIC sub-complexes, components containing TFIIB, and unphosphorylated RNAP II establish a closed complex at the promoter site upon TBP binding in a stepwise manner (Ranish JA *et al,* 1999). ATP hydrolysis melts the double stranded helix of the DNA at the transcription initiation site and the DNA interacts with the RNAPII (Murakami K *et al* 2013). Subsequent serine-5 phosphorylation of RNAP II relaxes the PIC and upon further hydrolysis of NTPs, RNAPII progresses along the DNA to synthesize mRNA. Upon termination of transcription, RNAP II is recycled by dephosphorylation at serine residues by FCP1 phosphatase (Cho H *et al* 1999). These main principles by RNAP II transcription represent a common molecular mechanism for mRNA transcription in all eukarya (Orphanides G *et al* 1996*)*.

The protein components of transcriptional machinery were mostly identified initially in transcriptional model systems such as the lac operon, trp operon in prokaryotes, regulation by Gal4 and cAMP in eukaryotes (Bertrand K *et al* 1976; Roesler WJ *et al* 1989). These studies showed fundamental differences between eukaryotic and prokaryotic transcription (Struhl K, 1999; Lin YS and Green MR 1989). Yet, structural studies pointed out that the components of eukaryotic transcriptional machinery were preserved in terms of protein sequence and architecture from yeast to human. Besides, successful implementations of heterologous assays like the yeast Gal4-UAS system in fish and drosophila suggested that the basic function of these components was also conserved in whole eukarya (Brand AH and Perrimon N *et al* 1993; Scheer N and Campos-Ortega JA *et al* 1999).

2.1.2.2. Components of Gene Expression

The assembly of RNAPII complex is controlled by regulatory elements. The DNA sequences that regulate transcription are widely classified into promoters, enhancers, silencer, insulators, and locus control elements (Maston GA *et al* 2006; Riethoven JJM 2010) (Figure 1).

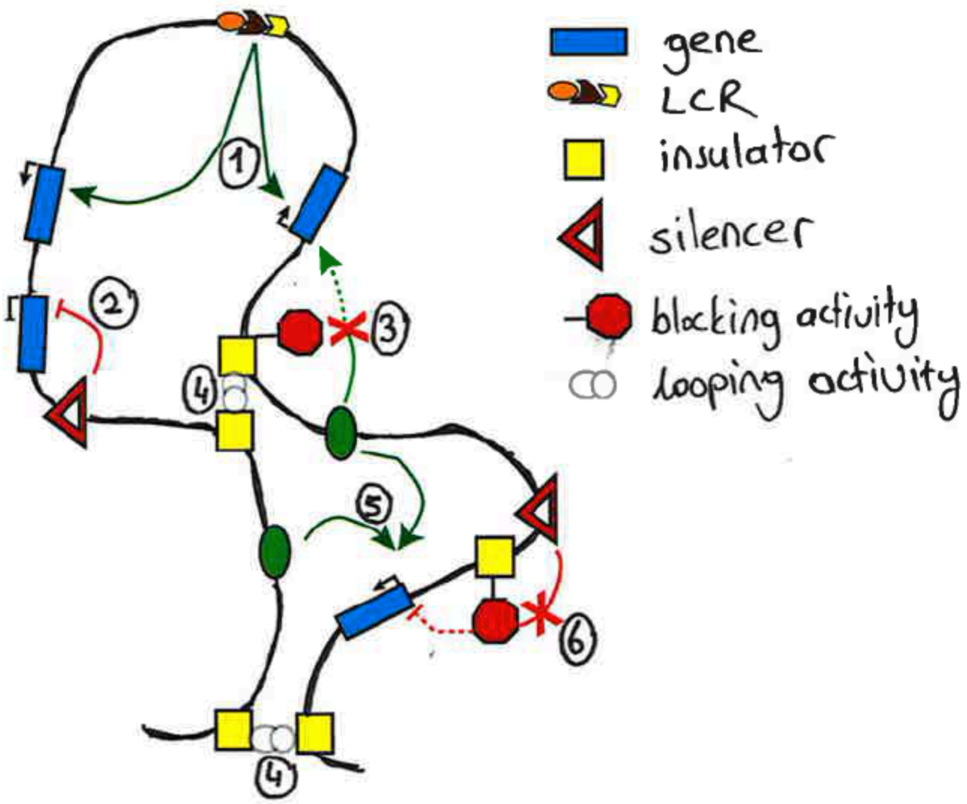


Figure1| Regulatory Elements in the Genome: 1) Locus Control Elements are composed of multiple coordinated elements and induces transcription of multiple genes in a locus. 2) Silencers repress the transcription of genes 3) Insulators block enhancer activity when they are in between enhancers and target promoter. 4) Insulators interact with each other and contribute to the architecture of a locus. 5) Enhancers induce transcription of target genes from long-distance. 6) Insulators also block progression of silencing activity and maintain the transcriptional state of a gene. Black arrows indicate the transcriptional activity in the promoters.

2.1.2.2.1 Promoters

Promoter sequences are located in the close proximity of transcription start sites (TSS) and mediate the initiation of transcription (Figure 1, black arrows). Promoters are composed of core promoter elements, which directly interact with PIC and proximal promoter elements, which possess binding sites for the activator proteins (Smale ST and Kadonaga JT, 2003). Apart from recruiting the PIC, the core promoter elements are proposed to determine the position of the TSS and the direction of the transcription. Several motifs have been identified in the core promoter region such as TATA box, Initiator (Inr), Downstream Promoter Element (DPE), Downstream Core Element (DCE), TFIIB recognition Element (BRE), Motif Ten Element (MTE) (reviewin in (Maston GA *et al* 2006)). In addition to these experimentally studied core promoter motifs several other motifs like motif8 and YY1 are identified via bioinformatic analysis (Xi H *et al* 2007; Lee MP *et al* 2005). Promoter motifs are widely shared among eukaryotes but with diversity. Approximately 77% of human promoters possess at least one of the given core promoter motifs. However, none of the motifs are present in more than half of the promoters. For example, Inr is shown to be the most abundant motif as it is found in 43% of the promoters, whereas TATA box is limited to 13% of the promoters (Gershenzon NI and Ioshikles IP 2005). In addition, CpG islands appear in the promoters more prominently than the given motifs and their methylation level reversely correlates with the activity of promoters (Ioshikhes IP and Zhang MQ,2000).

Even though there is no consensus on the precise position PIC binding, core promoter motifs have mostly been shown to locate 35bp upstream or downstream of TSS (Smale ST and Kadonaga JT, 2003). However, early studies by Steven McKnight and Robert Kingsbury demonstrated that elements between 110bp and 90bp far from TSS could increase the efficiency of transcription (McKnight SL and Kingsbury R, 1982). Such regions immediately upstream of core promoter elements were named "proximal promoter elements". These proximal regions contain common sequences for TFs such as NF-Y and CTF/NF-1 binding CCAAT box, which is present in 25-30% of all eukaryotic promoters (Jones KA *et al* 1985, Dorn A *et* al

1987, Bucher P, 1990) and SP1 binding GC boxes (Gidoni D *et al* 1984). These motifs in the proximal part of the promoter are shown to be activator-binding sites to augment basal transcription (Mantovani R, 1998).

Depending on the motifs in the core or proximal sites, promoters exhibit functional differences. For example, Promoters with a TATA box around -20bp and Inr motif are more likely to start transcription at a single nucleotide. This "focused" initiation takes place mostly in regulated genes in simple organisms rather than vertebrates. In contrast, the presence of CpG islands and BRE sequence in the promoters correlate with multiple transcription start sites within a 50-100bp window. The majority of vertebrate promoters and the promoters with constitutive activity have this kind of "dispersed" initiation. Besides, promoters with multiple weak and one strong start site are identified and they display both focused and dispersed nature (Juven-Gershon T and Kadonaga JT, 2010; Sandelin A *et al* 2007). In addition to the differences in TSS, some mammalian promoters are shown to start transcription in both directions. Common motifs like TATA box are underrepresented in these bidirectional promoters, whereas, a number of activator binding sites like E-boxes and partially palindromic sequences frequently appear in their sequence (Lin JM *et al* 2007). Moreover, depending on the transcriptional activators, promoters may work in a tissue invariant manner (in housekeeping genes with CpG-dependent promoters), or tissue specific manner (in YYI dependent B29 promoter) (Johnson P and Friedmann T, 1990; Hatch N and Sarid J, 1994). Apart from its role in transcriptional initiation, some promoters are shown to stabilize the transcripts (Kadonaga JT, 2012).

Promoters cannot account for all aspects of gene expression in mammalian genome. For example, extensive studies on the characterization of *c-Myc* promoter showed that the promoter structure hosts many different modules, transcription factor binding sites and epigenetic modifications. However, this promoter structure did not reveal the regulatory logic underlying the complex regulation of *c-Myc* in the course of development and tissue homeostasis (reviewed in (Wierstra I and Alves J, 2008). (Further information on ANNEX_Promoters)

2.1.2.2.2 Enhancers

Enhancers activate transcription of their target genes. They reside outside of the promoter regions and most commonly they are found in intronic and intergenic sequences. They are bound by transcription factors and they act on their target promoters from long-distance (Figure 1). Enhancers are associated with the presence of CBP/p300, H3K4me1 and H3K27ac, and the absence of H3K4me3 marks (Chen C *et al* 2012, Heintzman ND *et al* 2007). Enhancers show two distinct characteristics for transcription factor binding dynamics: Orchestrated Binding and Modular Binding:

Enhanceosomes are very well described examples of orchestrated binding of transcription factors to the enhancer elements. Enhanceosome activity exclusively depends on coordinated binding of its subunits to the enhancer sequence in a precise order (reviewed in (Struhl K, 2001)). The cooperative binding of enhanceosome subunits like ATF-2 and IRF-3 to the IFN-ß regulatory sequence is modulated by their DNA binding domains rather than their protein-protein interactions (Falvo JV *et al* 2000). Once the whole enhancer sequence is bound by transcription factors, co-activator protein CBP/p300 is recruited to enhancers to step up transcription. Therefore, a point mutation in the TF binding site of this enhancer sequence can completely halt the assembly of the whole enhanceosome (Thanos D and Maniatis T, 1995) and impede transcription.

Enhancers with modular binding characteristics are observed prominently in the genome from flies to mouse. In examples like Drosophila troponin T gene enhancers (Mas J *et al* 2004) or synthetic enhancers (Kulkarni MM and Arnosti DN, 2003), the binding dynamics of activator proteins correlates with intermediate level of target gene expression. For immune system related genes, enhanceosomes are critical for rapid and efficient response. However, intermediate activation is important for genes like *c-Myc*, as its level of expression is slightly modulated in a tissue and stage specific manner (Further information in ANNEX_Enhancers).

2.1.2.2.3 Silencers

In the broadest terms, silencers are the DNA sequences, which involve in downregulation of target genes (Figure 1). These sequences can act from long-distance as well as short distance. Therefore, they are shown to locate in many distinct regions such as introns, exons of genes, and upstream of promoters. Silencers have been reported to interfere transcription initiation, elongation, RNAPII pausing, and splicing (reviewed in (Ogbourne S and Antalis TM, 1998)).

Very few silencer sequences are characterized, as it is technically hard to characterize them. One of the biggest challenge to reveal the silencers is the fact that eukaryotic genome is repressive by default. Therefore, silencers are discovered by overexpression of neighboring promoter upon the deletion/mutation of the silencer sequence or the downregulation of a reporter gene in the presence of a silencer.
The mode of action of silencers is very diverse and ambiguous (Examples are given in Annex_Silencers).

Repressor proteins, which bind to silencers, are crucial to understand the silencing mechanisms. Studies to investigate the gene silencing mechanisms and the early studies on developmental genes in drosophila converged by the discovery of a very important repressor: Polycomb Group Proteins (Pc-G). Nobel Prize winning research by Edward Lewis genetically identified Polycomb locus as a repressor of developmental bithorax gene complex in drosophila (Lewis EB, 1978). Pc-G proteins join the structure of larger protein complexes called Polycomb Repressor Complex (PRC). These complexes bind to silencer Polycomb Response Elements (PREs). Depending on the subunits, there are two major PRCs: PRC1 and PRC2. PRC1 has two ubiquitin E3 ligases in the structure, which are responsible for compacting the chromatin via ubiquitylation of H2A Lysine 119 (reviewed in (Schuettengruber B and Cavalli G, 2009)). On the other hand, PRC2 engages in repressive activity via its methylase subunits, which acts on H3K27. H3K27me3 strongly correlates with repressive activity and co-localizes extensively with PRC2 (reviewed in (Margueron R and Reinberg D, 2011)). H3K27 methylation by PRC2 is shown to be mediated lincRNAs like HOTAIR or XIST (Tsai M *et al* 2010' Kaneko *et al* 2010) In addition, it's also been postulated that PRC2 silences transcriptional elongation by demethylating H3K36me3 via NO66 enzyme (Brien GL *et al* 2012). However, the PRC contribution to transcriptional silencing is reported not to be solely mediated their histone modifying activity to illustrate that there may be additional mechanisms of silencing to be elucidated. (Eskeland R *et al,* 2010). Moreover, the silencing effect of PRCs is neutralized by active non-coding transcription or trithorax-group (trxG) binding (Schmitt S *et al* 2005; Ringrose L and Paro R, 2007). Therefore, the interplay between PRC dependent silencing and anti-silencing contributed to various spatiotemporal expression patterns of many developmental genes including *Hox* genes.

2.1.2.2.4 Insulators

Insulator elements are initially defined with their blocking activity against activation and repression of a gene (Figure 1). Namely, they either interfere with gene activation via blocking enhancer-promoter interaction or they obstruct propagation of repressive of chromatin. Insulator elements are found from yeast to human. Most of the insulators found in S.cerevisiae work as a barrier against spreading of silence domains. For example, HMR-tRNA domain can interrupt spreading of heterochromain when it is between the heterochromatin and the target gene both in its genomic context and in the heterologous assays (Donze D *et al* 1999). In contrast, most of the insulators in higher eukaryotes have enhancer-blocking nature (West AG *et al* 2002). A retroviral sequence called *gypsy* is one of the most studied example of enhancer-blocking insulator. Gypsy retrotransposon disturbs the communication between yellow gene and its regulatory element in drosophila (Geyer P *et al* 1986). Nevertheless, the mode of action of insulators is yet solely speculative.

Enhancer blocking activity of insulators is investigated thoroughly. For example binding of Su(Hw) zinc finger protein to 340bp LTR sequence of gypsy retrotransposon is essential to prevent enhancer-promoter interaction both in endogenous yellow gene locus and also in heterologous assays. In addition, mod(mdg4) protein is also critical for insulator function of gypsy LTR through its direct interaction with Su(Hw) protein (Ghosh D *et al* 2001). However, gypsy does not only disrupt enhancer-promoter interaction but it has been also reported that in the presence of both Su(Hw) and mod(mdg4), gypsy insulator blocks propagation of repressive chromatin (Roseman RR *et al* 1995). Additionally, in the absence of insulator protein mod(mdg4), gypsy insulator works an enhancer for *yellow* gene but a repressor for *cut* gene (Cai HN and Levine M, 1997). Intriguingly, the copy number of gypsy insulator determines its mode of action. If there is a pair of gypsy insulator between an enhancer and promoter, the enhancer-promoter interaction is no longer disrupted. However, if the insulator pair is placed on two sides of the enhancer, the enhancer-promoter activity is more strongly blocked. This mechanism of gypsy insulator was shown for both for *zerknullt* and *eye* enhancer context (Cai HN and Shen P, 2001; Muravyova E *et al* 2001). The insulator sequence is proposed alter high order chromatin structure via formation of loops in a rosette-like structure, which excludes the components within the loop from the components outside the loop (Gerasimova TI *et al* 2000).

In vertebrates, insulator elements are found in many distinct domains like the X-chromosome inactivation related *Tsix* gene, differentiation related ß-globin locus, imprinting related *Igf2/H19* locus, and c-Myc proto-oncogene promoter (Kim TH *et al* 2007). 5'HS4 in chicken ß-globin locus is one of the most first identified insulator in vertebrates. Extensive studies on 5'HS4 chicken insulator demonstrated that it disrupts enhancer-promoter interaction and also impede heterochromatin spread (position-effect protection). Very interestingly, 5'HS4 chicken insulator is shown to be functional in human cell lines as well as drosophila, pointing out the deep evolutionary conservation in the insulation mechanisms (Chung JH *et al* 1993). Further characterization of molecular players that regulate insulator activity revealed that binding of an evolutionarily conserved 11 zinc finger nuclear protein called CTCF to 5'HS4 sequence is necessary for enhancer blocking activity (Yusufzai TM and Felsenfeld G, 2004). However, CTCF is shown to be neither necessary nor sufficient for prevention of heterochromatin spread (Recillas-Targa F *et al* 2002). Barrier activity of 5'HS4 correlates with the histone acetyltransferase protein recruitment to this site (Litt MD *et al* 2001). In addition to CTCF, *Rad21* subunit of mitosis related cohesin complex is shown to localize in vast majority of CTCF sites in the interphase of pre-B cells and downregulation of *Rad21* correlates with deregulation of enhancer-promoter interactions in the *Tcra* locus in mouse thymocytes (Parelho V *et al* 2008; Seitan VC *et al* 2011). Despite not sharing the exact molecular players, the mechanisms and insulator functions are conserved in vertebrates and ecdysozoans. (Further information in Annex_Insulators)

2.1.2.2.5 Locus Control Region

Locus Control Region (LCR) is a multi-component cooperative regulatory element that controls the expression of gene cluster in a tissue specific, position independent and copy-number dependent manner (Grosveld F *et al* 1987) (Figure1). Despite the most studied LCR is in mouse ß-globin locus, other LCRs in the vertebrates such as in human TCRa, CD2, APOE/C-1 locus are also studied extensively. ß-globin LCR is also found in different vertebrates like chicken, goat rabbit, and human (reviewed in (Li Q *et al* 2002).

Operational dissection of ß-globin LCR subunits revealed puzzling findings. The deletion of LCR drastically drops the expression of globin genes in erythroid cells in mouse. 4.5kb fragment of the ß-globin LCR and was enough to promote transcription from 1.5kb long ß-globin promoter in erythroleukemia cell lines but not consistently in the erythroid cells in transgenic mice. ß-globin LCR is only fully functional in transgenic mice when 14kb region, which contains 5 DNase1 Hypersensitivity Site (HS), is cloned upstream of a ß-globin locus gene (Grosveld F *et al* 1987). This LCR is able enhance the activity only in the tissues where promoters are active. For example, transgenic mice with lacZ reporter under the control of Hsp68 promoter show expression in the yolk sac (Kothary R *et al* 1987). When there is LCR in the upstream of Hsp68 and lacZ reporter gene, the lacZ expression does not fully recapitulate LCR driven expression in ß-globin locus, but strengthens the reporter expression in the tissues like yolk sac, in which Hsp68 is already shown to be active (Tewari R *et al* 1996). This suggests that the promoter identity contributes to tissue specificity of LCRs. Of these 5HS in LCR, HS2 and HS3 can act as an enhancer via NF-E2 and GATA1 activators (Talbot D and Grosveld F, 1991). Despite keeping the histone acetylation at the same level, binding of NF-E2 increases the expression of ß-globin gene more than 100 fold (Sawado T *et al* 2001). Deletion of core HS2 and HS3 element halts the spatiotemporal ß-globin gene expression (Peterson KR *et al* 1996; Navas PA *et al* 1998). Whereas, the deletion of whole HS2 does not affect the spatiotemporal regulation of globin genes despite slightly lower level of expression. This suggests that the LCR elements outside of HS2 core sequence can regulate globin expression but HS2 core sequence deletion is dominant negative (Bungert J *et al* 1999). By using proximity based ligation, Wouter de Laat lab showed that differential looping via CTCF (or the chromatin conformation) of ß-globin locus correlates with the transcriptional activity of globin genes (Tolhuis B *et al* 2002; Splinter E *et al* 2006). To sum up, so far, the studies demonstrated that LCR is a regulatory unit based on cooperation and synergy among the elements within the unit so that the function of the whole LCR is not sum of the functions of its elements. (Further Information in Annex_LCR)

2.1.3 Functional Annotation of the Genome

Evolutionary conservation and biochemical features like nucleotide modifications, nucleosome occupancy, histone modifications, and transcription factor binding are used to functionally annotate the genome.

2.1.3.1 Features Associated with Regulatory Activity

Extensive analysis of histone modifications like methylation, acetylation, and ubiquitylation revealed distinct features of chromatin. The combinations of histone marks are used to partition the genome. These small bins of DNA are correlated with previously identified regulatory elements. Very particular combinations appeared in IL2RA and IFNG enhances, which are specific to CD4+ T cells (Wang Z *et al* 2008). With an inductive reasoning, more enhancers can be revealed via ChromHMM-like Hidden Markov Model based algorithms (Ernst J and Kellis M, 2012). For example, different combinations of histone methylation patterns are shown to be predictive for active and inactive domains of the chromatin (Barski A *et al* 2007). Together with histone modifications, DNA nucleotide modifications like CpG methylation or hydroxymethylation are also classified in epigenetic modifications (reviewed in Branco MR *et al* 2012). Bisulfite conversion, or antibody mediated imunoprecipitation of modified nucleotides contributed to the partitioning of the genome and reflected the transcriptional activity in a given locus more precisely when analyzed together with histone modifications (reviewed in Zhou VW *et al* 2011). For example, H3K4me3 signal in a hypomethylated high CpG content promoter are enriched for RNAPII (Guenther MG *et al* 2007). Nevertheless, vast amount of these promoters are negative for H3K36 methylation and perform extremely low-level transcription. This indicates that the predictive power of ChIP data is yet far from being optimal, possibly due to currently unknown parameters.

In addition to histone modifications obtained in ChIP experiments, transcription factor binding is also a major source of information that reflects the regulatory activity in genome-wide scale. For example, in drosophila some activator proteins like Tinman were enriched in known heart specific enhancer sequences and repressors like Hairy were enriched in silenced ftz regulatory region (Jin H *et al* 2013, Li LM and Arnosti DN, 2011). In mammalian systems ChIP experiments pointed out that the activator protein p300 was found in many well-characterized cell-type/tissue specific enhancers including IFN-ß enhancer. Interestingly, the occupancy of the enhancer at least in IFN-ß locus correlated with the transcriptional activity of the gene (Merika M *et al* 1998). In addition to these active marks, Ring1B subunit of PCR1 together with H3K27me3 and H2AK119u1 occupies Hox gene cluster and their binding correlate with the repressed state of Hox genes in embryonic stem cells (ESC) (Endoh M *et al* 2012). Similarly HP1 binding together with H3K9me3 and H3K27me3 labels transcriptionally inactive heterochromatic regions.

Recently, a number of independent experiments demonstrated an unusual class of biochemical marks called bivalent marks, which does not reflect the activity of the locus at the current state but also informs about its regulation in the later stages of differentiation. Bivalent marks generally appear in the promoters but also in the distant regions and they carry both active and repressive marks. For example, the TSS of inactive Irx2 gene carries both active marks like H3K4me3 and inactive marks like H3K27me3 in ESC. However, in the differentiated cells where Irx2 is active such as mouse lung fibroblast, the H3K27me3 sign is not maintained and Irx2 is actively transcribed (Bernstein BE *et al* 2006). It has been shown that in fibroblasts that the enhancer occupancy by transcription factors like Oct4 can induce a bivalent state for the promoters repressed by Polycomb proteins and keeps its repressed state. However, binding of *MyoD1* transcription factor to the very same enhancer interferes with Polycomb binding and H3K27me3 state of the *MyoD1* promoter and lead to appearance of H3K4me3 active mark in the promoter, which correlates with transcriptional activity of the *MyoD1* gene. Similar bivalent modifications were investigated in the mouse genome and revealed promoters of other developmental genes in bivalent state and their enhancers in permissive state (Taberlay PC *et al* 2011). Consequently, bivalent marks do not only correlate with the transcriptional activity of genes but also have implications on the transcriptional state of the genes in its course of differentiation.

Some of the histone marks do not only have correlational relation with the transcriptional activity but also give mechanistic information in the biochemistry of region of interest. For example, H3K36me3 mark is shown to be directly proportional to the transcript level (Mikkelsen TS *et al* 2007). Extensive studies on the transcriptional machinery demonstrated that Set2, which interacts with elongating RNAPII methylates H3 at Lysine 36 position (Li B *et al* 2007). Therefore, this histone mark is the consequence of transcriptional activity. On the other hand, H3K36me3 has consequences as well. In cancer genomes, it's been shown that mutation rates are higher in H3K9me3 associated heterochromatin when compared to transcriptionally active euchromatin (Schuster-Böckler B and Lehner B, 2012). It's been also shown that DNA mismatch recognition proteins are recruited to the actively transcribed regions via H3K36me3 (Li F *et al* 2013). This recruitment may contribute the low mutation rates in the euchromatin. Similarly, H3K36me3 involves in splicing machinery via indirectly interacting with proteins, which regulate splicing (Luco RF *et al* 2010). These examples show that the histone modifications may not necessarily involve in regulation of transcriptional level but in genome stability or alternative splicing. However, still these marks are useful to partition the genome and make predictions on regulatory nature of the region.

In addition to ChIP, DamID experiments identified Lamina Associated Domains (LADs). LADs showed that nuclear membrane is resided by large regions in the human genome, which were from hundreds of kb to several megabase long stretches (Guelen L *et al* 2008). These regions generally corresponded to the repressed regions in the genome and they are enriched by repressive H3K9me2 (Wen B *et al* 2009). Moreover, further studies that recruit active regions to inner nuclear membrane via tethering lac operator to lac repressor or tetracyclin responsive element to rtTA, which are fused to inner membrane protein, resulted in the repression of the given sequences (Reddy KL *et al* 2008; Kumaran RI and Spector DL, 2008). This implicates causality between the repression and the nuclear organization more than causality. This is in support of the hypothesis based on Hutchinson-Gilford progeria syndrome, in which a mutation in lamin gene leads to loss of nuclear membrane associated heterochomatin regions marked by H3K9me3 (Shumaker DK *et al* 2006). Interestingly, despite the presence of CTCF sites within a LAD, a clear enrichment for the CTCF occupancy at the LAD boundaries has been reported (Guelen L *et al* 2008).

2.1.3.2 Identification of Enhancers

Several different methods have been used to identify the regulatory information provided by the "potential" enhancer sequence. First of all, the classical enhancer activity tests rely on a simple reporter-gene assay. In this assay enhancer sequence is cloned into the plasmid carrying a promoter and a reporter-gene like luciferase or lacZ. Then the activity or reporter gene is measured in transiently or stably transfected cell lines. (For example: Sakurai M and Strominger JL, 1988) Since the enhancer activity is mostly tissue specific, the cell type used for enhancer testing is important to get a meaningful readout. This simple enhancer assay has substantial limitations. First of all, the cell lines used for the assay are at least transcriptionally very different from the tissues that they represent. Therefore, the activity of developmental enhancers activity cannot be captured by using solely this classical assay (Zheng-Bradley X *et al* 2010). Nevertheless for high-throughput enhancer screens like STARR-seq, classical enhancer assay is used and 11/13 of the enhancers found in S2 cells gave tissue specific patterns in transgenic flies (Arnold CD *et al* 2013)

In order to circumvent the tissue specificity problem in a developmental context, transgenic mice are used. The sequences tested for enhancer activity is cloned upstream of Hsp68-lacZ sequence. Linearized plasmid is used for pronuclear injection to fertilized oocytes (Pennacchio LA *et al* 2006). Alternatively lentivirus is used for the transgenic assay in a similar fashion (Friedli M *et al* 2010). The zygotes are implanted back to the foster mothers and the lacZ activity is measured in a relevant stage (Poulin F *et al* 2005). This method is the basis of VISTA Enhancer Database and it was successful enough to reveal many sequences that can autonomously drive gene expression in a tissue specific manner. However, in this assay the main problem is that the enhancer is immediately upstream of the promoter. At least for the developmental genes, the critical tissue specific enhancers are proposed to be up to a megabase far from the target genes. This assay reveals the isolated function of the regulatory sequence, which may not reflect its role in its endogenous locus, where it works cooperatively with other enhancers. For example, Mirna Marinic and colleagues revealed tissue specific autonomous enhancer activity of many sequences in Fgf8 locus. However, some of this enhancer activity was not used by any of the neighboring genes. Therefore, autonomous activity of the enhancers may not correspond to their activity in vivo (Marinic M *et al* 2013). However, for compact genomes like drosophila, where the regulatory sequences are not as dispersed as mammalian genome, this kind of transgenic assay is very informative for enhancer activity. In flies, the regulatory elements like the enhancers of eve gene is cloned upstream of Hsp70 (and also eve promoter) and the plasmid as introduced to the fly via P element mediated transformation. The reporter readout recapitulated the endogenous eve expression pattern very successfully in a promoter independent way (Goto T *et al* 1989).

In order to monitor the enhancer activity in its genomic surrounding, instead of injecting a single enhancer, minimal promoter construct, bacterial artificial chromosomes (BACs) are injected to fertilized oocytes. This assay reflects the enhancer activity of a sequence in its native position. BACs with Fgf8 regulatory sequences and different lacZ insertions clearly demonstrated that the regulatory input depends on where the reporter gene is inserted in the BAC (Marinic M *et al* 2013). The prostate enhancer in human chr8 and the brain enhancers in Shh locus have been identified via BAC reporters (Wasserman NF *et al* 2010; Jeong Y *et al* 2006). Although yeast artificial chromosomes (YACs) are much less stable than BACs, they are also used to identify enhancers (McBride DJ *et al* 2011, reviewed in Lamb BT and Gearhart JD, 1995)). BACs also allow deletion of rapid deletion of regulatory sequences and monitor the influence of the deletion on the reporter gene to further characterize an enhancer element.

Enhancers are functionally identified upon deletions in the genome. For example in drosophila, the enhancer sequence of *eve* gene was shown upon a deletion of the sequence (Stanojevoc D *et al* 1991). This approach is very low throughput approach to find out which sequences work as enhancers upon deletions. Mouse heart enhancers in Cdkn2a locus were revealed upon knocking out a 70kb large block from the genome (Visel A *et al* 2010). Apart from being extremely slow, the deletion of an enhancer is not informative when the enhancer activity is redundant. For example, deletions of the ultraconserved elements, which clearly show autonomous brain or neural tube activity, do not lead to any visible phenotype. In this case, it is not clear whether their enhancer activity is redundant or the autonomous activity of a sequence does not indicate its enhancer activity in endogenous context (Ahituv N *et al* 2007).

Identification and characterization of shadow enhancers are other challenges for the current enhancer assays. Shadow enhancers contribute to the robustness of spatiotemporal gene expression only in the presence of environmental perturbation. Shadow enhancers are described by using reporter-BAC assays and shown for drosophila *snail* gene (Perry MW *et al* 2010).

2.1.3.3 Evaluating the Regulatory Potential of the Annotated Regions

So far, I have described certain ChIP (or DamID) based histone modifications or transcription factor binding profiles partitions the genome. Moreover, microarray and RNAseq data gave also global view on transcriptional status of the genes (Kogenaru S *et al* 2012). Therefore, these wet-lab methods combined with extensive computational algorithms indicated a number of correlations between the biochemical marks or evolutionary conservation and the transcriptional state of the loci. Some of these correlations are shown to be causal or consequential correlations, yet, majority of them remain as correlations. One of the main question is how much functional data can we obtain from these biochemical marks or evolutionary conservation.

In order to test whether in silico evolutionary conservation analysis accurately predicts enhancer activity, 167 of 3100 sites, which are conserved between human and Fugu are tested for in vivo autonomous short-range activity. 45% of these sequences showed reproducible tissue specific activity (Pennacchio LA *et al* 2006). Pairwise genome alignments revealed that between rodents and humans there are 256 ultraconserved sites, which are longer than 200bp. Slightly less than half of these sequences show short-range tissue specific activity via transgenic reporter assay. Similarly, when highly conserved sequences, which lack ultra conservation between human and rodents, are tested, again half of these sequences show tissue specific activity at embryonic day 11.5 of mouse development (Visel A *et al* 2008). Lowering the degree of conservation does not change the prominence of spatiotemporal activity of the sequences. Moreover, evolutionary conservation criterion is shown to be more predictive for some tissues when compared to others. For example, predictions solely based on conservation give heart specific expression only in 2% of the trials, whereas, this number raise to 15% for brain specific expression due to the evolutionary divergence of the enhancers in the given tissues (Blow MJ *et al* 2010). More importantly, there are no studies, which reported the enhancer activity of solely non-conserved sites in mouse; it is not clear whether the evolutionary conservation criterion improves enhancer identification.

P300 binding sites are associated with enhancer activity. The sites enriched for p300 occupancy are investigated for their enhancer activity via transgenic reporter assay. Between 80% and 90% of these sequences showed regulatory activity in the tissue that p300 is bound (Visel A *et al* 2009). In a heterologous assay, human heart p300 binding sites are tested for enhancer activity in mouse embryos. 43/65 sequences could autonomously drive expression of reporter gene in short range in e11.5 mouse embryonic heart (May D *et al* 2011). On the other hand, enhancer predictions upon histone marks perform worse than p300. For example, only 6 out of 30 tested H3K4me1 sites obtained in ENCODE data showed tissue specific activity in the mouse embryos by the transgenic reporter assay. However, in medaka, predictions upon histone marks have predictive value of about 50% (Yip

KY *et al* 2012). In drosophila, DNA sequences associated with active marks like K3K79me, K3K27Ac and RNAPII are shown to have predictive value for tissue specific regulatory activity in 8 out of 9 tested sequences (Bonn S *et al* 2012). Biochemical activity is translated into regulatory information in short range. The mechanisms that distribute this regulatory information in the genome has not been understood yet.

2.1.4 Genome Organization

2.1.4.1 Composition of Eukaryotic Genome

Unlike the conservation of transcriptional machinery, genome structure of eukaryotes is highly divergent. Both the size of the genome and the composition of the genome dramatically differ among eukaryotes. The genome size itself does not tell much about the organism due to the 'c-value enigma'. Even before the discovery of DNA, it was reported that the complexity of an organism does not correlate with the total DNA content (the c-value) of the organism. (Gregory TR, 2005). The genome size correlates well with the gene number in the unicellular organisms, but not in the animal and the plant kingdoms (Lynch M, 2006). In some extreme examples of animal kingdom the genome size ranges from $2x10^7$bp (in nematodes) to $132x10^9$bp (in lungfish) (Gregory TR, 2005; Pedersen RA, 1971). For example, the genome sizes of in marine sponge, C.elegans, and humans are $1.6x10^9$, $9.7x10^7$, and $3.3x10^9$, respectively (C.elegans Genome Consortium *et al* 1998; International Human Genome Consortium *et al* 2004; Imsiecke G *et al* 1995; www.genomesize.com). On the other hand, the gene number is relatively constant for in animal kingdom. For the given three species, the gene numbers are 18000, 19000, 21000, respectively. Besides, the length of the protein-coding sequences is also constant around $1x10^7$ to $4x10^7$. Therefore, genome size turns out to be a reflection of the composition of non-coding elements in the genome. According to the given coding sequence length and the genome size, only 1.5% of the human genome is protein coding. The rest of the genome is composed of various non-coding sequences including transposable elements and introns (Figure2). The contribution of this non-coding region expansion to the reservoir of regulatory elements in the genome is still elusive.

Non-coding regions were proposed to have a structural role in the genome just by spacing two genes rather than having an active role (Ohno S, 1972). Dr. Susumu Ohno also brought up the idea that the non coding genome is 'junk DNA', which were functional coding sequences in the evolutionary past of the organism but not anymore. This explanation was widely accepted for a long time. Although several studies addressed the functions of non-coding regions by assessing the phenotypes upon deletion of various non-coding regions, they couldn't find an effect of these regions on survival or fertility in laboratory conditions (Ahituv N *et al* 2007). For example, mice with deletions of 1.5 megabase, and 0.8 megabase of non-coding intervals were indistinguishable from the wild-type littermates, suggesting that at least some portion of the non-coding genome can be discarded without visible consequences (Nobrega M *et al* 2004). On the other hand, a computational study points out that 10% of the mammalian genome is estimated to be subject to evolutionary selection (Smith N *et al* 2004). In addition, there are several studies pointing out the presence of regulatory elements in the non-coding regions, which have functions in development and disease (Kleinjan DA *et al* 2008). Therefore, although there is evidence for existence of both functional DNA and "junk DNA" in the non-coding regions, functionally investigated portion of the non-coding genome is negligibly small and it requires extensive characterization.

Nevertheless, one clear consequence of genome size expansion is the increase of average distance between two genes. In nematodes, since the gene density is 30 times higher than humans, it is more likely that the regulatory elements are more dispersed in humans when compared to nematode. This adds up another layer to the complexity of gene regulation by introducing a gene regulatory mechanism that acts from long-distance. The mechanisms, which contribute to the interaction between distant regulators and their target gene, are poorly understood. This long-range activity is particularly important to understand the molecular mechanism of developmental processes and diseases.

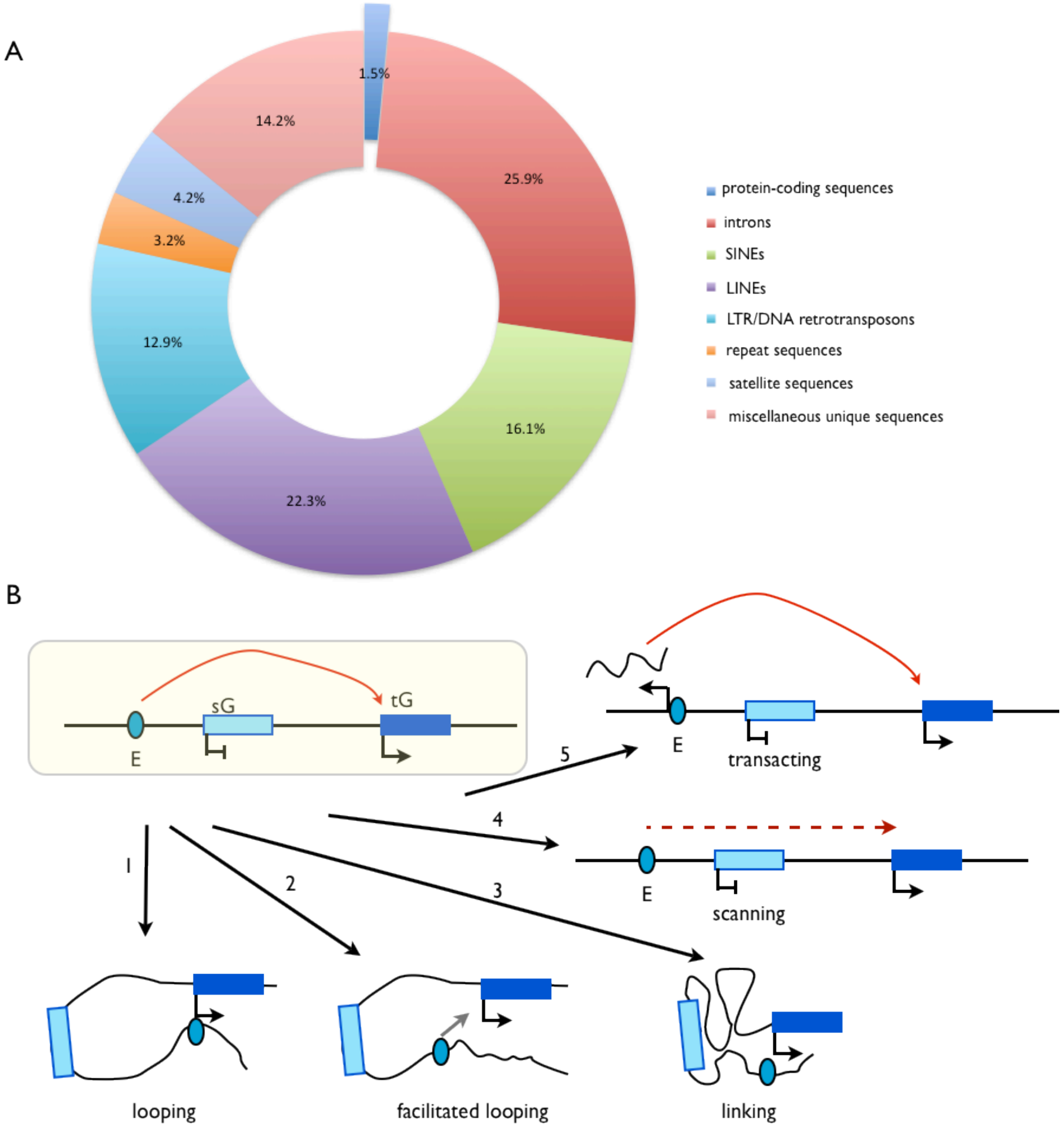


Figure2| Genome Organization: A) (Adapted From Patrushev LI *et al* 2008): The components of the human genome: A very small portion of the human genome encodes information for protein sequence. Non-coding sequences like transposable elements and introns constitute more than 75% of the whole genome. The Regulatory Element are in these large non-coding regions.

B) (Adapted from Dean A, 2006) The Models for Enhancer-Promoter Interactions: 1) A specific enhancer-promoter loop is shown in the looping model. 2) Enhancers are directed to promoters in the facilitated looping model. 3) The region between the enhancer and the promoter becomes compact and brings the enhancer to the promoter. 4) Enhancers scan along a locus to find their target(gene) according to the scanning model. 5) RNAs produced at enhancers work to enhance transcription of the target gene *in trans* according to transacting model. E: enhancer, sG: silent Gene, tG:target gene

2.1.4.2 Distant Nature of Enhancers

Human genetic rearrangements that give rise to developmental diseases indicated that developmental enhancers are often distant from their target genes. (Figure 1). For example, the deletion that causes small eye (Sey) phenotype is mapped around 100kb downstream of Pax6 gene (Fantes J *et al* 1995). Pax6 is known to be one of the most ancestral genes for the development of eye and haploinsuffiency of Pax6 gave rise to hypoplastic small eyes (Kozmik Z, 2005; Hill RE *et al* 1991, Ton CC *et al* 1991). Further studies showed that Pax6 eye expression is governed by multiple far downstream cis-regulatory sequences and deletion of these enhancers showed partial contribution of each enhancer to the Pax6 expression from long-range (Kleinjan DA *et al* 2001).

In addition to Pax6 locus, human diseases like Split Hand Foot Malformation (SHFM) indicated that certain regulatory sequences act on two different genes in the same tissue from a long distance. Dlx5 and Dlx6 transcription factors alone cannot give rise to SHFM but double knockout for these homeobox genes causes very severe phenotype including SHFM (Robledo FR *et al* 2002). Although the precise mechanisms are elusive, location of the chromosomal breakpoints in patients with SHFM suggests the presence of a megabase far distant enhancer, which can act both on Dlx5 and Dlx6. Futhermore, *Hox* gene clusters and *Myf5/Myf4* genes are controlled by shared regulatory elements from long distance (reviewed in (Spitz F and Duboule D *et al* 2008)).

Sonic Hedgehog (Shh) locus is another example of the development gene loci that has a long-range nature of regulation. Several different Shh coding sequence mutations and deletions are identified in a spectrum of Holoprosencephaly (HPE) patients (Nanni L *et al* 1999; Odent S *et al* 1999). A very similar HPE phenotype is also observed in patients who have chromosomal translocations in Shh locus with a breakpoint between 15-250kb downstream of Shh gene (Belloni E *et al* 1996). This suggests that there are distant regulatory elements, which are fundamental for proper Shh expression in the brain. In addition to the brain, Shh signaling is shown to be critical for limb development in mouse (reviewed in Nieuweinhuis E and Hui C, 2005). Some translocations and mutations almost a megabase far from Shh coding gene lead to a spectrum of limb related phenotypes (Lettice LA *et al* 2002; Sharpe J *et al* 1999). However, the nature of the mutations identified was also very intriguing. The deletions and translocation caused limb phenotype due to downregulation of Shh in the posterior limb bud, whereas, certain mutations lead to the ectopic expression in anterior limb bud and lead to preaxial polydactyly. In brief, the enhancers of developmental genes have often a long-range nature.

2.1.4.3 Communication between distant Enhancers and their Target Promoters

Enhancers, silencers, insulators, and the LCRs are the regulatory elements are dispersed in the vast non-coding regions of the genome. In the developmental loci these elements are shown to be away from their target genes (Marinic M *et al* 2013, Kimura-Yosida C *et al* 2004, Nobrega MA *et al* 2003). Moreover, in many of these loci, the target gene is not the closest one to the enhancer itself (Lettice LA *et al* 2002). Intriguingly, GWAS suggest that the long-range gene regulation is a very prominent phenomenon throughout the genome (Visal A et al 2009). This brings out one important aspect of gene regulation via distant enhancers: Specificity. How do enhancers communicate with the promoters and within each other? A number of different models have been proposed so far (Figure2B) In order to understand which model represents the nature of enhancer-promoter communication the physical structure of the genome and the interactions in the chromatin are essential to find out.

The molecular mechanisms of transcriptional activation via distant enhancers have been a long-standing discussion in the field (reviewed in (Atchison ML, 1988)). Activity difference between the promoters distal and proximal to the enhancer lead to the idea that enhancers function as an entry site to transcription factors, which slides along the DNA. Another model based on DNA accessibility around SV40 enhancer suggested that enhancers might influence changes in the chromatin structure or level of DNA supercoiling. A complete different model was proposed based on the correlation of transcription and nuclear positioning. Since enhancer also correlates with transcriptional activity, according to this "Nuclear Address Model", enhancers carry the promoters to transcriptionally active regions like matrix attachment regions (MARs). Last and still widely accepted looping model was initially proposed upon the observation that an enhancer sequence can activate its target promoter from a distance of 250-300bp only if they reside on the same side of the helix, which allows protein-protein interactions via looping (Dunn TM *et al* 1984, reviewed in Ptashne M, 1986) (Figure 2B).

The molecular mechanisms that steer the enhancers to their target promoters are still elusive. In drosophila antennapedia gene complex, regulatory elements selectively bring a distal enhancer to *Scr* promoter. This "tethering element" is a promoter independent element as it has been shown that when the tethering element in Scr promoter is cloned into another promoter such as *ftz* promoter, the enhancer specifically activates this new heterologous promoter (Calhoun VC *et al* 2002). Although, the tethering elements haven't been shown in the mammalian systems, the promoters are shown to communicate only with a subset of enhancers in a developmental locus like the *Fgf8* locus. The other genes in the very same locus are isolated from the activity of the enhancer (Marinic M *et al* 2013). In humans, chromosomal translocations lead enhancers to activate genes, which are silent otherwise, and cause human disease (Lettice LA *et al* 2011). This "enhancer adoption" suggests that the regulation of enhancer-promoter communication is not solely based on creating specific loops but also preventing ectopic enhancer-promoter communications. Therefore, the 3D structure of the genome and the other interaction partners of the enhancers and promoter are critical to understand the molecular mechanisms of enhancer-promoter interactions.

2.1.4.4 Physical Interactions and the Conformation of the Genome

There are two independent approaches to understand the chromatin organization in the nucleus. One of them is visualization of the nuclei via probes by microscopes to estimate physical distance (for methodology Annex_Methodology for Microscopy Based Studies) and the other one is biochemical methods based on proximity ligation to estimate physical interactions and overall conformation (for methodology Annex_Methodology for Chromosome Conformation Capture Based Studies). These methods revealed several features of the genome as a large polymer.

The chromosomes occupy a distinct space in the nucleus and only at the periphery of these territories, inter-chromosomal interactions are observed (Bolzer A *et al* 2005; Branco MR and Pombo A, 2006). The nuclear position of a locus influences its transcriptional activity and the repression of the transcriptional activity also changes the nuclear location of the gene (Lundgren M *et al* 2000). For example, the chromosomal territories in the close proximity of the nuclear lamina show decreased expression (Towbin BD *et al* 2012). Moreover, Wendy Bickmore lab has shown that displacement of a gene outside of its native chromosomal territory correlates with the increased activity of *Hoxb* gene. This suggests a looping-out mechanism for the transcriptional activation by distant enhancers (Chambeyron S and Bickmore WA, 2004). Furthermore, Peter Fraser's lab showed that distant genes colocalize in RNAPII foci in a transcription initiation manner (Mitchell JA and Fraser P, 2008; Osbourne C *et al* 2004). In an independent experimental system, ChIA-PET by PNAPII pull down also showed that intra-chromosomal promoter-promoter interactions are very prominent (Li G *et al* 2012). Even if the genes are not co-transcribed, they are clustered in a non-random fashion in the nuclei (Shopland LS *et al* 2006). However, colocalization of enhancer and promoter did not always lead to gene expression (Amano T *et al* 2009).

Chromosome Conformation Capture derived methods show the interaction profile of the regulatory regions and overall shape of chromosome conformation. In the regions where functionally critical candidate sites are known like in ß-globin locus, semi quantitative or quantitative 3C and 4C experiments showed differential interaction among regulatory elements upon differentiation (Tolhuis B *et al* 2002). Similarly, studies in H19/Igf2 loci, mouse X chromosome and drosphila PcG regulated genes supported the views that there are loops, which are organized or anchored at CTCF and cohesin bound DNaseI hypersensitive sites (Splinter E *et al* 2006, Simonis M *et al* 2006, Bantignies *et al* 2011, Parelho V *et al* 2008). However, interaction profile did not reflect the changes in the gene expression (Splinter E *et al* 2011). Besides, 5C and Hi-C experiments showed that the active genes cluster together and inactive genes cluster together. Chromatin modeling based on the physical proximity and knot-free structure of the genome Lieberman-Aiden *et al* postulated that the genome is organized in a fractal globule structure at megabase scale. Further Hi-C and 5C experiments with higher resolution revealed a sub-megabase structure called Topologically Associated Domains (TADs). Further analysis of regulatory landscapes in Spitz lab suggested that all the regulatory activity distribution is confined in TADs (Symmons O *et al* in review). However, the interaction profile within a TAD changes in a tissue specific manner (Philips-Cremins *et al* 2013).

2.1.5 Summary and Open Questions

Distant non-coding regions show very dynamic interaction profiles. 5C, HiC and 4C data show that the interaction partners of the promoters and regulatory elements are mostly confined in topological structures called TADs. 3C and 4C data revealed that both enhancers and promoters contact with multiple sites within the TADs. These findings postulate a facilitated loop structure between regulatory archipelagos and promoters in active transcription (Figure 2B). However, the molecular mechanisms that influence the interaction profile of the promoters and the enhancers are not known. The functionality of the interactions is yet elusive.

2.2 *c-Myc/Pvt1* Flanking Locus

In this section, I will introduce mouse *c-Myc/Pvt1* flanking locus and explain why it is a convenient model to investigate the role of regulatory organization in gene expression. *c-Myc* proto-oncogene is located in 8q24 region of human genome and chr15:61.8M region of mouse genome and it is the only protein coding gene in a three megabase window surrounded by *Gsdmc* cluster and *A1bg* genes in both human and mouse (Figure4). *c-Myc* is a non-classical developmental gene, which is expressed at detectable levels in all tissue types but has stronger expression in certain embryonic tissues like liver, somites and face. Despite being gene poor region, the sequence and synteny conservation between mouse and human is remarkably high (Figure4). Non-coding regions in human *c-Myc* locus are associated with many cancer, immunity or development related disease and disorders. There is no *c-Myc* coding sequence mutation reported in a developmental disease in humans. This is possible due to the fact that *c-Myc* is an essential gene and the null allele is embryonically lethal (Davis AC *et al* 1993). Intriguingly, *c-Myc* upregulation leads to tumorigenic phenotype (reviewed in Dang CV, 1999). Therefore, *c-Myc* level has to be tightly controlled.

Genome wide association studies (GWAS) indicate that changes in distant regions of this locus increase susceptibility to acquire tissue specific cancers. These associated linkage disequilibrium blocks are further away from *c-Myc* gene by tens and hundreds of kilobases. In addition to these associations, translocations 55-340kb far from *c-Myc* gene are shown to underlie Burkitt Lymphoma (Joos S *et al* 1992). Besides, chromosomal duplications in the telomeric end of this locus are reported in childhood acute myeloid leukemia (AML) (Radtke I *et al* 2009). The recent sequencing of HeLa genome revealed HPV-18 insertion 500kb upstream of *c–Myc* gene (Adey A *et al* 2013). These examples show that the prominence of long-range regulation in this region. Therefore, *c-Myc* locus is a very good model to investigate genome organization in a non-classical developmental gene context.

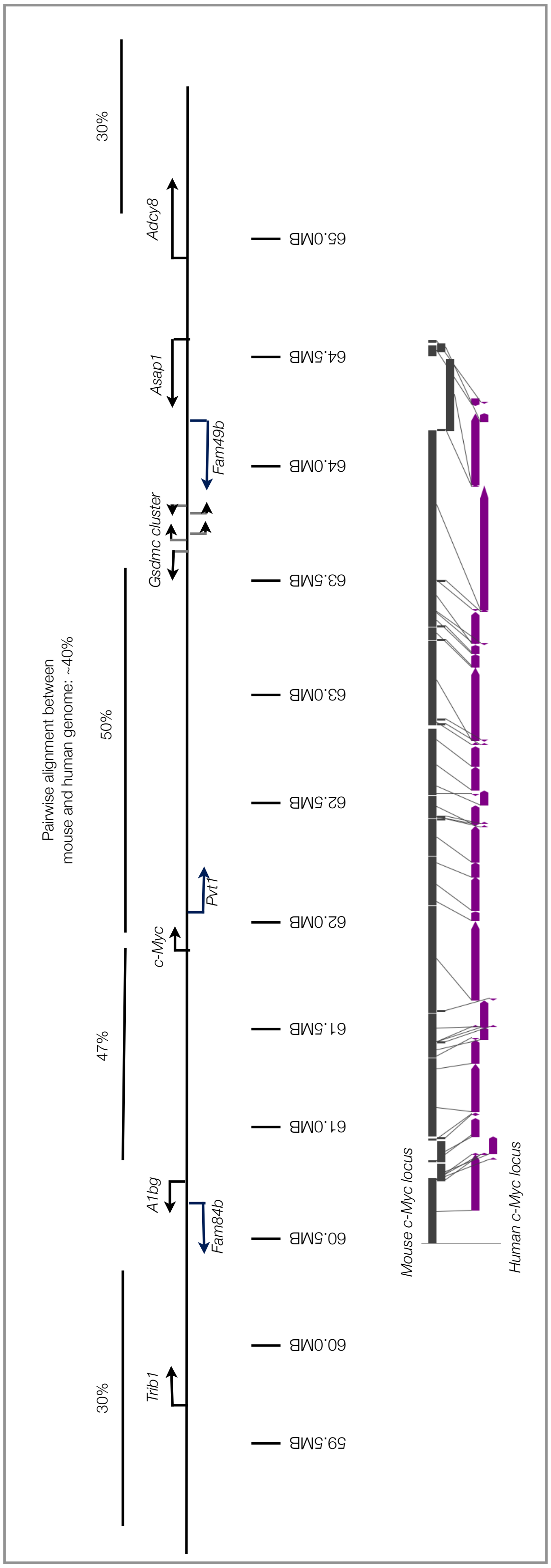


Figure3| Sequence and Synteny conservation in *c-Myc/Pvt1* flanking Locus: The *c-Myc/Pvt1* flanking locus is defined for the interval between *A1bg* gene and *Gsdmc.* Despite being gene poor, both the centromeric and the telomeric side of the *c-Myc/Pvt1* flanking locus show higher sequence conservation between mouse and human. However, the gene poor region on the centromeric side of *A1bg* and the telomeric side of *Adcy8* genes have lower overall sequence conservation between human and mouse according to pairwise alignment algorithm. In the lower panel the synteny of the locus is compared between human and mouse genome. The synteny of the locus is almost perfectly preserved between mouse (gray) and human (purple).

2.2.1 Organization and Evolution of the *c-Myc/Pvt1* Flanking Locus

*c-Myc (*cellular Myc) gene is discovered and named due to its resemblance to myelocytomatosis viral oncogene (Dalla-Favera R *et al* 1982). A weakly homologous gene to *c-Myc* is found in Drosophila melanogaster, called dmyc. Despite a number of overlapping cellular functions between *c-Myc* and dmyc, the deletion of these genes give rise to different phenotypes in mouse and in drosophila (Bellosta P and Gallant P, 2010). Nevertheless, *c-Myc* deficient mouse embryonic fibroblasts can be rescued by dmyc (Trumpp A *et al* 2001). Therefore, despite the limited sequence conservation, the protein is functionally preserved from ecdysozoa to deuterostomes.

c-Myc gene is found in amphioxus as well as in vertebrates. With the recent available whole genome sequencing, it appeared that not only the coding part of the gene but also some degree of synteny conservation dates back to 550 million years ago (Putnam NH *et al* 2008). Both in mouse and humans, *c-Myc* gene locus show substantial synthenic similarity. *c-Myc* locus is defined here as the region between the *A1bg* and the *Gsdmc* cluster (Figure3). Within the locus, on the telomeric side of *c-Myc* gene, there is a long non-coding RNA (lncRNA) gene *Pvt1*. Outside of the locus, F*am84b* and *Trib1* genes reside at the centromeric side*, Fam49b, Ddef1, and Adcy8* genes lie on the telomeric side. One of the very striking features is that the *Fam49b* and *c-Myc* synteny is conserved even in amphioxus. In zebrafish, the *cMyc* locus is duplicated and two paralogous Myc loci appear with *c-myca* and *fam49a*, and *c-mycb* with *fam49b* (Marandel L *et al* 2012). In the chicken *c-Myc* locus, the whole synteny from *Trib1* to *Asap1(Ddef1*) is preserved. Besides, the *c-Myc* locus is duplicated in chicken as well and gave rise to the paralogous *Mycn* locus. This relatively smaller locus possesses *Fam84a, Mycn,* and *Fam49a* genes. The synteny of this paralogous *Mycn* locus is preserved in humans as well as the rest of the mammals. In a knock in study in mouse, *Mycn* is shown to replace *c-Myc* without any survival or reproductive consequences (Malynn BA *et al* 2000). This suggests that not only the coding sequence but also the genomic context of both myc genes have functional roles and upon duplication they might have gained new isolated regulatory domains.

2.2.2. *c-Myc* Gene in Development and Disease

c-Myc is a cHLHLZ transcription factor that regulates multiple cellular processed like cell proliferation, growth, differentiation, apoptosis, and energy metabolism. It binds to a very prominent E-box motif, that thousands of promoters contain. It heterodimerizes with multiple proteins including Max and Mad and this heterodimer influences transcription of target genes depending on the interaction partner (reviewed in (Grandori C *et al* 2000)). One of the most well characterized cellular function of *c-Myc is* G1-S phase transition via activation of several cell cycle proteins like cyclin D1, cdk4, and cdk6 and inactivation of other cell cycle checkpoint proteins like p27^{KIP1} (Mateyak MK *et al* 1999, Obaya AJ *et al* 1999). In addition, *c-Myc* is known to increase transcription of ribosomal genes via RNAPI (Grandori C *et al* 2005). Besides, *c-Myc* is associated with a range of vital cellular activities like apoptosis and differentiation. ChIP analysis of different cell lines indicated that *c-Myc* binds to around 11% of the mammalian genes and influence a number of different cellular pathways (Fernandez PC *et al* 2003). Due to being in the hub of many cellular networks it has been shown that slight changes in *c-Myc* activity can lead to distinct cellular phenotypes. Short half-life of *c-Myc* mRNA and protein allows rapid and dramatic changes in the *c-Myc* activity. Therefore, it requires robust regulation at transcriptional, translational and post-translational level (reviewed in (Wierstra I and Alves J, 2008).

2.2.2.1 *c-Myc* in Mammalian Development

Functional characterization of *c-Myc* is mostly done in cell lines. It's role in many pathways including Wnt, Tgf-ß, and Notch (You Z *et al* 2002, Yagi K *et al* 2002, Palomero T *et al* 2006). However, myc biology is still a black box in developmental context. c-Myc knock out in mouse is lethal before embryonic day 10.5. Starting from fifteen to eighteen somite stage, c-Myc null embryos show overall growth delay when compared to wild-type littermates. Apart from the size problem that is common in all embryonic organs at e9.5, heart and neural tube show morphological problems in *c-Myc* null embryos. Heterozygous deletion of *c-Myc* does not reveal any clear morphological problems in this embryonic stage, whereas it causes reduced fertility in adult females (Davis AC *et al* 1993).

Andreas Trumpp obtained *c-Myc* alleles in mouse, where the expression is relatively lower than the *c-Myc* expression from the wt allele. In the lower *c-Myc* expressing genotype, the embryos from day 16.5 and the pups show significant size reduction. Unlike drosophila myc mutants, in which cell number doesn't change but cell size shrinks, mouse *c-Myc* mutants have less cells without change in the cell size. This reduction in the organ size is reported to be directly proportional for organs like liver, brain, and lungs. On the other hand, size of muscle, connective tissue, skin and skeleton was reduced more prominently than the other organs. This implies differential dependence of tissues to *c-Myc* levels. Embryonic fibroblasts from low *c-Myc* expressing mice exhibited slower proliferation when compared to MEFs from wildtypes (Trumpp A *et al* 2001).

In addition to the embryonic phenotypes, conditional deletions of *c-Myc* demonstrated its role in intestinal crypt formation, differentiation in the pancreatic cells, the mammary gland development in the pregnancy, and the melanocyte maintenance (Stoelzle T *et al* 2009, Pshenichnaya I *et al* 2012, Bonal C *et al* 2009, Bettess MD *et al* 2005). Apart from the downregulation of *c-Myc*, functional characterization of this proto-oncogene is carried out by overexpression models.

One of the first mice generated with *c-Myc* overexpression is a fusion between IgH enhancer and *c-Myc (Eμ-myc* mice). 95% of these mice develops severe lymphoma (Adams JM *et al* 1985). Further enhancer *c-Myc* fusions with Igλ and *c-Myc* also demonstrated lymphoma phenotype but with different pathology than IgH *c-Myc* or IgL *c-Myc* lymphomas (Kovalchuk AL *et al* 2000). This suggests that according to the specific cell type that *c-Myc* is overexpressed, the phenotype may differ up to certain extend.

2.2.2.2 *c-Myc* in Hematopoietic System

One of the most evident phenotype upon *c-Myc* overexpression and deletion is hematopoietic deregulation. Although deletion of *c-Myc* is lethal before day 10.5, embryos with epiblast-specific deletion of *c-Myc* via Sox2Cre survive up to e12. In other words, when *c-Myc* function in placenta is restored, the lethality is delayed. A striking phenotype between epiblast specific *c-Myc* after day 11.5 is that the embryos are completely pale and lack red blood cells both in liver and also in dorsal aorta (Dubois NC *et al* 2008). E11.5 is a critical stage where hematopoiesis moves from placenta to embryonic liver (reviewed in (Rieger MA and Schroeder T, 2012)). Therefore it implicates a possible involvement of hematopoietic deregulation in the lethality *c-Myc* null embryos. Moreover, hematopoietic system specific deletion of *c-Myc* via *Mx-Cre* driver give rise to acute anemic in mice due to the almost complete loss of differentiated hematopoietic cells (Wilson A *et al* 2004). Interestingly, as stated before, *c-Myc* overexpression leads to malignancies due to deregulation of hematopoietic system. Both overexpression and downregulation of *c-Myc* indicate it is a critical regulator of hematopoietic system.

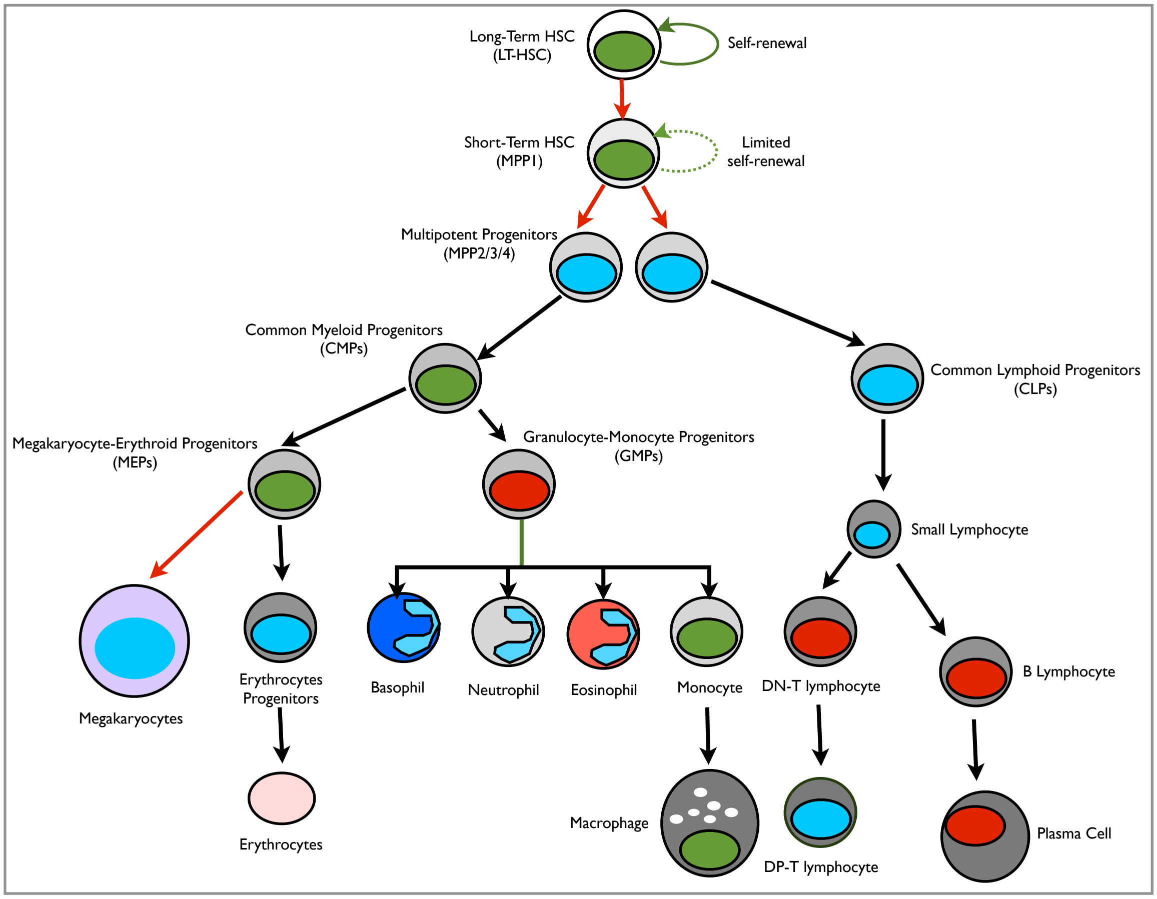


Figure4I Hematopoietic Lineage and the function of *c-Myc* in hematopoietic system: Red nuclei indicates a positive role for *c-Myc* in this lineage, green nuclei indicates an inhibitory influence of *c-Myc* in the given lineage, and blue nuclei indicates not enough or contradictory data for *c-Myc* activity. Red arrows indicate requirement of *c-Myc* through differentiation (Adapted from Delgado MD and León J, 2010, Wilson A *et al* 2008, Guo Y *et al* 2009 and from http://en.wikipedia.org/wiki/Haematopoiesis

The function of *c-Myc* in the hematopoietic cells is extensively studied in cell culture models as well as in vivo mouse models (Figure4). Among the cell types, in which the role of *c-Myc* has been studied, Long-Term Hematopoietic Stem Cell (LT-HSC and Short-Term Hematopoietic Stem Cell (MPP1) populations have shown to consistently respond to *c-Myc* levels in different *in vivo* contexts. Irving Weissman's lab showed that these HSCs are able to self-renew and replenish the whole hematopoietic system when they are transferred to irradiated mice (Sapngrude GJ *et al* 1988). In MxCre driven hematopoietic system specific deletion of *c-Myc*, LT-HSC population dramatically expands but loses its differentiation capability. Moreover, *c-Myc* overexpression via viral transduction lowered the number of HSCs (Wilson A *et al* 2004) (Figure 4). Not only the *c-Myc* transcriptional downregulation but also the post-translational destabilization of the *c-Myc* protein via hematopoietic lineage specific (MxCre driven) Fbxw7 ubiquitin ligase deletion, recapitulated the *c-Myc* deletion phenotype, and expanded LT-HSC population. On the other hand, the influence of *c-Myc* in differentiated blood cells is more complicated than it is in HSC. For example, overexpression of *c-Myc* in *Eμ-myc* mice hatched more pre-B cells and less mature B cells. For T-cell development, *c-Myc* deletion diminishes T-cell population in thymus but does not affect T-cell maturation (Dose M *et al* 2006). On the contrary to lymphoid cells, conditional *c-Myc* deletion lead to a significant increase in megakaryocyte population, whereas, a significant depletion of granulocyte, macrophage lineage. Nevertheless, despite being dispensible for maturation, *c-Myc* is essential for endomitosis and polyploidy of megakaryocytes (Guo Y *et al* 2009) (Figure 4).

2.2.3 Regulation of *c-Myc* from long-range

2.2.3.1 Retroviral Insertions and Chromosomal Translocations

One of the intriguing observations that links *c-Myc* and hematopoiesis comes from the retroviral insertion screens that induce blood cancer phenotypes. Moloney murine leukemia virus insertions that give rise to T-cell lymphomas in rats were mapped to the centromeric side of *c-Myc* gene between 50 to 270kb far from *c-Myc* promoter (Figure 5). These distant insertions upregulated *c-Myc* expression and caused T-cell lymphomas (Lazo PA *et al* 1990). As stated before, various breakpoints that induced Burkitt's lymphoma with t(8:14) translocation are mapped to 55-340kb centromeric side of *c-Myc* (Joos S *et al* 1992). Although these translocation brought Ig enhancers in *c-Myc* locus and induced Burkitt's lymphoma, the histology of these mice were different from *Eμ-myc* mice. Nevertheless, these examples showed that not only changes in or close to the coding sequence of *c-Myc* but also the changes in the *c-Myc* locus induce tumor formation (Figure 5).

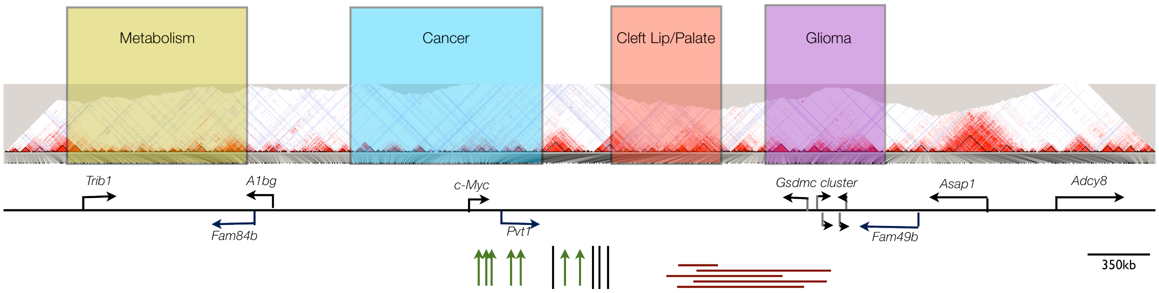


Figure5| Variations in the c-Myc/Pvt1 flanking locus and their phenotypic outcome: The rectangular blocks indicate genomic variations associated with disorders or malignancies. LD disequilibrium block is given in red triangles are for CEU population. Retroviral insertions that cause lymphoma in mice are represented in green arrows and Burkitt's lymphoma breakpoints in humans are represented in black perpendicular lines. Horizontal red lines represent the duplication regions in childhood acute myeloid leukemia patients and related cell lines.

2.2.3.2 GWAS in *c-Myc* locus

Recent GWAS point out various genomic alterations in humans increase the susceptibility to acquire certain cancers or developmental defects. Intriguingly, 40% of the GWAS hits do not overlap with any known exonic sequences (Visel A *et al* 2009). This shows that the regulation through long-range interaction is genome-wide phenomena. This abundance of GWAS hits in non-coding regions is very striking in *c-Myc* locus (Figure 5). According to the GWAS catalogue of UCSC genome browser, outside of *c-Myc* locus around TRIB1 gene, there are a number of SNPs are associated with macular degeneration, lipid metabolism, and childhood obesity. As soon as the SNPs on the centromeric side of the *c-Myc* locus are investigated, they are associated with increased risk of cancers in a tissue specific manner. More than 10 SNPs representing partially overlapping LD blocks coming from Japanese, Chinese, European population are shown to be a risk allele for prostate cancer.

Moving from centromeric to telomeric SNPs, very strong associations have been obtained between several other SNPs and breast cancer and especially colorectal cancer. In addition, a SNP 50kb far from *c-Myc* promoter is associated with urinary bladder cancer in European and Icelandic population. On the telomeric side of *c-Myc* gene in the locus ovarian cancer, lymphoma and autoimmunity related disorders are associated with genomic variations in Swedish and European ancestry population. In addition to these malignant susceptibilities, three independent groups showed very strong association between nonsyndromic cleft lip and palate (NSCLP) phenotype and a SNP that resides in 640kb LD block (Birnbaum S *et al* 2009; Mangold E *et al* 2010). This NSCLP allele is very common in European and Northern American population and the odd ratio for affected homozygous carriers of associated allele is between 3.5 and 6 for both of these populations. In contrast, Asian and African population does not show a strong associated for this SNP probably due to the scarcity of the risky allele in these populations (Beaty TH *et al* 2010). At the telomeric end of *c-Myc* locus there is strong association with glioma. Outside of the locus, on the telomeric side, changes in the genome are associated with neurological disorders like Multiple Sclerosis and Alzheimer's disease. To sum up, despite being hundreds of kilobases far away from any protein coding gene, changes in the *c-Myc* locus has phenotypic consequences. This clearly indicates an active long-range regulation, which takes place from the locus (Figure 5, full reference list is in Supplementary Table 1).

In order to identify the distant regulatory elements in the centromeric side of *c-Myc* locus, operational and functional assays are combined with GWAS data. BAC mediated enhancer traps showed prostate and breast specific enhancer activity in the centromeric side of *c-Myc* locus, where 440kb LD is associated with prostate cancer. Interestingly, the SNP that represents the risk allele in this LD resides in the prostate enhancer. The enhancer activity increased when non-risk allele T is replaced with risk G allele (Wasserman NF *et al* 2010). This SNP is also shown to be important for colorectal cancer susceptibility (Tuupanen S *et al* 2012). However, the in vivo deletion of this enhancer element that includes the SNP did not change myc expression in the colon crypts. Surprisingly, when the mice with enhancer deletion are investigated in colon cancer susceptible mice (APC$^{min/+}$), they were resistant to polyp formation in the colon (Sur IK *et al* 2012).

3. PROJECT MOTIVATION and OBJECTIVES

Chromosomal rearrangements in mammalian genome showed that the regulatory elements for the developmental genes are up to hundreds of kilobases far away from their target genes. Recent GWAS point out that long-range regulation is not restricted in the developmental genes but more than 40% of the genomic variations, which have been associated with a phenotype, are in the non-coding regions. Moreover, although 2% of the genome is protein coding, ENCODE project revealed that 80% of the genome exhibits biochemical activity, some of which are associated with regulatory activity. Therefore, the long-range regulation is a fundamental phenomenon that controls the spatiotemporal expression of genes.

c-Myc/Pvt-1 flanking locus in mouse is a very convenient model to investigate the role of genome organization in long-range gene regulation. First of all, this locus contains a 3-megabase long gene-poor interval, in which the *c-Myc* proto-oncogene is the only protein-coding gene. The sequence and the synteny of the *c-Myc/Pvt-1* flanking locus are highly conserved between mouse and humans. Furthermore, GWAS studies indicate that many variations in this locus (8q24) are associated with increased susceptibility to cancer and developmental problems in a tissue and stage specific manner. Moreover, the retroviral insertions, the duplications, translocations in this locus also caused cancers. These studies suggest that long-range gene regulation is very prominent in this locus, which makes it a good model to investigate the nature of gene regulation via distant enhancers. Last but not least, the characterization of the regulatory elements, the nature of the regulatory information encoded by these elements will contribute to the elucidation of the etiology of 8q24 related health problems.

c-Myc gene is a non-classical developmental gene, which has a widespread expression in the embryonic development but the expression level is higher in certain tissues. Downregulation of *c-Myc* leads to developmental problems, whereas, upregulation give rise to tumorigenic phenotype. Therefore, the regulation of *c-Myc* is tightly controlled. Extensive studies in *c-Myc* promoter suggested that the promoter structure itself was not enough to shed light on the complex regulation of *c-Myc* in embryonic development, implying the presence of more distal elements that involve in this complex regulation. Therefore, characterization of this locus will contribute to the understanding of the *c-Myc* biology and may give insights on the regulation of other non-classical developmental genes.

During the course of my PhD, I aimed to find out the functional regulatory regions that act on endogenous genes in the *c-Myc/Pvt-1* locus and the mechanisms that allow the communication between these regulatory elements and their target genes. For this purpose, I created large of panel of regulatory reporter insertions to monitor the regulatory activity in the locus. Then, I investigated the following questions in this locus:

- Where are the regulatory elements?
- What are the target genes of these elements?
- What are the phenotypic consequences of changes in the regulatory regions?
- How do the regulatory elements communicate with target genes?

4. MATERIALS and METHODS:

4.1 Materials

4.1.1. Instruments

4.1.1.1 Centrifuges

ID	Company	Notes
Microcentrifuge 5424	Eppendorf	Benchtop Centrifuge
Microcentrifuge 5415 R	Eppendorf	Refrigerated Centrifuge
MultiFuge 2SRT	ThermoScientific	Centrifuge for Falcons/Plates
Megafuge 2.0R	Heraeus	Centrifuge for Falcons/Plates
Sorvall RC6T	ThermoScientific	Centrifuge for >100ml /Maxi-MiniPrep

4.1.1.2 Thermocyclers

ID	Company	Notes
PTC-200 DNA engine	BIO-RAD	Standard PCR
C1000 Thermal Cycler	BIO-RAD	Standard PCR
C1000 Touch ThermoCycler	BIO-RAD	Standard PCR
S1000 Thermocycler	BIO-RAD	Standard PCR
ABI7500 Light Cycler, 96 Well	Applied Biosystems	Real Time PCR
StepOne™ Real Time PCR Systems, 96 well	Applied Biosystems	Reat Time PCR

4.1.1.3 Microscopes and Light Sources

ID	Company	Notes
KL1500LCD	Schott AG, Meinz	Light Source
Leica M80	Leica	Dissection Microscope
Leica M16	Leica	Dissection Micropscope
Leica M16F	Leica	Dissection and Documentation
Zeiss Axiovert200	Zeiss	Widefield, ALMF

Zeiss Cell Observer HS Automated	Zeiss	Widefield, ALMF
PE Ultraview VoX	Perkin Elmer	Spinning Disk, ALMF
PE Ultraview ERS	Perkin Elmer	Spinning Disk, ALMF
Bioptonics 3001M	Bioptonics	OPT

4.1.1.4 Incubators

ID	Company	Notes
Hybridization Oven	Helmut Saur Laborbedarf	65°C incubator
Hybridization Oven	Binder	55°C incubator
HyBaid Shake'n'Stack	Thermo Scientific	65°C shaking incubator
TW8	Julabo	Water Bath

4.1.1.5 Histology Equipment

ID	Company	Notes
Tissue-TEK®II	Sakura	Slide Staining Tanks
EG1160	Leica	Paraffin Embedding Platform
HI1220	Leica	Hot Plate
HI1210	Leica	Hot Water Tank
RM2255	Leica	Paraffin Block Sectioning Tool
VT1200S	Leica	Vibrotome

4.1.1.6 ChIP Equipment

ID	Company	Notes
DynaMag™2	Invitrogen	Magnets for Beads
Bioruptor®	Diagenode, Belgium	Sonicator Water Bath
Test Tube Rotator	Labinco	For ChIP and 4C

4.1.1.7 Spectrophotometers

ID	Company	Notes
Ultraspec3000	Pharmacia Biotech	Spectrophotometer with cuvettes
NanoDrop 8000	Thermo Scientific	Spectrophotometer 8x
ND-1000	NanoDrop®	Spectrophotometer 1x

4.1.2 Chemicals

Unless indicated otherwise, all chemicals were supplied by Sigma-Aldrich (Steinheim)

4.1.3 Buffers

Unless indicated otherwise, all solutions are prepared according to (Sambrook and Russell 2006)

4.1.4 Commercial Kits

ID	Catalog No	Company
EpiTech Bisulfite Kit	59104	QIAGEN
QIAquick PCR Purification Kit	28104	QIAGEN
QIAquick Gel Extraction Kit	28704	QIAGEN
illustra GFX Purification Kits	28-9034-70	GE Healthcare
Rneasy Mini Kit	74104	QIAGEN
MinElute Gel Extraction Kit	28606	QIAGEN
NucleoBond®-Xtra Midi Prep	740410.1	Macherey-Nagel
NucleoBond®-Xtra Maxi Prep	740416.1	Macherey-Nagel
NucleoBond® Xtra Maxi EF	740424.1	Macherey-Nagel
Dual Luciferase Kit	E1960	Promega
PCR DIG Probe Synthesis Kit	11636090910	Roche
DIG RNA Labeling Kit (SP6/T7)	11175025910	Roche
Dneasy Blood and Tissue Kit	69506	QIAGEN
pGEM®-T Easy Vector System I	A1360	Promega

4.1.5 Enzymes

4.1.5.1 DNA Polymerases for genotyping and transposon mapping

ID	Supplier	Notes
In House Taq	Self Made	The construct is provided by EMBL Protein Expression and Purification Core Facility
Expand Long Range dNTPack	Roche	Long-Range Genotyping
Platinum® Taq	Invitrogen	Transposon Mapping
Phire Hot Start II	Thermo Scientific	Quick PCR from yolk sac

4.1.5.2 Cloning Related Enzymes:

ID	Company	Notes
All restriction Enzymes	Fermentas	currently Thermo Scientific
T4	Roche	Ligase
T4	Promega	Ligase
LA Taq	TaKaRa BiO Inc.	PCR for Cloning
Phusion Taq	Finnzymes	PCR for Cloning
Klenow Frangment	Fermentas	currently Thermo Scientific

4.1.5.3 Other Enzymes

ID	Company
Rnase A	QIAGEN
Dnase 1	Roche

4.1.6 Bacterial Strains

Strain	Source	Notes
Dh5-alpha	Invitrogen	Standard cloning
Stbl3	Invitrogen	Lentiviral cloning
EL250	Lab Stock	BAC engineering
DY380	Lab Stock	BAC engineering

4.1.7 Plasmids and BACs

4.1.7.1 Plasmids

Plasmid	Source	Notes
pRRLSIN.cPPT.PGK-GFP.WPRE	Trono Lab	Lentiviral Cloning
RRL(Pme1-ßlab) INS	Trono Lab	Lentiviral Cloning
pMD2.G	Trono Lab	Lentiviral Cloning
pCMVR8.74	Trono Lab	Lentiviral Cloning
pGPS1	Lab Stock	BAC engineering
pRS16(ßlac frtKANfrt)	Lab Stock	BAC engineering
pGEM®-T easy	Promega	PCR product cloning
pCAGGS-puro linker	M.Treier	Expression vector

4.1.7.2 BACs

ID	Backbone	Locus (mm9)	Company
350P5	RP23	chr15:62,844,349-63,059,778	CHORI

4.1.8 Oligonucleotides

All Oligonucleotides are ordered from Sigma Aldrich and the stock concentration is 100μM.

4.1.8.1 Oligos for Genotyping

Primer ID	Sequence	DB number
179039 L	GGTTCTCTCTCATGGAGTGTATCAGG	2432
180206 R	ATACCACCATGCTTGGCTTGAC	3155
184347 R	ATCCCATGAAAGGCATGGAGAG	2431
188217 L	TGCATTTTTCACTCAGCCTAA	2418

188217 R	AAATAAAAGGGGCCATGGAG	2419
188152 L	AGGTCGGCTAACACGGTATG	2379
188152 R	AAACGAAGCACAGGCAGATT	2415
188150 L	ATGGTTGGCCAAAGAAGTTG	2373
188150 R	AATGTGGCCACTCTCTTTGC	2374
188219 L	CGGTCTAGTTTGGCAAAAGG	2375
188219 R	TGTGAAATGCCACAGACCTC	2376
188220 L	TGCTAGGCATTCAAGAGGAAG	2377
188220 R	GGGTATGTCAGGCAAGCACT	2378
188226 L	AGGTGACTGTTGCTCCATCC	2416
188226 R	ACGATGTTACCTGACAAGCA	2417
188313 L	CAAAATCAAATGTGGGCACAAGA	2651
188313 R	ACTCTGGATCATCTGCCCAAGAA	2650
190909 L	ATCCCATGAAAGGCATGGAGAG	2431
192331 L	CAGGAGGCTTTGGACTCAACACT	2645
192331 R	CATCGTCGTCGTGCTTTGAAAT	2644
192331R	CCTCTTTTGCCAACGTCTTCC	3575
192333 L	GCTGTGCACCATCTGAACATAGG	2627
192333 R	TGAAGTGTGCTGATGGTGATGC	2626
192334 R	CAATTGACATGCCATCTTGTGG	3021
192339 L	AATGCCAAAGACAAGGACTCCAG	2647
192339 R	GATGGGACTTCCCACATAACAGC	2646
192341 L	TGTGGGGGTTAAATTGGGTGTAA	2649
192341 R	TCAGGAGCTAGGATTCTAGAGGGTCT	2648
192566 L	CATAGCTCTGAGTGCCTCCAAAAG	2811
192566 R	TGAGTATTTTTGCATCGATATCATAACA	2812
192571 L	ACCCTTGGCTGAAGACATACCA	2813
192571 R	CAGGACTCCAGTCATGTGATGC	2814
192627 L	CGGAGCCACTAGGGCTATGAG	2815
192628 L	TCAAGCTGGCCTATGGGTACA	2816
192642 L	GGAAAAGAGGAAGCAGCCCTTA	2817
192644 L	AGGGCACAGGAGAATGAAGGAC	3020
192842 L	GCCTCGAGTGCTCGATAAGGT	2818
192857-R	GGTTCTCTCTCATGGAGTGTATCAGG	2432
193047 R	CATCGTGGTGATTTTGCACTG	2819
193058 L	TTGGGTACATCTGTCACCAGAGTC	2820
193058 R	TCAGAGTGTGGTCAACTGTGGAA	2821
193083 L	CCTCTGTGTCCTTCCCCTACCT	2822
193284 L	GATAAGTTTCCTTCCCCCATCG	2823
193284 R	CACATTAGTGCGACCCATTCAA	2824
193315 L	TAACCAGTACCCCCTGGAGCTT	2800
193315 R	ATCCACTCCCCTCTCTGTTTCC	2801
193317 L	CAGAGACAGAGTGCACAGCACA	2802
193428 L	GGTGACAGGCTTGCTCTGTGT	2803
193637 L	GCATGGATTCTATGGGTGTTGG	3030
193637 R	CCTCCTGGGATTTCCATGACTC	3031

193642 L	TTCTCTGGGTTGGAAGCTGTG	2804
193642 R	AATCGGCCCACAGTTCTGAAT	2805
193944 L	GGTGTTGAAGTCAGAGCCTGGA	2809
193970 L	TGCTCAGTCCAGTGGATGACTATG	2810
193970 R	TGGTTGCCTTTTTGTCTGATTGT	2837
194242 L	CCCTCCCCGCTAGGTAAATCTC	3025
194272 L	CCTGTGTGTTCATCTGGGGAGTC	3026
194278 L	GACACAATGTGCACAAGCATCAG	3027
194278 R	AGACAAACAGATGCTTCGAACTCC	3028
194279 R	TCAACTGGAACTGTGTGGATGAAC	3029
194575 L	CCACAAGTAGATCAGCCACAAACC	3032
194575 R	ATTGTTGGCAAAACACAACAGG	4185
194577 L	CTTCACACTTGACAAGGGGTGTG	3042
194577 R	TGTGTTTGGACACGGAAAATGAC	3043
194578 L	CAGGAGTTTGCCAATCAACAGTG	3033
194578 R	GAAAGCAAGTGGGGAAGTCAGAG	3034
194832 R	CCTCTCATGTTGACAGTCAAGACG	3044
195052 L	TTGGAATTTGAAAACGACATTGG	3038
195052 R	GCAGTCTGCTTGTTTGTTTGTTTG	3039
195297 L	GGATCATTGCCCAGTTTTAGTGG	3048
195304 R	TGGGACACACTTCCATTGACAAG	3040
195308 L	CCATGAGAGCTGGAGAGAGTCTTG	3049
195308 R	AGTTATTGTCCGGTCAGGCAAAG	3050
195469 R	TTTTCAACCCCAAGCATATGGAG	3059
195474 L	GAAGCAACATTTTAGAAGGCCTGTG	3051
195474 R	GTTCATCACTGGTTCCTTCACTCC	3052
195496 L	GCGCTGGAGAAATCTCAAGAAAG	3053
195496 R	GGCCAAATGTATTGTATGGCAAAAG	3054
195936 L	TGTGCAACCTGGCGTACATTAAC	3062
195936 R	ATGGAGGAGAGGGATTCAAGAGC	3063
195964 L	AACCCCACTTCCTGAACCACTG	3055
195964 R	TGTGTCACACTGGTGGAAAAGAAAC	3056
195968 L	TCACATGTCATACAGAGGCAAAGC	3066
195968 R	CTAAGCTTTGCATGGCTCATCAC	3067
195975 L	TTAGTATCCCAACTGGCCCAATC	3057
195975 R	CTTGAAAAGGAGTGGTCCTGAGC	3058
196231 L	CCTGGAATCTCCTTTTGTTCAGG	3180
196231 R	ATCCAACACCCCCTTTCTGC	3580
196337 R	TATCCATAAGGGATGGCAATGG	3485
196554 L	GGCCTAATCCCCTGTAATGACC	3178
196554 R	AAGGGGGCTTGATTTGAATAGC	3179
196558 L	CATCTGCCACAAAACAGACAGC	3176
196558 R	GTCTCTGTTTCCCAAGCACTGG	3177
196560 L	CCAAGAGCAGGCTAGCTTTTCC	3175
196918L	ATGAGTCATGCAGGCTCTCTCC	3367
196925L	TCTCTGAGTCCCCCAAGACTCC	3368

197270 R	TTGCTCAGAGGGGGAAAATAGC	3174
197272 L	CAACTCTCTGCTCCACTGATGC	3172
197272 R	TGGTCTTGAAGCATCCTCTTCC	3173
202151 L	GGTTATGGTTTTGCGTTTCAGC	3581
202151 R	TTACAGGAGCACCACCACTTCC	3582
202846R	GATGACTCTGGGGAATGGTCAC	3574
205877L	GACACAGTGCAAGCAAAATCG	3638
205877R	ACAGAAGGGAAACTTGCTGTGG	3637
205880 L	GGACGTTTTGTGCTGAGAAAGG	4186
205880R	TCCTGGAACCTTCTGAAACAGG	3639
211201-L	GAAGCAAAACTGGTTTGTGTGG	4270
195802 L	TGCCAGTTGCTTTTTACTCATCC	3436
195802 R	GGGCGTGGATCATCTTTGG	3437
179039 R	GTTCCTCCCAAGGTTCATGCTC	790
180206 L	GGCTTTGACCCTGACTTTAGG	994
186894 L	CCTTGCCATTGTGTTCTGAG	2369
186894 R	TGATGTGGTGACTGACATCTGA	2370
196234 L	CATTGTTGGCTGTTGGTATTTGG	3484
198932 L	GCATTTATGCATTAGAATGTCTTGG	3434
198933 L	GAGGAGGGGTACACAGAGAGAGC	3469
198934 R	CATCTTTCCACCTTTCCACTGC	3435
209545 L	TTCATGGCAGGTATGATTGTGG	4266
211151 -L	TCACCTGAGCAAGTCTGTCTCC	4267
211151 -R	ACAGGAGGACCCATTAAACACC	4268
211198 -L	TTTTCCCACGACTTCTTTTTGC	4269
SB-L3	AAGTAGATGTCCTAACTGACTTGC	426
SB-R3	TCCTAACTGACCTAAGACAGG	429
SRY1(F)	GTGAGAGGCACAAGTTGGC	3089
SRY2(F)	TCTTAAACTCTGAAGAAGAGAC	3090
SRY4(R)	GTCTTGCCTGTATGTGATGG	3091
ZYF3(F)	AAGATAAGCTTACATAATCACATGGA	3092
ZYF4(R)	CCTATGAAATCCTTTGCTGCACATGT	3093

4.1.8.2 Oligos for Real Time qPCR

Primer Name	Sequence	DataBase_ID
qPCR Csf-1 F	GGAGCTCTGGGACCTGCTC	4943
qPCR Csf-1 R	CTACGTCCCGGTGGATGC	4944
qPCR ApoE F	ACCCTGGAGGCTAAGGACTTG	4945
qPCR ApoE R	TCATCTTCGCAATTGTGATTGG	4946
qPCR Tmem16a F	AAGTAAACGGCGGAAGTGTGG	4947
qPCR Tmem16a R	CATAGTCCCCATCGTGCAGAG	4948
qPCR Nr2F1 F	CATCGTGCTATTCACGTCAGATG	4949
qPCR Nr2F1 R	GATTTCTCCTGCAGGCTTTCG	4950
qPCR Tpi1 F	CTTCGTTGGGGGCAACTG	4951
qPCR Tpi1 R	CGGTGCACAAACCACCTC	4952
qPCR integrin b3 F	ACACCAGTGGGAGGGCAGTC	4953

qPCR integrin b3 R	TATCAGGACCCTTGGGACACTC	4954
qPCR Sox11 F	GGAGCTGAGCGAGATGATCG	4955
qPCR Sox11 R	AACACCAGGTCGGAGAAGTTCG	4956
qPCR nbl-1 F	CCCAGTCCATGTGGGAGATTG	4957
qPCR nbl-1 R	ACAATCTTCTCTACCAGCTTGTCC	4958
qPCR A1bg-201 F	TGGAGCTGCGGGTGAATG	4959
qPCR A1bg-201 R	CCAGATGTACTGTGCTTTTCCAC	4960
qPCR Fam84b F	CCAGGGAAAGGATTCAATTAAGG	4961
qPCR Fam84b R	CACAACAGCAGGCCAAAAACA	4968
qPCR AC164597.1 F	CAGATTTCCAGGAAGAGGACAGA	4962
qPCR AC164597.1 R	CTCCTCCAGTCAAGTCGTGAAG	4963
qPCR AK089020 F	GGGAAACAACCAATTGGGTAAA	4964
qPCR AK089020 R	GGTGGCATAGATCCCATACCC	4965
qPCR eGFP F	GGGCACAAGCTGGAGTACAAC	4743
qPCR eGFP R	CTGCTTGTCGGCCATGATATAG	4744
qPCR cMyc F	CCCTAGTGCTGCATGAGGAGACAC	3006
qPCR cMyc R	CCACAGACACCACATCAATTTCTTCC	3007
qPCR Pvt1 F	CTGAGGTGGAGGAAGTTGCCCTTG	3012
qPCR Pvt1 R	GGCCACCTCAATCAGGCAGTGTC	3013
qPCR Ddef1 F	AAGAACGGGATCCTGACCATCTCC	3014
qPCR Ddef1 R	TGGCAGGTGAGGAGGTTTAACTTAGC	3015
qPCR AK040104.1 F	TGGGAAGTCCTGAGTGAATTACATGC	3016
qPCR AK040104 R	TGCAGTCACTCAGCAAATCTGTAAGG	3017
qPCR Gsdmc F	GCAATCAAAGGGATCATCAACCAG	3018
qPCR Gsdmc R	TGAATCTGTTTTCTCTGTTTGCCACTG	3019
qPCR GusB F	CTCTGGTGGCCTTACCTGAT	1317
qPCR GusB R	CAGTTGTTGTCACCTTCACCTC	1318
qPCR Pdhb F	TGTTGTCCACTCCCTACCCTAGATAC	2900
qPCR Pdhb R	CATTCTTATCTTGCCCCTTCCAGTG	2901
qPCR lacZ F	GCGTTAACTCGGCGTTTCATC	2241
qPCR lacZ R	GCGCTCAGGTCAAATTCAGAC	2242
qPCR Ribosomal Rplp1 -F	CCTGGCTTGTTTGCCAAGG	-
qPCR Ribosomal Rplp1 -R	GCAGTGGATGGAGCAGCAC	-
qPCR Ribosomal Rps20 -F	CCTGACTCACCGCTGTTCG	-
qPCR Ribosomal Rps20 -R	CGTCTTTCCGGTATCTTTAAATGC	-
qPCR Krt18 -F	AGATGACACCAACATCACAAGG	4258
qPCR Krt18 -R	TCCAGACCTTGGACTTCCTC	4259
qPCR Vimentin -F	AGCGTGGCTGCCAAGAAC	4260
qPCR Vimentin -R	GCCTCAGAGAGGTCAGCAAAC	4261
qPCR E-cadherin(Cdh1) F	ATCCTCGCCCTGCTGATTC	4262
qPCR E-cadherin(Cdh1) R	CTCTTTGACCACCGTTCTCC	4263
qPCR Twist1 -F	AGCTACGCCTTCTCCGTCTG	4264
qPCR Twist1 -R	TTCTCCTTCTCTGGAAACAATGAC	4265
Hif1an_Mus_qPCR_F	GTGTACAGTGCCAGCACCCATAAG	2239
Hif1an_Mus_qPCR_R	TAATTTCCTCCCTGTTGGACCTTG	2240

4.1.8.3 Oligos for Cloning

Lentiviral Enhancer Screen	Sequence
8_riskint F	ACTGGCTAGCCAAGAACAGTGCTATCAAGC
8_riskint R	TAGCGTCGACGAGACTACTCTCTGGGAACC
Hit2 Kerstin F	ACTGGCTAGCCATCATGCATATTCACAC
Hit2 Kerstin R	TAGCGTCGACTGAATATTCCATGAACCCG
inDepHumCLP F	ACTGGCTAGCTGGGAGTAGGGAGGCTGTAG
inDepHumCLP R	TAGCGTCGACTAGAAGGTGCAATGAAAGG
(viki)humTrioDel F	ACTGGCTAGCTGCCTTTCCTGTCCATGTTGAC
(viki)humTrioDel R	TAGCGTCGACGTCATCCCCTATCACTTCATCC
veff F	ACTGGCTAGCCCCTTTCTTTACATTGCCTC
veff R	TAGCGTCGACCTCAGACACACAACACTTGC

4.1.8.4 Oligos for Bisulfite Sequencing

ID	Sequence
SB9 promoter (F-alt1) bisulfite	TGTAAAACGACGGCCAGTAAAGGGGGATGTGTTGTAAGG
SB9 promoter (F-alt2) bisulfite	GTATTTGTTAGTTTGAGGGG
SB9 promoter (R-alt1) bisulfite	CAGGAAACAGCTATGACCATCAAAAATTTACATACACTC
SB9 promoter (R-alt2) bisulfite	CTTAAATCAATTAAAATCACCAC
SB9 promoter compl F-1 bisulfite	TGTAAAACGACGGCCAGTGTTTTTAAATTGTTTAATTTGGG
SB9 promoter compl R-1 bisulfite	CAGGAAACAGCTATGACCAATCRCACTCCAACCAACTTTC
179039R bisulfite F	TGTAAAACGACGGCCAGTGTTTAAATGTATTTGGTTAAGGTG
179039R bisulfite R	CAGGAAACAGCTATGACCCCTCTCTCTCCCTTCAAAATTCTAC
179039L+R bisulfite F	TGTAAAACGACGGCCAGTTTGATTAGTGTTTTTGGGATGGG
Chr15-63572030-63572700 (F)-1	ATGTAAAACGACGGCCAGTGAAAGAGTAAGTTTTTATAGTGGG
Chr15-63572030-63572700 (F)-2	ATGTAAAACGACGGCCAGTGAGGGGAAAAAAAAAGGAGAG
Chr15-63572030-63572700 (R)	ACAGGAAACAGCTATGACCCTATAAACTATCTAAACCTACCAC
chr15-63602900-63603650 (F)	ATGTAAAACGACGGCCAGTTGTAGAAGGATTTAATGGTTGG
chr15-63602900-63603650 (R)	ACAGGAAACAGCTATGACCCCAAAACCAAATATAAAACACC
chr15-63553100-63553800(F)	ATGTAAAACGACGGCCAGTGTTTTAAAGATGGTTTAGGTG
chr15-63553100-63553800(R)	ACAGGAAACAGCTATGACCACAAATATTAATACAAACCC
chr15-63553100-63553800.comp(F)	ATGTAAAACGACGGCCAGTGGGTGTAGGGGA

	GATAGGTTG
chr15-63553100-63553800.comp(R)	ACAGGAAACAGCTATGACCCTCTTAAATCCCTCTCCTCC
SB9prmDistinctiveSeq(F)	CCAACCAACTTTCCAACACC
SB9prmDistintiveSeq(R)	AAATTGTTTAATTTGGGTTA
newCTCF2(63572kb)compF1	CCTATTTTTTCCAATTCAAAAC
newCTCF2(63572kb)compR1	TTTGGTATTTAGTGGGAGTTAA
newNestedCTCF2(63572kb)compF2	TCAAAACACTTTCCAAAACTATC
newNestedCTCF2(63572kb)compR2	GATTTTTATTTTTTTGGGGT
newCTCF2(63572kb) F1	GGGAAAAAAAAAGGAGAGAATAA
newCTCF2(63572kb) R1	CTAAACTAAAAACCCATAAAC
newNestedCTCF2(63572kb) F2	AGTTTGTTTTTTTTAGTTTAAG
newNestedCTCF2(63572kb) R2	CTAAATTTACCAACCTACCTAA
newCTCF3(63603kb) F1	ATGAAGATGGAAAAAGAGATTAG
newCTCF3(63603kb) R1	CCAAAACCAAATATAAAACACCA
newNestedCTCF3(63603kb) F2	TTAATTTAATAGGGAGGGAAGA
newNestedCTCF3(63603kb) R2	CAACACTAAACTAACTTAAAAAC
newCTCF3(63603kb)comp F1	ATAATATTAGATTGGTTTGGAGG
newCTCF3(63603kb)comp R1	ACAAACAACTTACTACTAAA
newNestedCTCF3(63603kb)comp F2	GGGTATTTAAGATTATATTATTG
newNestedCTCF3(63603kb)comp R2	ACTTACTACTAAACAATACT

4.1.8.5 Oligos for ChIP

Primer ID	Sequence
CTCF(-)Ndn qPCR(F)	GGTCCTGCTCTGATCCGAAG
CTCF(-)Ndn qPCR(R)	GGGTCGCTCAGGTCCTTACTT
ctcf(-)neg3-chr12:61mb F	CGGTTTTACTTGTCCCATTTTCC
ctcf(-)neg3-chr12:61mb R	GATCTGAAAATACAGGTGAACTATTGG
ctcf(-)neg4-chr1:174mb F	AAGTGATTTTCAGTGGTCTCAGC
ctcf(-)neg4-chr1:174mb R	TGGAGAAAGACAAATGAGACACC
CTCF(-)Ndn qPCR(F)	GGTCCTGCTCTGATCCGAAG
CTCF(-)Ndn qPCR(R)	GGGTCGCTCAGGTCCTTACTT
ctcf1-L(chr15:63553217) qPCR	GGAGGCAGTAGTGGCCTGTTC
ctcf1-R(chr15:63553217) qPCR	CCTACCTCCATCTTCCCACCTG
ctcf2-L(chr15:63572126) qPCR	TGAAGACCTCATCCAGATGTACCC
ctcf2-R(chr15:63572126) qPCR	TGGTGCACTTGAGGTGGTAAG
ctcf3-L(chr15:63603177) qPCR	TTGCCCAGTTTCTCCATTCC
ctcf3-R(chr15:63603177) qPCR	CCCCCACTCTCCCCTTACAC
ctcf(-)neg1(L)(chr15:63562332) qPCR	TGTAGGAGGTGTGGTCTTCAACAG
ctcf(-)neg1(R)(chr15:63562332) qPCR	AGCAAGGACACCACCTCCCATAG
ctcf(-)neg2(L)(chr15:63345746) qPCR	TTAGTGCTTGTCTGAAATCCTTTCC

ctcf(-)neg2(R)(chr15:63345746) qPCR	ACTGGGAGTGAGAAGTAGTCAAAGC
ctcf(+)cmyc(L)(chr15:61818529) qPCR	CGCCTCGGCTCTTAGCAGAC
ctcf(+)cmyc(R)(chr15:61818529) qPCR	GAATCGCCATCGGCCTTG
ctcf(+q)H19(chr7:149766207) F	CACATAACAGCTTCTATGCCTTCC
ctcf(+q)H19(chr7:149766207) R	GGGGTCCCTTTGGTCACTG
ctcf(+q)Shh(chr5:28,777,841) F	TTGGGTCCACAAGTCTTTTTCC
ctcf(+q)Shh(chr5:28,777,841) R	CATTCTTCCCTGCGTGGAG
ctcf(+q)Bmp7(chr2:172656602) F	AACCTGTGACAAGGCTGGTG
ctcf(+q)Bmp7(chr2:172656602) R	GCATGCTTCTCAAGGATGTGC
chr15:63,479,735-63,480,115 F	AAAGGAGAAGGGGAGTCAGG
chr15:63,479,735-63,480,115 R	GAAAACAGGAGTTGCCCTTG
chr15:63,489,046-63,489,627 F	GGTTCCTCTGGGGACTCTTC
chr15:63,489,046-63,489,627 R	GCCGTTTCAGGAATTAGCAC
chr15:63,489,582-63,490,224 F	ACGGTTTTCACACCCAAAAC
chr15:63,489,582-63,490,224 R	GGGGCTATGTCTCCTCCTTC
chr15:63,448,020-63,448,742 F	AGGCTTTCTGGACAATGGAG
chr15:63,448,020-63,448,742 R	GCAGGAAGAGGCGATAACAC
chr15:63,535,635-63,536,411 F	CAAGGTGGTTCAGGAGAAGC
chr15:63,535,635-63,536,411 R	TCTTGCCTGTCTGGTTTGTG
chr15:63,536,183-63,536,808 F	AATGTGGAGACACGGGAGAG
chr15:63,536,183-63,536,808 R	TTTGGAGCGTTAGAGTGCTG
chr15:63,502,284-63,502,886 F	CTGTCTTCTCTGCCCTGGAG
chr15:63,502,284-63,502,886 R	TGTTGCATAAACAGGGGTTG
chr15:63,315,714-63,316,495 F	AACTCCCACCCCCATAATTC
chr15:63,315,714-63,316,495 R	TATCAGCAGCCAATGCAAAG
Liver chr15:61,930,089-61,931,390 F	TGCTTGGGTTAGGTGAGGAC
Liver chr15:61,930,089-61,931,390 R	GCACCAGAAGCTGGAGGTAG
Liver chr15:61,953,735-61,954,415 F	AGTGCCATCGAGGAAAGTTG
Liver chr15:61,953,735-61,954,415 R	GAGGCCAACAGAGCTAGGTG
Liver chr15:61,989,790-61,990,273 F	GCTAATTGTCTCCCCAATGC
Liver chr15:61,989,790-61,990,273 R	ACAGCCATATGCCCAGAGTC
Liver chr15:61994092-61994479 F	CCTGCTTTCCCATTTCAGAG
Liver chr15:61994092-61994479 R	CAGGGCTCTGTGGGTAACTG
Liver chr15:62,021,531-62,022,050 F	CAGAGAGGTGTGAGGGAAGC
Liver chr15:62,021,531-62,022,050 R	CCTCTCAGAGCTGGACCTTG
Liver chr15:62,052,045-62,052,364 F	GCTGGCTGTGGGGTATAGAG
Liver chr15:62,052,045-62,052,364 R	GTCCACAGGAGGCAGAAGAG

4.1.8.6 Oligos for Transposon Mapping

Primer Name	Sequence	DB no:
SB-L1	CTGGAATTGTGATACAGTGAATTATAAGTG	424
SB-L2	CTTGTGTCATGCACAAAGTAGATGTCC	425
SB-L3	AAGTAGATGTCCTAACTGACTTGC	426
SB-R1	CTTCTGACCCACTGGGAATGTGATG	427
SB-R2	GTGGTGATCCTAACTGACCTAAGAC	428
SB-R3	TCCTAACTGACCTAAGACAGG	429
KmonP-N7-CTCAG	GTACGAGAATCGCTGTCCTNNNNNNNCTCAG	526
KmonP-N7-TCCAG	GTACGAGAATCGCTGTCCTNNNNNNNTCCAG	527
KmonP-N7-TCCTG	GTACGAGAATCGCTGTCCTNNNNNNNTCCTG	529
R/L CommonP	AGTGTATGTAAACTTCTGACCCACTGG	2796
R/L CommonP 2	TGTATGTAAACTTCTGACCCACTGG	2797
KmonP-N8-TTAAG (Rnew)	GTACGAGAATCGCTGTCCTNNNNNNNNTTAAG	2999
KmonP-N8-TAATG (N8)	GTACGAGAATCGCTGTCCTNNNNNNNNTAATG	2795
KmonP	GTACGAGAATCGCTGTCCT	471

4.1.8.7 Oligos for *in situ* Probes

Primer Name	Sequence	DB number
Pvt1 (f)	TACCTCTTGGTCCCTGATGC	1731
Pvt1 (r)	TAGGTTCAACATGGCTGCTG	1732
Gsdmc (f)	CTTGCTGGAAGGATGGAAAG	1922
Gsdmc (r)	CATGTGCAGGAAACTGGAGA	1923
Asap1 (f)	CCAACATCCCACCTGAGACT	1924
Asap1 (r)	GCCTGGCAGTCATAAATGGT	1925
AK015428.1 (f)	GGCCACTGCTCTCTTGAAAC	1926
AK015428.1 (r)	GTCCCAGAGGAACTGCAGAG	1927
Pvt1 new(f)	TTGTCATCTCTCGGGCTACC	1928
Pvt1 new(r)	CACCTTTCCCAGTTTCAGGA	1929
AK089020.1 (f)	TGTTGGAAAAGCTGCACATC	1930
AK089020.1 (r)	CCCACTGTTGGACCTTTTTG	1931
cMyc-F	GGAACTATGACCTCGACTAC	1533
cMyc-R	CTCCACAGACACCACATCAA	1534

4.1.9 Antibodies

4.1.9.1 Primary Antibody

Cleaved Caspase 3 Antibody (Cell Signaling, #9661)

Histone Phospho H3 antibody (Milipore, #06-570, Lot:JBC1863310)

c-Myc Antibody (Epitomics , Cat.No #1472-1, Rabbit Polyclonal)

CTCF ChIP Antibody (Millipore, Upstate #07-729, Lot:2142232)

H3K4me1 Antibody (Abcam, ab8580)

H3K27Ac Antibody (Abcam, ab4729)

DIG Antibody (Roche, 11093274910)

4.1.9.2 Secondary Antibody

Goat anti-rabbit Alexa 594 conjugated (Invitrogen, A11012 Lot:57911A)

4.1.9.3 Beads for ChIP Antibodies

Dynabeads® Protein A, Life Technologies (Cat.No: 10001D, Life technologies)

4.1.10 Animals

Breedings, Plug Checks and Maintenance of the mice are performed by Andrea Schultz, Michaela Wesch, Silke Brohn and Marika Krudwig in the animal facility (LAR) of EMBL.

The animals with transposon insertion have C57/BL6 background. For lacZ staining CD1 outbred females are used.

4.1.11 Software

Software	Company	Notes
MacVector 11.0.2	MecVector Inc.	DNA/RNA sequence Analysis
StepOne™ RT PCR Software v2.0	Applied Biosystems	qRT-PCR analysis
FileMaker Pro9	FileMaker Inc	Mouse, BAC, plasmid, oligo database
Application Suite V3	Leica	Leica MZ16F microscope software
AxioVision c4.8	Carl Zeiss	Zeiss Cell Observer HS software
ZEN	Carl Zeiss	Zeiss Cell Observer HS software
Volocity	Perkin Elmer	PE Ultraview Vox/ERS Software
ImageJ 1.46a	Wayne Rasband, NIH	Image Analysis
Cell Counter Plug In	Ian Levenfus	Image Analysis
BiQ Analyzer	MPI Informatik	Bisulfire sequence analysis
IGV	MIT	Visualization of Genomic Data
MicroWin 2000	Berthold	Luciferase Reads
Nrecon	Skyscan	3D reconstruction
Skyscan v1.3.13	Skyscan	OPT Control Software
Amira™ 5.4.2	FEI	3D Image analysis

4.1.12 Internet Resources

Tool	Address
UCSC genome browser	www.genome.ucsc.edu
ENSEMBL genome browser	www.ensembl.org/index.html
NCBI databases and Tools	www.ncbi.nlm.nih.gov
JAX Lab Mouse DB	www.jax.org
Primer3 primer design	http://bioinfo.ut.ee/primer3-0.4.0/
Pyrat LAR DB	www.pyrat.com/pyrat/cgi-bin/login.py
Bisearch Primer Design	http://bisearch.enzim.hu/
Double Digest	www.thermoscientificbio.com/webtools/doubledigest/
TAD Data Visualization	http://chromosome.sdsc.edu/mouse/
Universal Probe Library for qPCR primers	www.roche.com
FIJI Build-In Functions	http://rsbweb.nih.gov/ij/developer/macro/functions.html
TRACER Database	www.ebi.ac.uk/panda-srv/tracer/index.php
JASPAR, TF Binding Prediction Tool	http://jaspar.genereg.net
EMAGE In Situ Database	www.emouseatlas.org/emage/

4.1.13 ImageJ Macros

ImageJ Macros and Plug-ins	Author
Cell Counter	Ian Levenfus
MultipleAnalysisFolderFrontal.ijm	Veli Vural Uslu
PH3 Analysis.ijm	Veli Vural Uslu
NostrilSizeMacro.ijm	Veli Vural Uslu
ParticleCounterOverAreaV1-1.ijm	Veli Vural Uslu

4.2 Methods

4.2.1 Mouse Maintenance

4.2.1.1. Maintenance and Documentation of Mouse Strains with Insertions

Mice Strains with transposon Insertions are bred with C57/BL6 mice to backcross them into C57/BL6 background. lacZ staining of these lines done on embryos coming from Insertion male and CD1 female breedings. Useful insertion lines have been secured by sperm or embryo freezing.

4.2.1.2 Breedings for Remobilization

Males with heterozygous insertion are bred with carrying Prm16-HSB transposase line. The insertion and transposase double positive males are used for remobilization. In addition, the double positive females are bred with male with heterozygous insertion to obtain males with homozygous insertion with transposase to obtain more remobilization from a single male.

4.2.1.3 Breedings for Chromosomal Engineering

4.2.1.3.1 Deletion and Duplication Lines

For chromosomal engineering there are two insertions, which have loxP site at the same orientation and a recombinase needed. One of the insertions in first bred with HprtCre, which is carried on X chromosome. The mouse line with Insertion and HprtCre is bred with another mouse line with the insertion at the other breakpoint of chromosomal modification. Upon this breeding, I seek to obtain a "transloxer" animal, which carries two insertions and *cre* recombinase. The transloxers are bred with wild type C57/BL6 animals and the progeny is screened for deletion and duplication lines.

4.2.1.3.2 Inversion Lines:

To create an inversion, two insertions, which have loxP sites are the opposite orientation and *cre* recombinase are required. Two insertion lines are bred together to obtain a transloxer line. The transloxer is set up with HprtCre mice in order to obtain a cis-loxer (two insertion in cis configuration) with HprtCre. Due to the presence of HprtCre, the mice will be mosaic of cis and transloxer configuration. In the next generation mosaic animals without HprtCre are kept. These mosaic animals are bred with C57/BL6 line to screen for stable inversion lines.

4.2.2 Molecular Biology Methods

4.2.2.1 DNA Extraction

4.2.2.1.1 Low Purity DNA Extraction

For the samples that will directly be used in straightforward and robust PCR reactions low purity DNA extraction is used. A common stock of Quick Lysis Buffer is prepared [0.05M KCl, 0.01M TRIS pH:8.0, 0.002M MgCl2, 0.01% Gelatin (Sigma G7765), 0.0045 (v/v) NP40 (IPEGAL, Sigma I-3021), 0.0045 (v/v) Tween-20 (Sigma, P2287) in distilled water]. 0.03mg Proteinase K (PK) (Sigma, P4850) is added to 150μl Quick Lysis Buffer and used for one tail or one 11.5 embryonic yolk sac sample. Higher volumes of lysis buffer are used for larger samples. The samples are kept in 55°C oven overnight (O/N) and inactivated at 95°C for 20 minutes. Lysed sample is centrifuged for 1min at 13000rpm. 1μl clean solution from the top is used as a template for PCR reactions.

4.2.2.1.2 High Purity DNA Extraction

Tissues are lysed in SDS Lysis Buffer (100mM TRIS pH:8.50, 5mM EDTA, 0.2% (w/v) SDS and 200mM NaCl) 0.02mg PK is added for each 100μl SDS Lysis Buffer. 150-250μl lysis buffer is used for each sample. The Lysis reaction is done in 55°C oven O/N or 2-4 hours in a shaking (>800rpm) heat block at 56°C. The lysed sample is mixed with equal volume of isopropanol and centrifugated at >13000rpm for 10 mins. The supernatant is discarded and the pellet is washed with 500μl 70% Ethanol.

4.2.2.1.3 DNA extraction by Phenol Chloroform

Tissues are lysed in 150μ-250μl Phenol-Chloroform Lysis Buffer (20mM TRIS, pH:8.0, 5mM EDTA, 400mM NaCl, 1% (w/v) SDS, 400μg/ml PK, PK is always added fresh) overnight at 55°C. Equal volume of Phenol-Chloroform (AppliChem, A088910.0100) is added and the tube is rigorously mixed. It is centrifugated at >13000 rpm for 5 minutes and the the supernant and transferred to a new 1.5ml eppendorf tube. 34μl, 3M NaAc (pH:5.20) and 1.5ml 100% ethanol are added per 300μl supernatant. It is kept in -80°C for 30 minutes and centrifugated at

>13000rpm for 20 minutes at 4°C. The pellet is washed 2 times with 500μl 70% Ethanol. The pellet is resuspended in 150μl dH2O.

4.2.2.2 Total RNA extraction from Embryonic Tissues

Total RNA is isolated by RNeasy Mini Kit (QIAGEN). Embryonic tissues are homogenized by mortar and pestle and the user manual of the kit is followed. RNA is eluted in 50μl RNase free water. PCR amplification of genomic DNA (primer no:790+2432) is used to verify complete elimination of genomic DNA from RNA.

4.2.2.3 cDNA Synthesis

Protoscript MuMLV Kit (NEB) is used for 250-750ng total RNA with random hexamers. The manufacturers manual is used, everything is scaled down to 60% of the volumes suggested by the manufacturer without any loss of efficiency.

4.2.2.4 Embryonic Tissue Lysis for Protein Assays

Whole tissue is homogenized in RIPA Buffer 50mM TRIS, pH:8.0, 150mM NaCl, 0.1% SDS, 0.5% Sodium Deoxycholate, 1% Triton-X, Protease Inhibitors (cOmplete ULTRA Tablets, Roche, Cat NO:05892953001). The lysate is boiled in sample buffer at 95°C for 5 minutes in 1:1 (v/v) mixed 2x Laemmli Buffer (4% SDS, 10% ß-mercaptoethanol, 20% glycerol, 0.004% bromophenol blue, 0.125M TRIS, final pH:6.8)

4.2.2.5 Polymerase Chain Reaction

4.2.2.5.1 Standard Genotyping PCR

For each PCR reaction, 0.4μl in house Taq-polymerase, PCR Buffer (Final Conc. 50mM TRIS, pH:9.5, 15mM (NH4)2SO4, 1.75mM MgCl2 in dH2O), 0.2 mM dNTP mix (peqGOLD dNTP-Set, Peqlab) is used with 0.5μM primer concentration.

5-200ng template gDNA is used per reaction in 200μl wells. The following PCR program is used for Genotyping PCR.

Step:	Temperature	Time (seconds)
Initial Denaturation	95°C	300s
Denaturation	95°C	30s
Annealing*	60°C	30s
Elongation*	72°C	70s
goTo Denaturation	35 times	
Final Elongation	72°C	420s
Hold	16°C	∞
END		

All of the primers for genotyping work at 60°C annealing temperature. Nevertheless a few primer pairs have lower background at higher annealing temperatures. In house Taq polymerase can amplify sequences up to 1300bp. Elongation for 70s is enough for amplicons shorter than 750bp. If the amplicon is between 750 and 1300, elongation time is extended to 100s.

4.2.2.5.2 Long Range PCR

4.2.2.5.2.1 Long Range PCR for genotyping

Roche dNTPack Long Range PCR reaction is used. 20μl final volume, which is 40% of the suggested reaction volume, is used per reaction. 4μl Expand Long Range Buffer with 12.5mM MgCl2, 500μM dNTP mix, 0.3μM F primer, 0.3μM R primer, 3%DMSO, 1.4u Expand Long Range Enzyme Mix is used with 5 to 200ng genomic DNA as template in the following PCR program:

Step:	Temperature	Time (s)
Initial Denaturation	92°C	120s
1st Denaturation	92°C	10s
1st Annealing	60°C	15s
1st Elongation	68°C	60s/kb
goTo 1st Denaturation	10 times	
2nd Denaturation	92°C	10s

2nd Annealing	60°C	15s
2nd Elongation	68°C	60s/kb+20
goTo 2nd Denaturation	25 times	
Final Elongation	68°C	420s
Hold	16°C	
END		

4.2.2.5.2.2 Long Range PCR for cloning

TaKaRa LA Taq® is used to high fidelity amplifications of DNA for cloning. 20μl final volume is used, which is 40% of the suggested reaction volume. For 20μl, 0.2μl LA Taq, 2μl 10X LA PCR Buffer II, 3.2μl dNTPmix (2.5mM each), 0.1μl F primer (100μM stock), 0.1 μl R primer (100μM stock) and 5-200ng template gDNA is used.

Step:	Temperature	Time (seconds)
Initial Denaturation	94°C	60s
Denaturation	94°C	25s
Annealing&elongation	68°C	60s/kb
goTo Denaturation	30 times	
Final Extension	72°C	600s
Hold	16°C	∞
END		

4.2.2.5.2.3 Asymmetric PCR for Mapping:

Step	PCR1	Temp(°C)	Time(s)	PCR2	Temp(°C)	Time(s)
1	Denaturation	95	180	Denaturation	95	240
2	Denaturation	94	20	Denaturation	94	20
3	Annealing	63	60	Annealing	63	45
4	Elongation	72	180	Elongation	72	180
5	goTo	2	10x	Denaturation	94	20
6	Denaturation	94	30	Annealing	63	45
7	Annealing	35	60	Elongation	72	180
8	RAMP	to 72	0.3°C/s	Denaturation	94	20
9	Elongation	72	180	Annealing	53	45
10	Denaturation	94	20	Elongation	72	180
11	Annealing	63	45	goTo	2	9x
12	Elongation	72	180	Denaturation	94	20
13	Denaturation	94	20	Annealing	63	45
14	Annealing	63	45	Elongation	72	180
15	Elongation	72	180	Denaturation	94	20
16	Denaturation	94	20	Annealing	53	45
17	Annealing	44	60	Elongation	72	180
18	Elongation	72	180	goTo	12	8x
19	goTo	10	15x	final extension	72	420
20	final extension	72	420		END	
		END				

4.2.2.5.3 Quantitative Real-Time PCR

Materials for qRT-PCR are provided by GeneCore Facility at EMBL. Primers for the transcript of interest are designed by Universal Probe Library (Roche). SYBR-Green is used for quantitation of amplification. Reactions are done in 100μl 96 well plates in StepOne™ and 200μl 96 well plates in ABI7500 Light Cycler. Total Volume of reaction is 20μl. 0.5μM F primer, 0.5μM R primer, 1μl cDNA template is used with 10μl 2x SYBR Green PCR Master Mix (Applied Biosystems) are used. Gene expression is normalized to *Gusb* for all of the genes, using Microsoft Excel and StepOne™ Real-Time PCR Software v2.0. StepOnePlus™Real-Time PCR Systems (Life Technologies) and AB7500 Real-Time PCR Systems (Applied Biosystems) are used.

Two to three technical replicates, and two to five biological replicates are used for each sample.

4.2.2.6 Genotyping Strategies

4.2.2.6.1 Genotyping for Insertions, Deletions, and Duplications

The presence of the insertion is confirmed by PCR pairs (#3 and #11) or (#6 and #114). After the presence of the insertion is confirmed, primer pairs specific for insertions are used. Primer IDs are given in the table above and the left (“L”) side is used with (#426) and the right (“R”) side is used with (#429). For deletions, if the insertions are in the positive orientation, genomic sequence on the centromeric side of the centromeric breakpoint is amplified by #426 and the genomic sequence on the telomeric side of the telomeric breakpoint is amplified by #429. For duplications, if the insertions are in the positive orientation, genomic sequence on the telomeric side of the centromeric breakpoint is amplified by #429 and the centromeric side of the telomeric breakpoint is amplified by #426.

Homozygosity is determined by negative PCR product. If an insertion is homozygous deletion, the PCR with primers annealing to the “L” and “R” of the genomic region of the insertion does not work. In order to check the homozygosity of the deletion lines, primer pairs from the deleted region is used for PCR and absence of PCR product indicates homozygous deletion.

Insertion Genotyping (Ins)	PCR	PCR	PCR	PCR
PCR #3 and #11	-	+	+	+
Ins L and 426	-	+	+	-
Ins R and 429	-	+	+	-
Ins L and Ins R	+	+	-	+
Genotype	wt	Ins het	Ins hom	Transposed Insertion

DEL/DUP Genotyping	PCR	PCR	PCR	PCR	PCR
PCR #3 and #11	+	+	+	+	+
Cen_Ins L and 426	+	+	-	+	+

Cen_Ins R and 429	-	-	+	+	+
Tel_Ins L and 426	-	-	+	+	+
Tel_Ins R and 429	+	+	-	+	+
Cen_Ins L & Cen_Ins R	+	-	+	+	-
Tel_Ins_L & Tel_Ins_R	+	-	+	+	-
Copy Number	het	hom	het/HOM	het	hom
Genotype	DEL	DEL	DUP	INVers. Cis/Trans	INVers. /Cis

The duplications are genotyped by qRT-PCR for homozygosity by using lacZ primers. DNA is extracted from the tails by High Purity (4.2.1.1.1.2) protocol and from the yolk sac by Phenol Chloroform (4.2.1.1.1.3) protocol. 25ng genomic DNA is used as template and the lacZ amplification is normalized to *Hif1an* genomic region amplification.

4.2.2.6.2 Genotyping for Inversions

The only way to distinguish inverted configurations from cis configuration is using Long Range PCR (4.2.2.5.2.1). The primer, which anneals to the centromeric side of the centromeric insertion, is used in combination with the primer, which anneals to the centromeric side of the telomeric insertion. As a complementary PCR, the primers, which anneals to the telomeric side of the centromeric insertion, is used in combination with the primer, which anneals to the telomeric side of the telomeric insertion. Positive result of these two PCRs indicate inverted configuration.

INV/cis (Long Range)	PCR	PCR	PCR	PCR	PCR
Cen_Ins L and Cen_Ins R	+	+	-	-	+
Tel_Ins_L and Tel_Ins_R	+	+	-	-	+
Tel_Ins_L and Cen_Ins_L	-	-	+	+	+
Cen_Ins R and Tel_Ins R	-	-	+	+	+
Genotype	cis config	cis config	INV contig	INV contig	Mosaic

4.2.2.7 Sanger Sequencing

Sequencing of the PCR products and plasmids has been performed by GATC Biotech. The samples are purified and transferred to 1.5ml eppendorf tubes before shipment with an exception of plasmid sequencing on 96 well plates, in which the bacterial colonies are stabbed into ampicillin agar media plates provided by GATC.

4.2.2.8 Bisulfite Assay

4.2.2.8.1 Bisulfite Conversion

EPITECH Bisulfite Kit (QIAGEN) is used for the bisulfite conversion of 250ng-1μg genomic DNA obtained by high purity purification and the protocol given by the manufacturer is exactly followed except the final elusion is done two times in 20μl nuclease free water.

4.2.2.8.2 PCR Amplification of Bisulfite Converted Regions

Nested Primers are designed manually. In house Taq is used to amplify regions with the modified PCR conditions established by Tugce Aktas in the lab.

1st Step PCR	Temperature	Time (seconds)
1st Denaturation	95°C	240s
1st Annealing*	59-61°C	120s
1st Elongation*	72°C	120s
goTo 1st Denaturation	2 times	
2nd Denaturation	95°C	60s
2nd Annealing*	59-61°C	60s
2nd Elongation*	72°C	120s
goTo 2nd Denaturation	35 times	
Final Elongation	72°C	600s
Hold	16°C	∞
END		

2nd Step PCR:	Temperature	Time (seconds)
Initial Denaturation	95°C	240s
Denaturation	95°C	60s
Annealing*	60°C	90s
Elongation*	72°C	120s
goTo Denaturation	35 times	
Final Elongation	72°C	420s
Hold	16°C	∞
END		

The amplified products are cloned into pGEMT-easy vector and transformed into DH-5α. While colonies are send for sequencing in 96-well format by stabbing them in the ampicillin agar plates. They are read by SP6 primers.

4.2.2.9 BAC targeting

BAC clones are ordered from CHORI (Oakland, California, USA). They are electroporated into EL250 strain or DY380 strain. 50 base pair homology arms are used to introduce the sequence of interest with Kanamycin selection. Kanamycin is flipped out in EL250 strain.

4.2.2.10 RNAseq Experiment

Embryonic faces from 4 del(8-17) homozygous and 4 wildtype littermates are dissected. 72000 drosophila S2 cells are added to each tube as spike. Total RNA is isolated by RNeasy miniprep kit(QIAGEN). RNA quality and quantity is measured by 2100 Bioanalyzer (Agilent Technologies). RNA samples are prepared according to Tru-Seq RNA sample preparation guide (Illumina). 50bp+, single end sequencing is performed by Illumina HiSeq. 8 samples are barcoded and run on the same Flow cell.

4.2.3 Cell Culture Experiments and Lentivirus Production

4.2.3.1. HEK293T cell culture

HEK293T cells are cultured for used for Lentivirus production. The cells are kept in MEF medium [High glucose DMEM (Cat No: 41965, Gibco), 10% heat inactivated Serum Supreme (Cat.No:BW12-492F, Lonza BioWhittaker), PSQ mix [1% Penicillin/Streptavidin (Cat No: 15070-063, Gibco) and 1% L-glutamine (Cat No: 25030-081, Gibco). The cells are kept at 37°C and 5% CO2 incubator and they are passaged two times. The first in first passage the cells are splitted 1:2 and the second one is 1:6 to obtain 12 of 10cm Nunc Nunclon™ dishes (Cat.No: YO-01930-23, Thermo Scientific) at the end. The cells are treated with Trypsin-EDTA (Cat.No: T3924, Sigma) for 5 minutes in 37°C to have them detached from the bottom. They are washed with MEF medium and resuspended in 10ml medium. Cells are counted by haemocytometer and $4x10^6$ cells are plated per dish.

4.2.3.2. Drosophila S2 cell culture

Drosophila S2 cells are cultured to use as spike for RNA-seq experiment. Wild type S2 cells are kindly provided by Vasily Sysoev from Anne Ephrussi Lab. Cells are cultured in Schneider's Drosophila Medium (21720-024, Gibco) and kept at 25°C, 5%CO2 incubator. These S2 cells grow in the suspension and they are splitted only once 1:2 into 10cm Nunc Nunclon dishes.

4.2.3.3 Lentivirus Production

Confluency of HEK293T cells are checked under the microscope. 10 plates are selected with over 80% confluency. The medium is refreshed with MEF medium 2 hours before transfection. The plates are taken from S1 cell culture to S2 cell culture. 105μl of 1μg/μl lentiviral construct is mixed with 36.75μl pMD2.G (1μg/μl stock concentration), 68.5μl pCMVR8.74 (1μg/μl stock concentration), 651μl CaCl2 (2M), and 4390μl dH2O. This mixture is slowly vortexed and meanwhile 5250μl 2x HBS is added onto mix drop-wise. The final mixture is incubated at room temperature under the hood for 20 minutes. 1ml from the mixture is added for each 10cm dish drop-wise and left at 37°C incubator for 24 hours. The second day, 10mM Sodium Butyrate is added to OPTI-MEM medium (Cat.No: 51985034, Life

Technologies) and 10ml of this mix is used to change the medium of the plates. The cells are kept at 37°C incubator for 24 hours with the new medium. On the third day, the supernatant is collected in falcon tubes and kept at 4°C fridge. 10ml OPTI-MEM, Sodium Butyrate mixture is added to each plate. The dishes are kept in 37°C incubator for another 24 hours period. On the last day, the supernatant is collected. All supernatant collection is passed through a Milipore Steriflip 0.45μm filter (Cat.No SE1M003M00, Milipore) to remove the dead cells. The virus is concentrated via Amicon Ultracell-100 Membrane (Cat.No: UFC910096 or Centricon Plus 70 Filter Devices (Cat.No: UFC10008, Milipore). The concentrated virus is aliquoted into 20μl batches and kept in -80°C freezer.

Further Lentiviral injections to mouse embryos have been performed by Katja Langenfeld.

4.2.4 Embryo Preparation

4.2.4.1 LacZ staining

Embryos are collected from stage e11.5 to e13.5, in which plug dates are taken as day zero. e11.5 Embryos are fixed in 4% PFA (pH:7.40) for 30' and e13.5 Embryos are fixed for 40mins. They are washed 2 times with PBS (pH:7.40) on ice and once at room temperature. The staining is performed with standard solution (0.01% Na-Deoxycholate, 0.02% NP40, 2mM MgCl2, 1% Spermidine, 10mM K3Fe3(CN)6, 10mM K4Fe2(CN)6, 2mM X-gal(in DMSO)) at 37°C overnight. The embryos are preserved in 0.4%PFA/PBS solution.

4.2.4.2 *in Situ* Hybridization

4.2.4.2.1 DIG-labeling of probes

Probes are amplified from e11.5 whole embryo cDNA stock by standard genotyping PCR protocol. The amplicons are cloned into pGEM®T easy vector. The ligated plasmid is linearized by restriction digestion and blunted by Klenow fragment. SP6 or T7 RNA polymerases (Roche) are used to produce DIG-labelled complementary RNA probes via in vitro transcription. RNA is cleaned up by RNeasy Kit (QIAGEN) and eluted in 50μl RNase free water. 5μl of the eluted sample is run on

RNase free 1% agarose gel and if the quality is good, the samples are kept in -20°C freezer.

4.2.4.2.2. Sample Preparation for *in situ* hybridization

Embryos dissected at e11.5 are fixed overnight in 4%PFA/PBS. They are washes by PBS-Tween(0.1%) three times, 15 minutes per wash and dehydrated in 25%,50%,75% and 100% Methanol/PBS-T, respectively. They are kept in -20°C freezer until the experiment. At the day of the experiment, they are rehydrated by 75%, 50%, 25% Methanol/PBS-T, respectively, and washed 3 times in PBS-T on ice. After the washing steps, embryos are treated with freshly prepared 6% H2O2/PBST on ice until the color of the embryos turn to white. Then, H202/PBS-T is discarded and the embryos are washed 2 times with PBST on ice and once at room temperature. Embryos are treated with 1μl Proteinase K in 1ml of PBST for 11-12 minutes according to the room temperature. PK digestion is blocked by 5 min glycine/PBS-T (2mg/ml) treatment on ice. Then, the embryos are washed 3 times with PBS-T on ice and fixed with 4%PFA at room temperature for 20 minutes. PFA is washed off by 4 consecutive PBST washes on ice and 1 last PBST wash at room temperature. The embryos are washed with W1 Buffer (5x SSC, 50% deionized Formamide, 1%SDS, 0.1% Tween-20) at 65°C oven for 10 minutes and treated with Hybridization Buffer without the probe (0.5% (w/v) Torula Yeast RNA, 5%(w/v) Heparin in W1) for 2 hours in 65°C oven. 1.5μl and 2.5μl *c-Myc* probe is added to a fresh Hybridization Buffer and the embryos are hybridized overnight at 65°C oven. The following day the embryos are washed 3 times with W1, 3 times with W2 buffer (2x SSC pH4.5, 50% deionized formamide, 0.1% Tween20) and 3 times with W3 buffer (2x SSC pH:4.5, 0.1%Tween20) in 65°C, respectively. After 3 TBST washes, the embryos are kept in blocking solution (2% FCS, 2%BSA in TBST) for 2 hours at room temperature and incubated with anti-DIG-AP antibody (1:3000 in blocking solution) overnight at 4°C. The day after the embryos are washed with extensively in TBST the whole day for 8 times and incubated in TBST at 4°C overnight. On the final day of in situ protocol, the embryos are washed in NTMT solution (100mM TRIS pH:9.5, 100mM NaCl, 1% Tween 20) three times at room temperature and stained with NBT&BCIP (in NTMT) in dark at room temperature. The staining is stopped by PBS and a few drops of 4%PFA when the intensity is informative enough.

4.2.5 Histological Methods:

4.2.4.1 Paraffin Embedding:

Embryos dissected at e11.5 are dissected and kept in 4%PFA overnight and transferred into 70%Ethanol/PBS for overnight incubation. Then, the dehydration is completed the day after in 85%, 95% and 100% Ethanol series, respectively. The embryos are incubated in Ultraclear for 15 minutes at room temperature and then transferred to 65°C oven in 50% Ultraclear-50% paraffin mixture. Next, the embryos are treated with 100% paraffin at 65°C 2 times for 3 hours and overnight respectively. The samples are embedded into paraffin blocks and paraffin sections are obtained at 5μm thickness.

4.2.5.2. Immunostaining of Paraffin Sections

5μm paraffin sections are deparaffinized in ultraclear and rehydrated in 100%, 85%, 70% Ethanol and dH2O series. Antigen retrieval is done in citrate buffer pH:6.0 for PH3 and CC3 antibodies and in alkali Tris-EDTA buffer for c-Myc antibody. Blocking of the sections is done in 10% FCS solution in PBS-Tween(0.2%). Antibody hybridization is performed overnight at 4°C (1:500 for CC3, 1:500 for PH3 and 1:50 for c-Myc antibody). Secondary antibody (Alexa conj. Anti-Rabbit 1:5000 +1x DAPI) is incubated for one hour in room temperature in dark. After 3 times PBS washing the sections are mounted.

4.2.5.3 Haematoxylin-Eosin Staining of Paraffin Sections:

5μm sections are deparaffinized in ultraclear solution and gradually hydrated in 100%, 95% and 80% Ethanol solutions and dH2O, respectively. Slides are treated with Haematoxylin for 5 minutes and Eosin for 30 seconds. Then, they are dehydrated in gradually increasing ethanol solutions and finally treated with ultraclear before mounting.

4.2.5.4 Skeletal Stainings

Skulls from 3 week old and 5 week old animals are prepared in standard 0.3%Alcian blue/0.1% Alizarin Red protocol. Skulls are treated with 1% KOH for 5 days and kept in 100% glycerol at the time of measurements.

4.2.5 Biochemical Methods

4.2.5.1. Western Blot

Pre-cast SDS-PAGE gel system is used (NuPAGE, Invitrogen). Samples are prepared as given in section 4.2.2.4 Embryonic Tissue Lysis for Western Blot.10μl sample is loaded to the wells. Gel is run according to the manufacturers protocol and then blotted to PVDF membrane (Immobiolon, Molopore) with Xcell Lock blot module (Life Sciences) according to manufacturer's protocol. The membrane is treated with Ponceau red to check for the quality of the transfer. The membrane is blocked in 5% milk/PBS-T(0.03%) for 1 hour at room temperature and primary antibody hybridization (1:200 *c-Myc,* 1:1000 FLAG antibodies) is done in blocking buffer. The membrane is washed in PBS-Tween (0.03%) for three times and incubated with rabbit or mouse HRP conjugated secondary antibody for 1 hour. The membrane is treated with ECL (Milipore) and the chemiluminiscence is detected in Kodak X-Ray film.

4.2.5.2 Chromatin Immunoprecipitaiton (ChIP)

ChIP is exclusively performed on e11.5 embryonic faces (46-48 somite stage). Each dissected face is minced into small pieces on a clean petri dish and then collected into a low binding 1.5ml tube in 200μl autoclaved PBS (Cat.No: 022431081, Eppendorf). The cells are 100μl trypsin-EDTA is added to each tube and it is kept at 37°C shaking (1000rpm) heat block for 120 seconds. Then the tubes are centrifugated in in 600g for 5minutes in 4°C refrigerated centrifuge. The Trypsin/PBS mix supernatant is discarded and 300μl dispase (Cat.No 07923, Stem Cell Technologies) is added to the tubes. The tubes are put back on 37°C heat block for 10 mins. The cells are briefly and gently pipetted up and down for facilitate detachment of the cells while incubating them on the heat block. The single cell suspension is centrifugated at 600g for 5min in 4°C refrigerated centrifuge and the supernant is discarded. 973μl PBS and 27μl 37% freshly prepared PFA is added to the cells are resuspended for fixation. The tube is put on a shaker or rotator for 10 minutes are room temperature. 100μl 1M Glycine is added to the tubes in order to quench the reaction on the tubes are transferred on ice for 2 minutes. The cells are again centrifugated with the same conditions, supernant is discarded and cells are snap-frozen in liquid nitrogen until the genotyping is done. Once the genotyping is done, the cells are thawn on ice and resuspended and incubated in 1ml Buffer A(10mM HEPES pH:8.0, 10mM EDTA, 0.5mM EGTA, 0.25% TritonX-100, 10μl PMSF) for disruption of membranes in the cold room on the rotator for 10 minutes. Buffer A is discarded after centrifugation and replaced with 1ml Hypotonic Lysis Buffer B(10mM HEPES pH:8.0, 200mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.01% TritonX-100, 10μl PMSF). The tubes are put back to the rotator in the cold room for 10 minutes. Buffer B is discarded after centrifugation and replaced by 300μl sonication buffer (10mM TRIS pH:8.0, 1mM EDTA, 0.1%SDS, 1x PIC). Sonication is done by Bioruptor® Sonicator Water Bath at 4°C, 24 cycles of 30 seconds pulse, 30seconds rest. After sonication, the tubes are centrifugated at 6000g at 4°C and the supernatant is transferred to another Low binding Tube and equilibrated into RIPA buffer. 1/10 of the sonicated chromatin is aliquoted as input. 0.1μl CTCF Antibody per e11.5 face, 0.3μl Histone antibody is added to the tubes and incubation is done overnight at 4°C on rotating wheel. Antibody is pulled down by DynaBead® Protein G magnetic beads in modified RIPA Buffer (140mM NaCl, 10mM TRIS pH:8.0, 1mM EDTA, 1% TritonX-100, 0.1%SDS, 0.1% Sodium Deoxycholate, 1% PMSF) and the beads are washed in 1ml RIPA for 4 times and once in Tris-EDTA. Beads are transferred to a new low binding tube and kept in 65°C overnight for elution of ChIPed chromatin. 4μl PK is added and 1:600 RNase A stock is added and incubated in 55°C on the heat block. The DNA is eluted by QIAquick PCR purification kit into 50μl dH2O. 0.5-1μl DNA is used for subsequent qPCR reactions.

4.2.5.3 Circularized Chromosome Conformation Capture (4C)

e11.5 face samples are collected and minced into small pieces. Single cell suspension is done by 2 minutes typsin (37°C) and subsequent 10minutes dispase

(37°C) incubation. The cells are fixed in 1%PFA in 10%FCS/PBS and quenched by glycine. The supernant is discarded and snap frozen in liquid nitrogen. The samples are shipped to Wouter de Laat lab in dry ice.

4.2.5.4 FACS Anaylsis for Hematopoietic System

The hematopoietic system related experiments are performed in collaboration with Lisa von Paleske from Andreas Trumpp lab in DKFZ. Here I clarify the personal contributions to the project. I generated the mice strains at EMBL and I performed the genotyping of the mice at EMBL. I did the dissections of these mice according to the instructions given by Lisa von Paleske at EMBL. The bone marrows are extracted in DKFZ with a joint afford of I and Lisa von Paleske. The FACS Experiment and Analysis is done by Lisa von Paleske at DKFZ (including lacZ staining on hematopoietic cells) The cells are sorted in DKFZ and subsequent RT-qPCR and ChIP experiments are done at EMBL by me and Massimo Petretich, respectively. The results are evaluated by a joint afford of Lisa von Paleske, I, Andreas Trumpp and Francois Spitz.

4.3 Image Analysis

4.3.1 Image Analysis for Proliferation Assay

Images are obtained in 10x and 20x magnification in Zeiss CellObserver HS Automated widefield microscope. 4 to 10 images stitched together to obtain a full e11.5 embryonic face image. DAPI filter is used to visualize the nuclei (D). GFP filter is used to visualize the background autofluorescence(G) and Ds-red filter is used to capture the signal(R). Image Analysis is done by FIJI. An automated macro is written to minimize subjective errors. GFP filter image is subtracted from Ds-red filtered image (S). S is subjected to background rolling, gaussian blur, and Renyi Entropy dark threshold, respectively. Then, particle analysis is done with size restriction (pS). pS is overlaid on top of D and pS signal, which is too elongated and which does not overlap with DAPI signal was eliminated manually(pSc). To find the right parameters for particle count in 10x and 20x widefield images, 63x magnification images of the 2 sections that cover the whole face are obtained by Spinning Disk Microscope (PE Ultraview VoX). Over 80 image/section are stitched together by

Volocity (Perkin Elmer). Mitosis-phase particles are counted manually and the parameters of automated 10x and 20x image analysis is adjusted accordingly. Proliferation unit is taken as pS-c divided by the total number of nuclei(nD). In 10x and 20x magnification particle count cannot be done solely based on DAPI staining. Current FIJI Plug-Ins for DAPI signal counting were computationally inefficient for large samples like face. Therefore, I used DAPI area(aD) as a measure of cell number. In order to verify the linear relationship between DAPI area and cell number, FIJI "Cell counter" plug-in by Ian Levenfus is used. The accuracy of the plug-in is assigned by comparing the manually counted DAPI signal versus DAPI signal counted by "Cell counter". Linear correlation between DAPI area and cell count done by "Cell counter" plug in is >99.5% in the range of 20 to 8000 cells. Therefore, proliferation rate is measured as (pSc)/aD.

4.3.2. Image Analysis for Morphological Measurements in Embryos

24 landmarks are used to determine the facial morphology. The highlighted measurements are emphasized in the text.

Measurement ID	Symbol	Landmarks & calculations
Dorsal Ventral Length	DV	3-4
Total Side-to-Side Length	TStS	1-2
Left Diagonal Length	Ld	1-3
Right Diagonal Length	Rd	2-3
Left Lateral Side Mesenchyme Length	LLsM	1-5
Left Lateral Diagonal Length	LLd	1-9
Left Medial Diagonal Length	Lmd	3-12
Left Medial Frontal Mesenchyme Length	LmfM	3-13
Right Lateral Side Mesencyme Length	RLsM	2-18
Right Lateral Diagonal Length	RLd	2-22
Right Medial Diagonal Length	Rmd	3-19
Right Medial Frontal Mesenchyme Length	RmfM	3-23

Medial Total Ventral Mesenchyme Length	MtVM	12-19
Medial Total Dorsal Mesenchyme Length	MtDM	8-15
Epithelial Thickness Frontal Left	ETfL	13-14
Epithelial Thickness Frontal Right	ETfR	23-24
Average Frontal Epithelial Thickness	AeET	(ETfL+ETfR)/2
Epithelial Thickness Lateral Left	ETlL	5-6
Epithelial Thickness Lateral Right	ETlR	17-18
Average Lateral Epithelial Thickness	AlET	(ETlL+ETlR)/2
Epithelial Thickness Medial Left	ETmL	11-12
Epithelial Thickness Medial Right	ETmR	19-20
Average Medial Epithelial Thickness	AmET	(ETmL+ETmR)/2
Average Epithelial Thickness	AET	(AeET+AlET+AmET)/3
Frontal Distance	fD	13-23
Medial Total Mesenchyme	MtM	(MtCM+MtDV)/2
Ratio of Medial Mesenchyme to Total Length	rMM	(MtV/TStS)
Left Lateral Mesencyme Average	LLMa	(LLd+Lmd)/2
Ratio of Left Mesencyhme to Total Length	rLM	LLMa/TStS
Right Lateral Mesencyme Average	rLMa	(RLd+Rmd)/2
Lateral total Side Length	LtS	(LLsM+RLsM)
Ratio of Medial Mesenchyme to Lateral Mesenchyme	rML	MtM/LtS
Ratio of Left diagonal Medial to Lateral	rLd	Lmd/LLd
Ratio of Right diagonal Medial to Lateral	rRd	Rmd/RLd

Many of the facial features changes from anterior to posterior part of the face. In order to determine the average anterior/posterior plane, dorsal-ventral (DV) face length is used. The DV length is the highest in the anterior part and it is zero in the anterior part as the nasal processes are still separate. Measurements are done only for the sections, in which DV is greater than zero. The comparison between homozygous del(8a-17a) DV length and wildtype littermates DV length show that there is no significant change in the anterior/posterior plane between these two genotypes. (p>0.40)

Face Measurements to calculate lateral relies on the sum of right and the left side lateral mesenchyme. Due to possible tilt in the cutting plane from one side to another also leads to differences between left lateral mesenchyme and right lateral mesenchyme. In order to work in the average plane, right and left mesenchyme lengths are summed up. Epithelial thickness of nostrils is measured by using 12 landmarks and 6 different lengths from 3 different directions.

4.3.3. Image Analysis for Adult Skeletal Stainings

18 landmarks are used for skeletal images:

Landmarks	Code
Distal tip of the median line	2
posterior tip of the nasal bone	3
right anterolateral corner of the frontal bone	4
left anterolateral corner of the frontal bone	5
right interorbital	6
left interorbital	7
anterolateral tip of right parietal bone	8
anterolateral tip of left parietal bone	9
posterior top of frontal bone	10
anterior tip of interparietal bone	11
right lateral-posterior zygomatic arch	12
left lateral-posterior zygomatic arch	13
right incisor Tooth	14
left incisor Tooth	15
right angular process	16
left angular process	17
right cartilage of hyoid bone	18
left cardilage of hyoid bone	19
Measurements:	

Nasal Bone Length	2_3
Frontal Bone Length	3_10
Parietal Bone Length	10_11
interorbital distance	6_7
interorbital distance distal	4_5
cheek to cheek	12_13
right cheek bone	4_12
left cheek bone	5_13
mandibular bone right	14_16
mandibular bone left	15_17
mandibular side to side	16_17
hyoid bone	18_19

The images are taken by Leica M16F microscope with 0.71x magnification in 100% glycerol. The coordinates of given landmarks are extracted by using FIJI. The calculations are done in Microsoft™ Excel. Wildtype and del(8a-17a) population is compared by student t test.

4.3.4 Image Analysis for Optical Projection Tomography of Embryos

OPT Images are obtained in Bioptonics 3001M OPT with 1024x1024 resolution with 5µm/pixel magnification, 0.45° rotation. 800 images are collected and reconstructed by NRecon® after manual fine-tuning of the registration. The Images are analyzed in Imaris and Amira v5.4.2 by Isosurface. Images are cropped 50 pixels from each direction. Isosurface is applied with a manually determined downsampling paratemeter (between 4 and 6). Threshold is also manually determined in a way that the facial landmarks are easily distinguishable. All of the measurements are done by 3D ruler in Amira.

5. RESULTS

5.1 Regulatory Landscapes in the *c-Myc/Pvt1* Flanking Locus

5.1.1 Mapping Regulatory Landscape with GROMIT

I investigated the regulatory activities in the *c-Myc/Pvt-1* flanking locus. For this purpose I used the Genome Regulatory Organization Mapping by Inserted Transposons (GROMIT) approach developed in the lab. GROMIT is based on the use of a regulatory sensor, constituted by a minimal promoter and a lacZ reporter gene (Figure 6A). The minimal promoter is composed by the 50bp sequence proximal to the human ß-globin transcriptional start site (Yee SP and Rigby PW, 1993). It does not have a strong activity by itself when inserted in the mouse genome. However, lacZ reporter gene is expressed in the presence of an endogenous regulatory input, corresponding to the activity of nearby or remote but long-range acting *cis*-regulatory elements (Ruf S et al*.* 2011). The use of lacZ as a reporter, which has no background in mouse embryos, and which has high sensitivity and spatial resolution, allows detection of a range of strong and weak regulatory inputs. Therefore, lacZ staining on embryos reveals the tissues in which endogenous regulatory activity is detected at the site of insertion.

The regulatory sensor in a locus reveals the regulatory input only at the site of insertion. In order to reveal the structure and extent of regulatory landscapes throughout the locus, multiple insertions of the regulatory sensor are required. To generate, in a simple and efficient manner, such a series of insertions in a locus of interest, the regulatory sensor was cloned into a *Sleeping Beauty* transposon (Figure 6A). This transposon allows distribution of this integrated reporter, in the mouse genome, without the time consuming procedures of mouse knock-in procedures. *Sleeping Beauty* is a cut and paste transposon (Ivics Z *et al* 1997), meaning that the number of copies of the transposons does not amplify upon transposition (remobilization). Besides, in contrast to other systems like piggyBac (Li MA *et al* 2013) and retroviral insertions (Mitchell RS *et al* 2004), *Sleeping Beauty* showed no bias with regards to the TSS, gene bodies, or intergenic sequences (Horie K *et al* 2003, Yant SR *et al* 2005; Ruf *S et al 2011*). However, *Sleeping Beauty* tends to reintegrate close to its starting site (Luo G *et al* 1998). This property, referred to as "local hopping" (Keng VW *et al* 2005) implies that upon remobilization of an insertion in a locus of interest, new insertions will be enriched at this locus. Moreover, *Sleeping Beauty* has been shown to work efficiently in mammals (Luo G *et al* 1998).

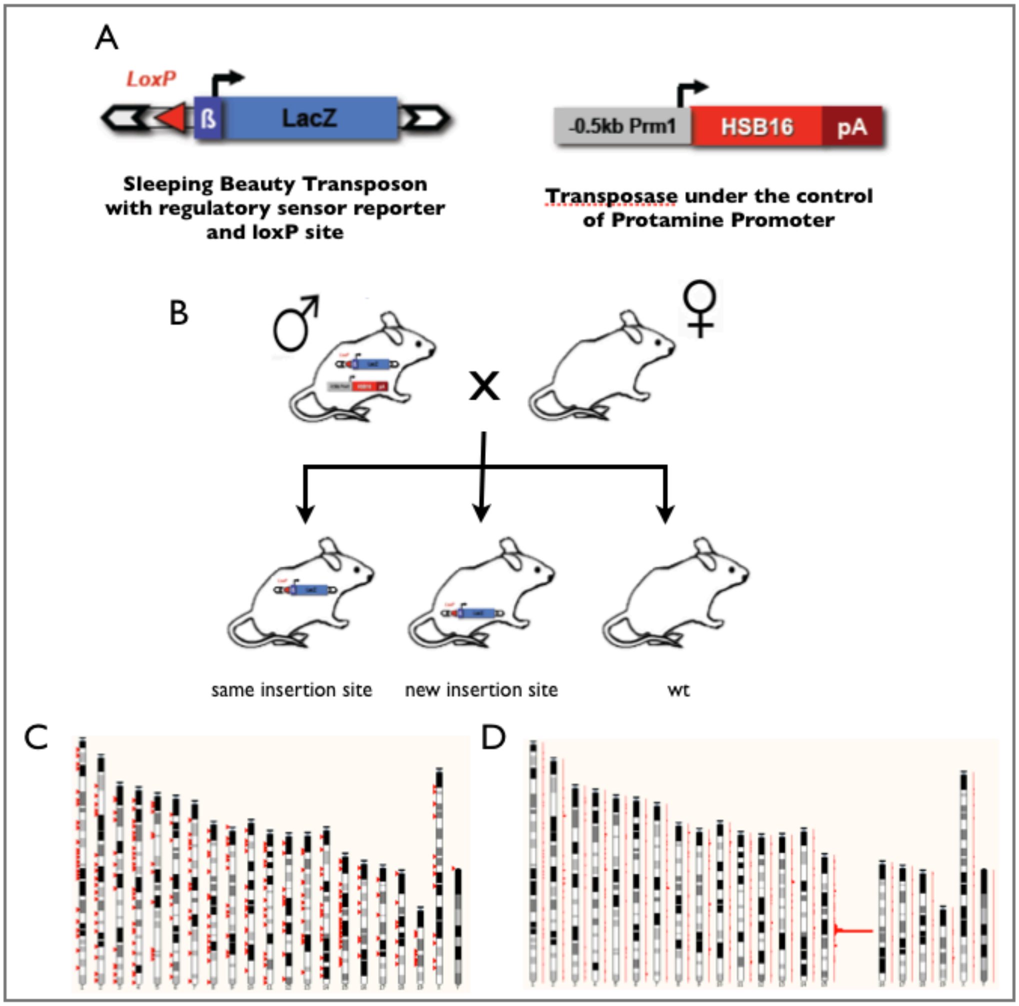

A
LoxP
ß
LacZ
Sleeping Beauty Transposon with regulatory sensor reporter and loxP site
-0.5kb Prm1
HSB16
pA
Transposase under the control of Protamine Promoter
B
♂
x
♀
same insertion site
new insertion site
wt
C
D

Figure6| Transposition by Sleeping Beauty System: A) The regulatory sensor cassette (left) and the transposase construct (right) are given. B) Males positive for transposon and transposase are bred with wildtype females. The transposition event is detected when the progeny is positive for Sleeping Beauty Specific PCR but negative for insertion site specific PCR. C) Transposition from *c-Myc/Pvt1* flanking locus distributed the transposon to all the chromosomes in the genome. D) However, quantitatively eight percent of the mapped transposition events, transposon jumped locally in this locus.

The system we use in the lab restricts the expression of a hyper-active form of the transposase (Baus J *et al* 2005) to the second meiosis stage in male spermatogenesis, by the use of a transposase under the control of protamine promoter (Ruf S *et al* 2011) (Figure 6A). New insertions are obtained simply by mating males carrying a transposon insertion in the locus of interest and the transposase transgene with wildtype females (Figure 6B). As the transposase is not expressed/active in somatic tissues, F1 animals will carry a stable, non-mosaic insertion, allowing either direct analysis, or efficient establishment of a stable line (Figure 6C-D).

5.1.2 Starting Insertion

An insertion (179039) obtained by Sandra Ruf as a remobilization from a concatemer of transposons on chromosome 9, was mapped to chromosome 15. LacZ staining on embryos from the 179039 line showed very prominent expression in the medial face, proximal limb and somites. It is located in a large gene desert, with few neighboring genes. *c-Myc* is located is 1.7 megabase (Mb) far from the insertion 179039 (Figure 7A).

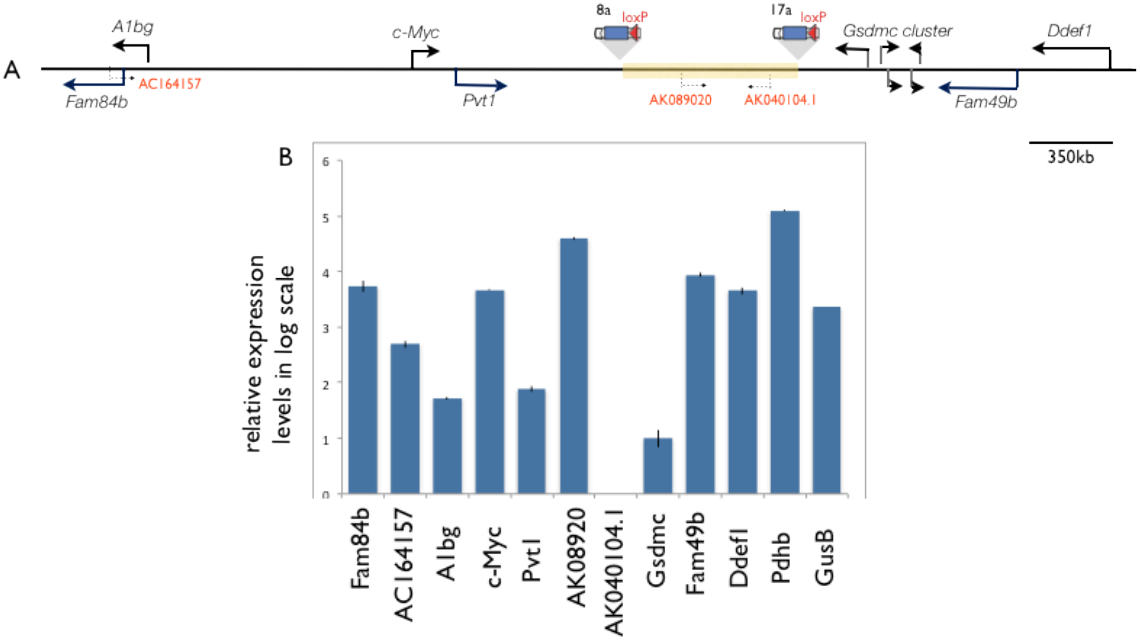


Figure7|the *c-Myc/Pvt1* Flanking Locus: A) The genes and annotated transcripts in the *c-Myc/Pvt1* flanking locus. The Genes above the line are protein-coding, whereas the genes below the line are non-coding genes/transcripts. The red colored transcripts are annotated only in a subset of genome browsers. The transposon constructs indicates the relative position of two insertions: 8a and 17a. B) The expression levels of the genes and transcripts are given for the e11.5 face of the C57/BL6 wt embryos. The expression levels are normalized to *GusB* and relative values are indicated with respect to the expression level of *Gsdmc* in logarithmic scale.

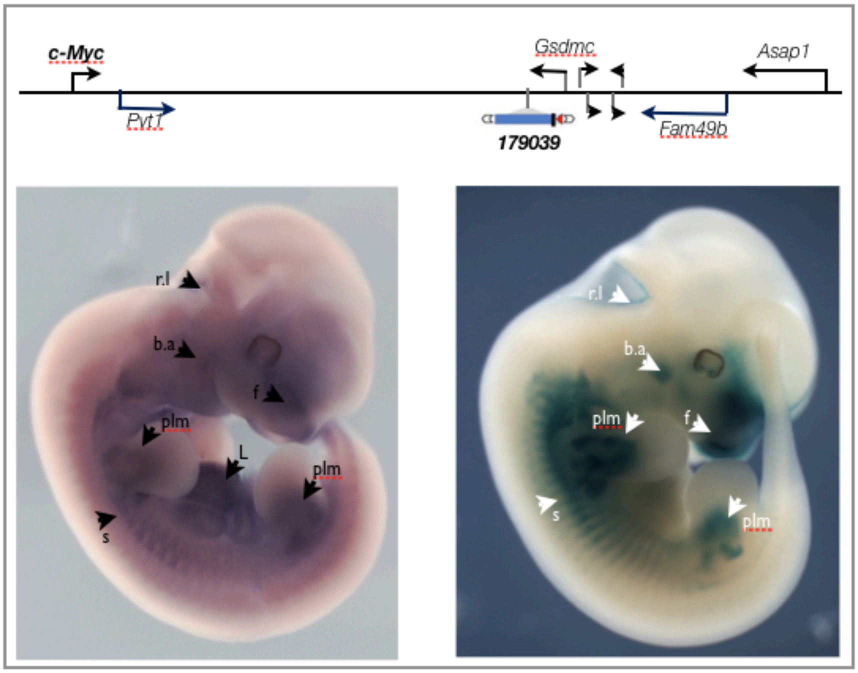


Figure8|Starting Insertion: 179039 is the first insertion in the locus. The position of the insertion is shown in the sketch above. On the right side, lacZ staining of 179039 is shown. On the left side, whole mount *in situ* hybridization for the *c-Myc* is shown on the left rl:rhombic lip, ba:branchial arch, plm:proximal limb mesoderm, s:somites,L:liver, f:face

In e11.5 embryos, *c-Myc* antisense probes showed broad overall staining *in situ* hybridization experiments (as a control, I did not detect any signal with a *c-Myc* sense probe), showing that *c-Myc* is expressed widely. However, *c-Myc in situ* signal was stronger in few tissues including branchial arches, proximal limb mesoderm, somites, liver and face (Figure 8). Interestingly, these stronger domains in the face, somites and limb mesenchyme resembled strikingly with the LacZ pattern shown by 179039. There were, however, differences between 179039 and *c-Myc*: 179039 did not show LacZ staining in the liver, contrarily to *c-Myc in situ* and the reporter gene was also expressed in the rhombic lip, where no specific enrichment of *c-Myc* was detected. The other surrounding genes, *Pvt1* and *Gsdmc* have low level of expression: *in situ* hybridization with probes for these genes did not reveal any tissue specific pattern and by RT-qPCR analysis showed overall low expression values (Figure 7B). The overlapping expressions suggested that despite their distance, *c-Myc* and 179039 may respond to the same regulatory input. However, because of the widespread expression of *c-Myc,* direct evidence was needed. Therefore, to better reveal the regulatory landscapes for defining new long-range enhancers and shedding light on its orthologous 8q24 locus in humans, I set up a large remobilization effort from the starting insertion 179039 (Figure 6C-D).

5.1.3 Transposition in the *c-Myc/Pvt1* Flanking Locus

Remobilization from 179039 (thereafter named as 17a insertion) has been very efficient. From this line, 234 new insertions were mapped to the genome corresponding to a remobilization (transposon reintegration) of 41% (Figure 9). 31 of these 234 insertions were obtained within the three megabase interval corresponding to the *c-Myc/Pvt-1* flanking locus. Eight of these insertions are used as a starting site for further remobilization to get a better coverage across the region. For example, remobilization from 184347 (thereafter named as 8a insertion as it is 8 hundred kilobase far from *c-Myc* promoter) gave similar general and local transposition efficiency with the 179039 line (Figure 9). In total, 53 insertions were obtained in the *c-Myc/Pvt-1* flanking locus and 20 additional insertions have been mapped to 1mb neighborhood of this locus (local transposition). Most of the insertions in *c-Myc/Pvt-1* flanking locus were obtained in the 1.8mb long telomeric side of the *c-Myc* gene. This corresponds to an average density of one insertion every 40kb, but with a large spread as the local hopping has a bell curve distribution: the largest gap in between two insertions is 250kb and the smallest one is 50bp. Distribution of 53 insertions showed 30-50kb islands, where multiple insertions are obtained from different start sites, suggesting that some regions may be preferentially targeted by the transposon. All of the lacZ data is stored in TRACER database (Chen C *et al* 2013).

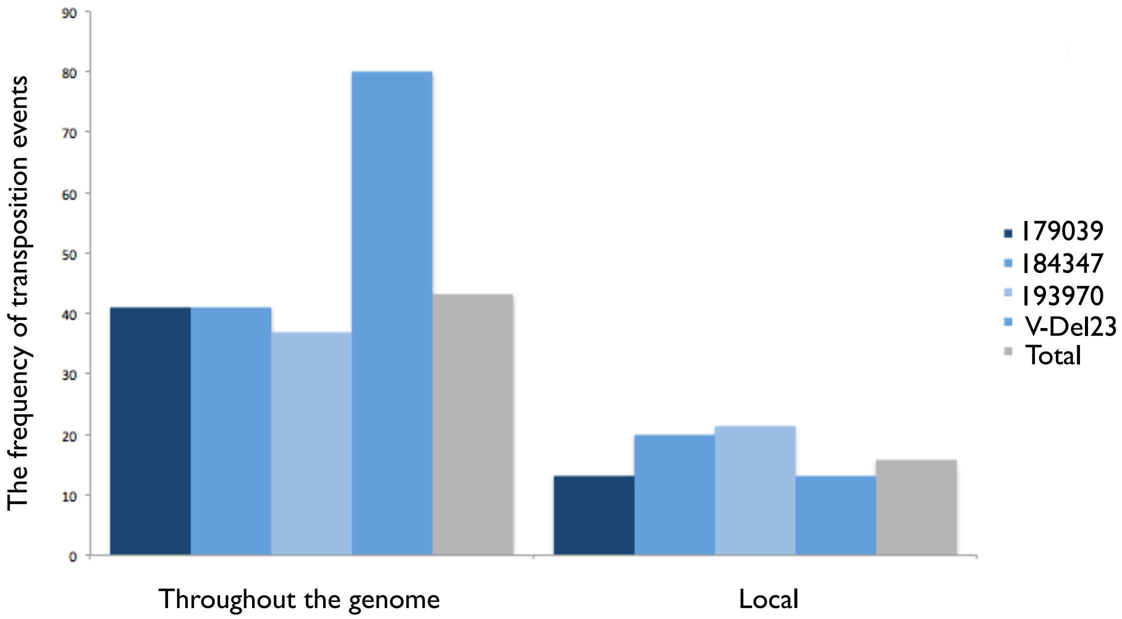


Figure9| Transposition from *c-Myc/Pvt1* flanking region: The percent efficiency of all transposition events for the given start sites is shown on the left panel. The percent efficiency of local transposition event, where the transposon jumped out from the original site to somewhere else in the *c-Myc/Pvt1* flanking locus is shown on the right panel.

5.1.4 Regulatory Landscapes in the *c-Myc/Pvt1* Flanking Locus in e11.5

I analyzed the expression of the regulatory sensor at 50 insertion sites spread along the locus and its 1Mb neighborhood by carrying out lacZ staining on mouse embryos, which are obtained from a cross between males heterozygous for insertions and wildtype females. This was done for most at e11.5, even though additional stages were analyzed for several insertions. Here, I will mostly discuss the data obtained at e11.5 (Figure 10).

LacZ staining provides spatial regulatory information on the whole-mount embryos. I also used the intensity of the staining, which was very reproducible for each given time (across experiments, litters) as a proxy for accessibility of the insertion site. I defined 5 different categories for the intensity of lacZ expression: Strong (4), average (3), weak (2), faint (sometimes hardly visible on photos of whole-mount embryos) (1), and not expressed at all(0). Altogether, the analysis of the 50 different "viewpoints" revealed a subdivision in large domains of shared expression, corresponding to the "regulatory landscapes" described initially at the

HoxD locus (Spitz F *et al* 2003). The density and the number of insertions allowed me to get a more detailed view of the organization and properties of these landscapes (Figure 10).

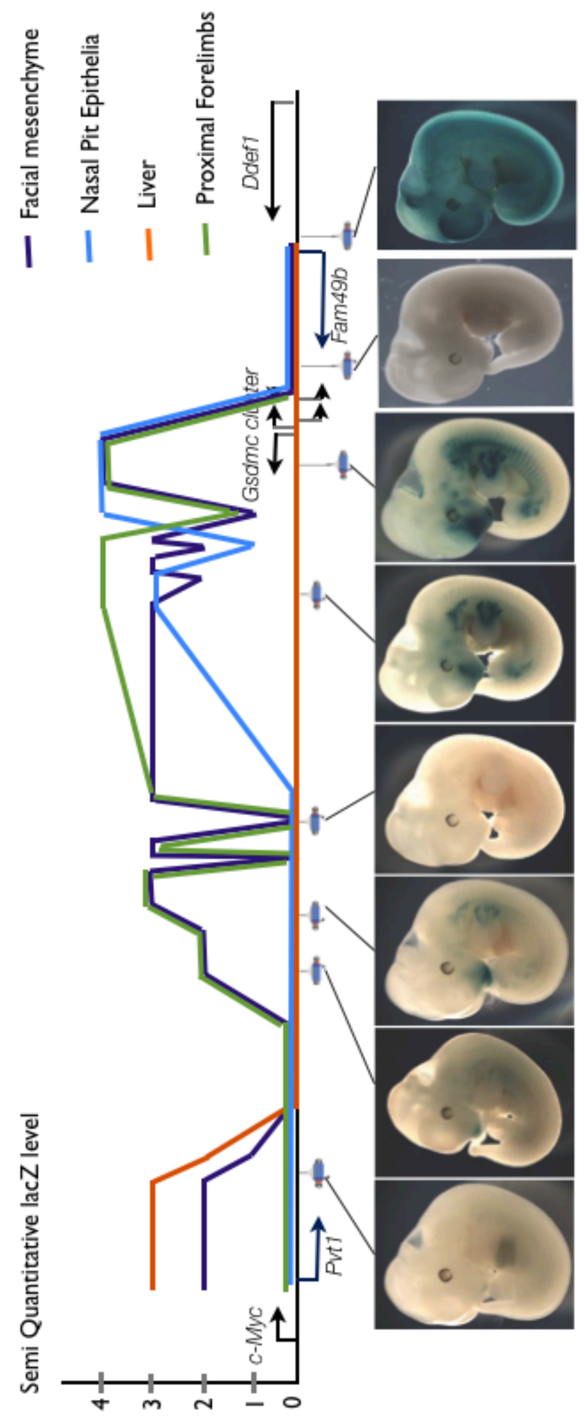


Figure10|Regulatory Landscapes in the *c-Myc/Pvt1* flanking locus: All the fifty transposon insertions in this locus are semi-quantitatively scored according to their lacZ expression in the facial mesenchyme, nasal pit epithelia, liver, and proximal limb mesoderm. Then, these scores are indicated in dots in the y axis at their position in the locus, which is in x axis. These dots are linked by a straight line to reveal the shape of the regulatory landscape. Several representative lacZ staining on e11.5 embryos are shown in the lower panel. These landscapes show that in the *c-Myc/Pvt1* flanking locus there are overlapping but distinct regulatory landscapes and there is a different regulatory landscape the beyond the *Fam49b* gene.

5.1.4.1 General Features of Regulatory Landscapes in the *c-Myc/Pvt1* Flanking Locus

I found that very often, adjacent insertions had the very same expression patterns, revealing some organization in the distribution of the regulatory potentials present in the region. Noteworthy, the orientation of the regulatory sensor (with respect to plus or minus strand) appeared to have no effect on the expression pattern detected. For example, we obtained from the insertions 179039 and 184347, extreme local hopping events with the transposon jumping back to its starting position but in opposite orientation: in these two cases, I observed identical pattern and intensity of LacZ for the two orientations (Figure 11).

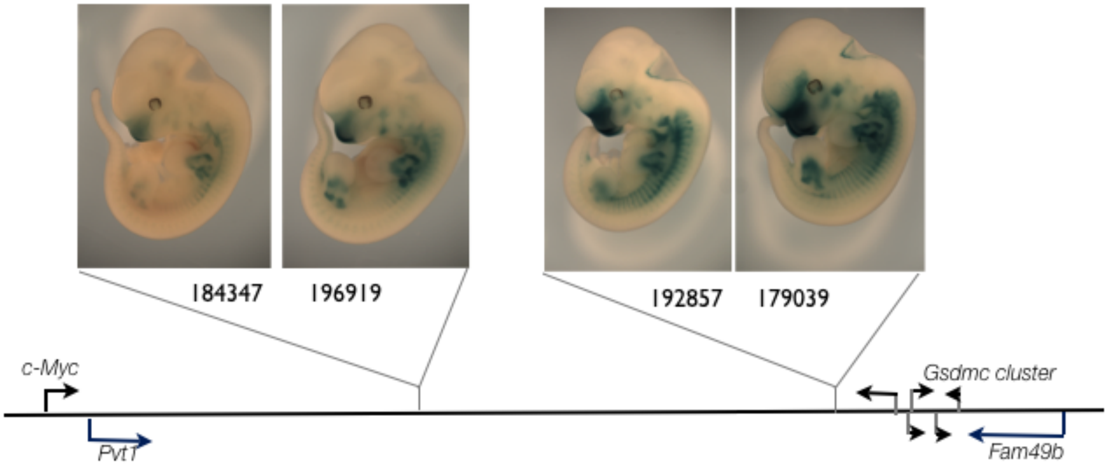


Figure11|The Effect of Orientation on Regulatory Input: lacZ staining of the embryo pairs where the insertion sites are the same but the orientations are opposite show that regulatory input is independent of the direction of transcription

At a large scale, lacZ reporter expression patterns, which are obtained from 30 different insertions strictly in the telomeric gene desert flanking *c-Myc/Pvt-1,* reveal three regulatory landscapes corresponding to activity in the embryonic face, proximal limb mesoderm (PLM), and somites*.* On the centromeric side of *c-Myc* gene, three insertions showed either no expression or expression patterns were different from the domains observed on the telomeric side (Figure 10). Therefore, embryonic face, PLM and somite landscapes extend from the *Pvt1* gene, centromerically to the *Gsdmc* cluster, telomerically. The *Pvt1* region corresponds to a different landscape (liver), which partially overlaps with the three other ones, but does not extent up far in the gene desert. The overlap between these landscapes seem to correspond to a progressive / reciprocal decrease in the strength of regulatory input. The telomeric end of the landscapes is much more sharply defined: around the *Gsdmc* cluster, just adjacent to the insertions showing very strong expression, I delineated a region of 160kb (with two insertions), where no positive regulatory input could be detected at e11.5. Further telomeric to this region, starting from centromeric end of *Fam49b*, I observed a new landscape with three insertions characterized by widespread LacZ expression, with specifically strong expression in the tail bud ectoderm, forebrain, and midbrain-hindbrain boundary (Figure10).

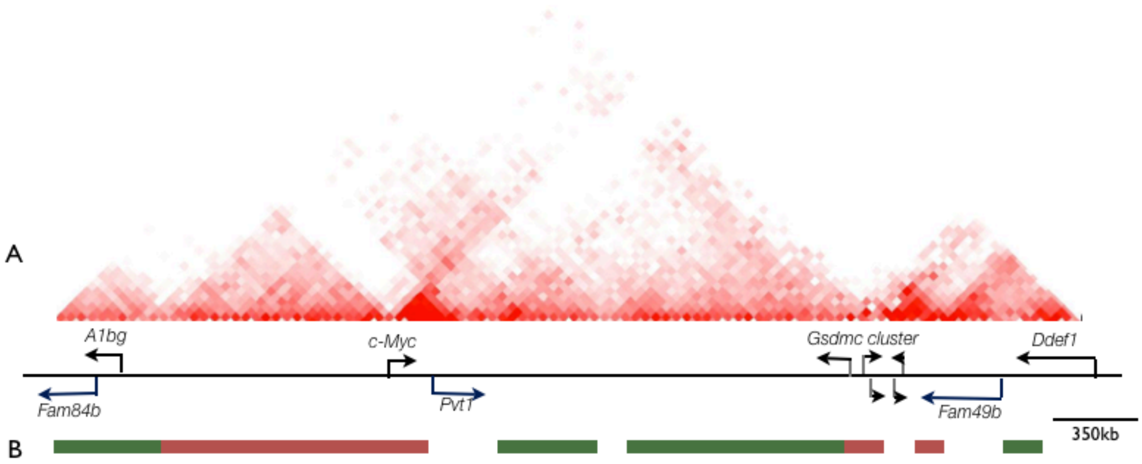


Figure12|3D organization of the *c-Myc/Pvt1* flanking locus: A) Topologically Asscociated Domains are obtained from Dixon JR *et al* 2012. The strength of the red correlates with the number of the readcounts of ligation products. B) Hidden Markov Model of TAD structures deliniating the well-structured, less well structured and the TAD boundaries.

HiC experiments in mouse cells and tissues have identified several topologically associating domains (TADs) that subdivided this region (Dixon JR *et al* 2012) (Figure 12). *c-Myc* is the border of two TADs. The telomeric one, is 1.8-megabase long and ends up in *Gsdmc* cluster. This cluster, which is a less well-structured (or maybe its repetitive nature impaired a clear definition of its structure) region, contains the non-expression interval and separates tbe *c-Myc* TAD from the *Fam49b* landscape, which corresponds to another clear TAD on the HiC maps.

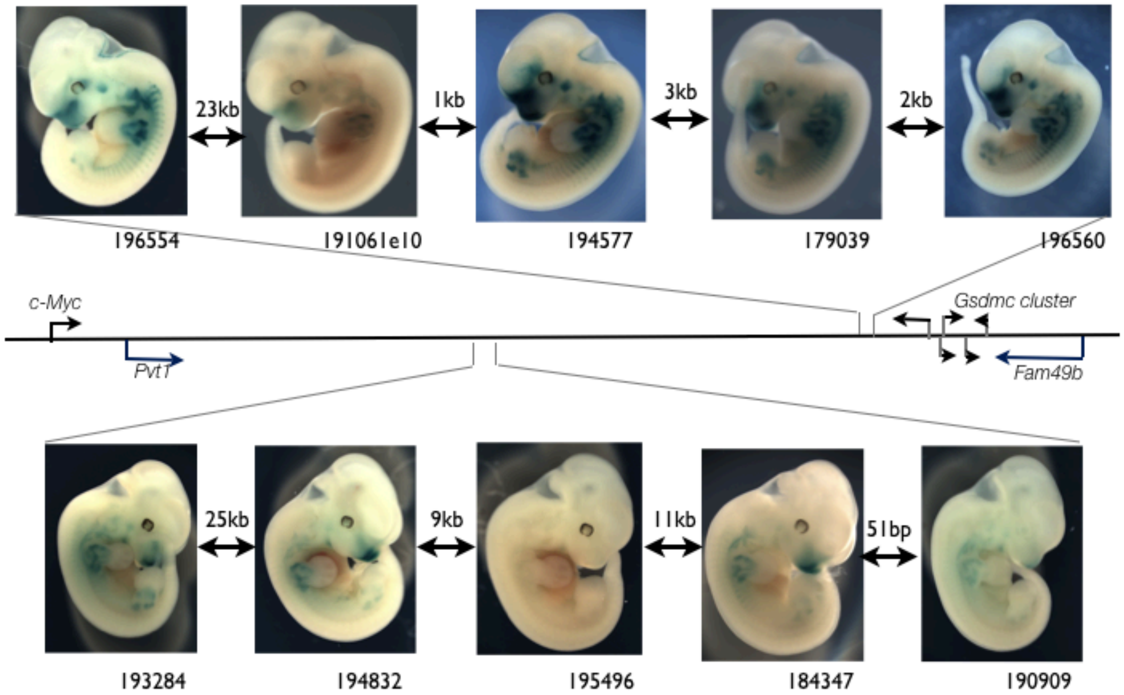


Figure13|Change in the regulatory input in short range: lacZ staining of neighboring regulatory sensors around 179039 insertion site (top panel) and around 184347 insertion site (bottom panel)

The comparison of the regulatory domains with the structural TADs showed that the region can be subdivided in structural TADs, which the *c-Myc/Pvt1* flanking locus regulatory domains overlap and share the same boundary regions. However, the topologically associated domains and regulatory domains (RDs) are not equivalent: The telomeric TAD contains different RDs with varying stretches (Figure 10-12). For example, the liver regulatory domain is covering the centromeric region around *Pvt1,* whereas the other activities are concentrated on the more central and telomeric parts of the TAD, with the expression of LacZ reported in the limb and in the face being the strongest at the telomeric end of the TAD. Together with the *c-Myc/Pvt1* flanking locus, by employing a large number of regulatory sensor insertions, I contributed to studies that show a large overlap between RDs and TADs in many other loci (Symmons O *et al* submitted). Importantly also, within a given regulatory domain (RD), the regulatory inputs are not distributed homogenously.

Expression of the reporter genes fluctuates in intensity all along the locus (Figure10). Usually, insertions close to each other displayed identical or at least similar expression pattern, yet intensity of the reporter activity can vary also on few regions between insertions 1kb far from each other (even shorter) and transition from expressed to non-expressed can occur across 10kb (Figure13). Such discontinuities are not specific to this region and have been observed in other loci. (Symmons O *et al* submitted).

5.2 Distant Regulatory Regions in the *c-Myc/Pvt1* Flanking Locus

5.2.1 Embryonic Face-Specific Regulatory Elements

The genome-wide association studies (GWAS) have shown the presence of a common risk allele for cleft lip and palate (CLP) on 8q24, which is orthologous to the *c-Myc/Pvt1* flanking locus (Figure 5, Figure 14A). The presence of regulatory activity in the embryonic face for this region was therefore interesting. In humans, the medial nasal process, together with maxillary processes gives rise to lip and palatal shelf development between 4th and 6th weeks of pregnancy. Defects in the growth and fusions of the different processes taking place at this stage will give rise to cleft lip and palate (CLP) (reviewed in Dixon MJ *et al* 2011). In the mouse, the corresponding morphogenetic events take place between e10 and e15. At e11.5 the medial facial mesenchyme (MFM), which is the region restricted by nostrils, the lateral face mesenchyme (LFM), which lies on the side of the nostrils, and maxillary processes fuse to form the palatal shelves in mouse (Figure 14B-C-D-E). Most of the insertions throughout the *c-Myc/Pvt1* flanking locus captured regulatory activity in MFM, notably at its posterior and frontal edges, where active fusion of the nasal processes takes place. (Figure_14D). In addition to the MFM staining, the insertions close to the telomeric boundary of the locus show expression in the nasal pit epithelial (NE), particularly on the anterior part of the face (Figure 14E).

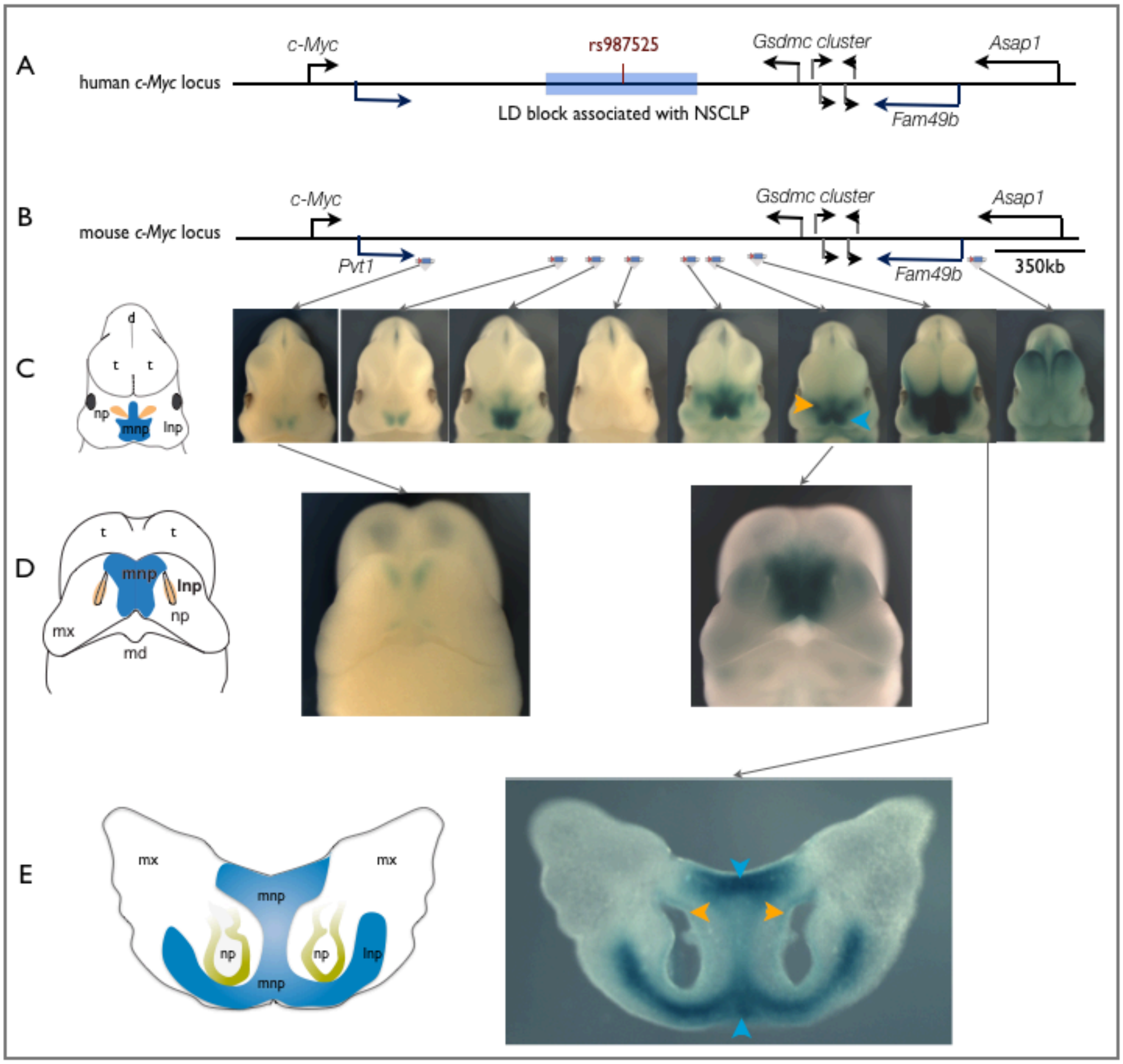


Figure14|Embryonic Face Regulatory Domain: A) The SNP rs987525 (and the Linkage Disequilibrium Block) is associated with NSCLP. B-C) lacZ staining of the insertions in mouse orthologous region at e11.5 reveals the regulatory landscape. D) The lacZ staining of 15a insertion in the face is more precisely in the medial nasal process and the nasal pit. E) The vibrotome sections show that the expression in the nasal processes is mesenchymal and the nasal pit is epithelial. Blue arrows indicate the medial face mesenchyme and the yellow marks indicate the nasal pit epithelia Mnp:medial nasal process, np: nasal pit, mx: maxillary arch, lnp: lateral nasal process, md: mandibular, t:telencephalon, d:diencephalon

In concordance with this developmental trajectory, lacZ staining for 17a position persists in the face at e12.5. In the development course of palatal shelf at e13.5 and e14.5 lacZ staining appears strongly in the palatal shelf and the facial mesenchyme that give rise to skeletal structures and muscles. In brief, regulatory activity detected by the insertions in the mouse *c-Myc/Pvt1* flanking region, which is orthologous to the CLP susceptibility locus in 8q24 in humans, was coherent with the postulated presence of a regulatory element(s) in the embryonic face. Altogether, the expression data in mouse and embryonic origins of CLP suggest that the expression domains correspond to the activity of a long-range regulatory element(s), which may constitute the target of variants causing 8q24 associated CLP (Figure14).

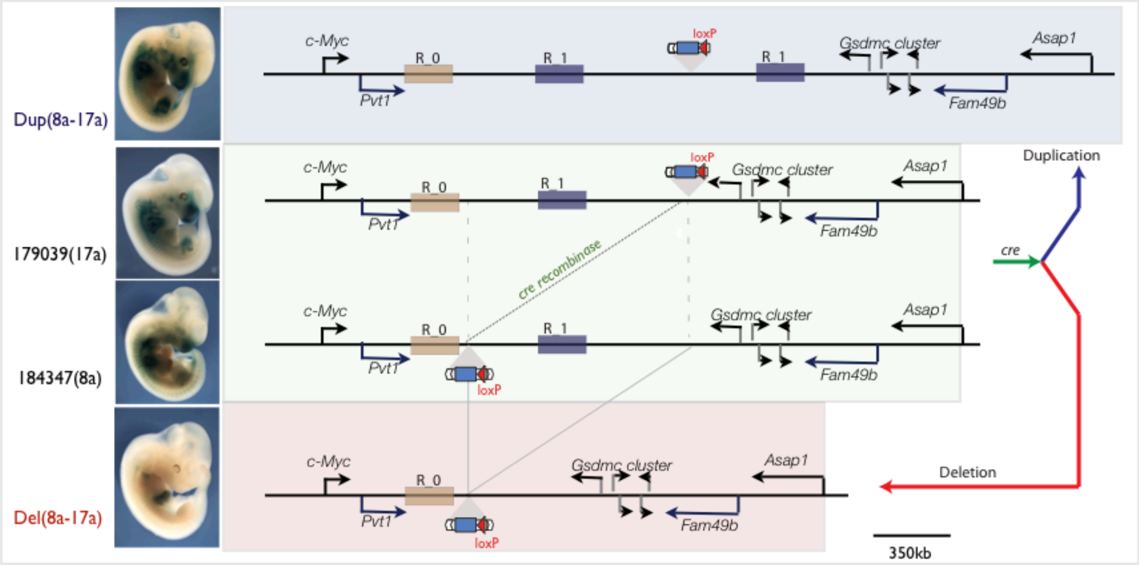


Figure15|Chromosomal Engineering in the *c-Myc/Pvt1* flanking locus by TaMaRe: In the animals triple positive for two insertions in trans and *cre* recombinase (highlighted in yellow), *loxP* site recombination leads to the duplication (highlighed in purple) and deletion (highlighed in red) of the regions between two insertions. Upon recombination loxP sites are reconstructed therefore, lacZ staining reflects the regulatory input at the breakpoints(left panel).Ro and R1 represents two hypotherical distant regulatory element in the non-coding regions.

The identification of a regulatory activity driving gene expression in the developing medial nasal mesenchyme supported that the cleft lip/palate susceptibility was due to a regulatory variant. However, several questions were to be answered:

- the more precise localization of the cis-regulatory element(s) responsible for this activity
- the endogenous gene(s) under this control

- the biological role of this regulation for facial development.

To address these, I generated a series of overlapping deletions and duplications over this interval. Using the loxP sites in the Sleeping Beauty Transposon, I could use Targeted Meiotic Recombination (TAMERE) to create new rearrangements (Hérault Y *et al* 1998, Ruf S *et al* 2011) (Figure_15) *Cre*-mediated recombination products were obtained with a frequency of 5% to 15%, similar to what has been described in Ruf S *et al* 2011). Noteworthy, the recombination reconstitutes the reporter, which allows us to compare the expression before at the breakpoints and after the recombination.

5.2.2 Locating Embryonic Face Specific Regulatory Elements

I used these rearrangements to localize the elements that were responsible for the regulatory activities ascribed to this large region. Deletion of the region in between 8a and 17a {del(8a-17a)} leads to complete loss of regulatory input on the minimal promoter in the tissue (Figure 15). On the other hand, the lacZ staining obtained in del(7a-8a) is exactly like 7a, and it does not lead to a loss of regulatory input sensed by the transposon. del(c8-7a), which spans the *c-Myc* gene and covers 1.5mb region centromeric to 7a still show strong expression in the MFM , whereas in a mosaic manner. del(17a-21a), on the telomeric side of critical (8a-17a) region, does not change the regulatory activity in the embryonic face. The analysis suggests that the elements associated with MFM and NE, are within (8a-17a) interval. (Figure 15)

In order to map their location more precisely, I analyzed lacZ expression in the alleles with only partial deletions of the (8a-17a) region: del(15a-17a) did not alter any of the expression domains in the embryonic face (Figure 16A). The deletion of (14b-15a) and the deletion of (14c-17a) both did not change MFM expression but lead to a loss of NE expression domain. The deletion of three overlapping regions, (8a-14b) and (8b-13a), and (7a-14a), lead to the complete loss of regulatory input on MFM but NE staining was retained upon deletions. In conclusion, the critical regulatory region for NE expression is located in the (14b-15a) region, whereas, MFM expression requires element(s) located between 8a and 13a insertion sites (Figure 16 B-C-D).

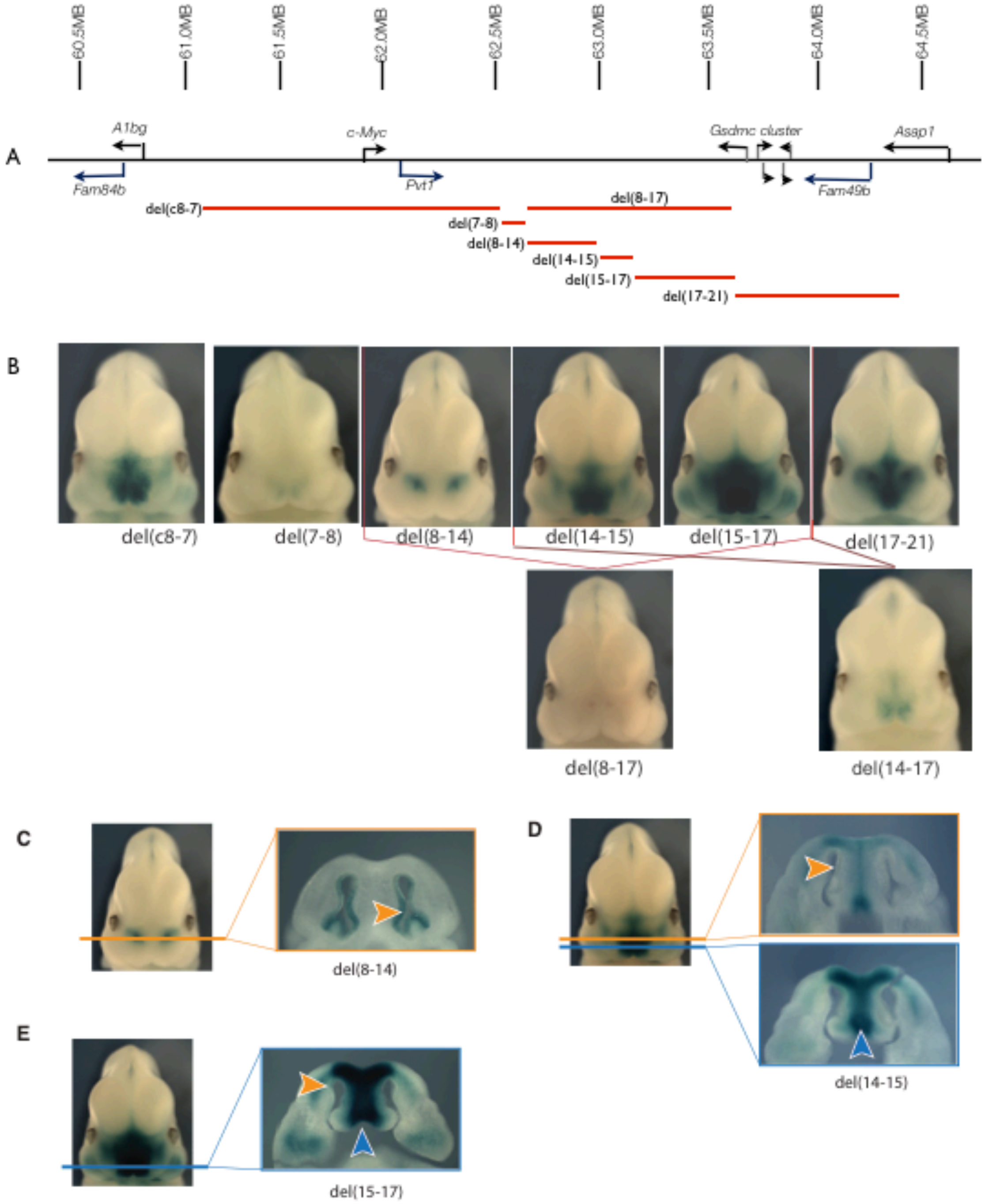


Figure16|Locating Regulatory Elements by Systematic Dissection of the Locus: A) The *c-Myc/Pvt1* flanking locus and the relative positions of the deletions are shown in red. B) The regulatory activity -reveal by lacZ staining- acting on the regulatory sensor at the breakpoints are shown C-D-E) Vibrotome sections on the embryos reveal that the regulatory regions for MFM is in the (8a-14a) interval and for NE is in the (14-15a) interval.

After showing the necessity of the (8a-13a) region and the (14b-15a) regions for regulatory activity upon deletions, I used two duplications to investigate whether these regions were sufficient to drive expression in the MFM and NE. The duplication of (10-20a) showed that the telomeric region of 10a has an autonomous regulatory activity, which is sufficient to provide regulatory input for the MFM. On the other hand, the dup(13a-20a) indicated that the telomeric side of the 13a insertion is sufficient only for the NE expression. Consequently, the enhancer(s) that regulates MFM expression appeared to be in the in the 250kb long (10a-13a) region. The enhancer(s) that regulates the NE expression is in the 100kb long (14a-15a) region, which is sufficient for this regulatory activity. These regulatory elements can still function even though they are centromerically not in their native genomic context.

In order to narrow down the critical elements for regulatory activity in MFM, which is the critical tissue for lip and palatal shelf formation, Massimo Petretich conducted a ChIP-seq experiment on the e11.5 embryonic face to find out the active enhancers in the region by using two enhancer marks: H3K4me1 and H3K27Ac. Furthermore, publically available EP300 binding site datasets from facebase.org are obtained to characterize (10a-13a) region biochemically in the embryonic face. According to the H3K4me1, H3K27Ac and EP300 marks indicate presence of multiple potential enhancer sites in (10a-13a) region, some of which are located on conserved blocks (Figure 17). Two sequences (#1 and #7) were selected among seven candidates and none of them showed reproducible regulatory activity in MFM and NE in lentivirus injections performed by Katja Langenfeld. However, the human orthologous sequence of a slightly larger fragment of #1 (hs1870) is tested by VISTA enhancer project and showed reproducible short-range activity in MFM (Attenasio C *et al* 2013). In addition, a mouse BAC(RP23-350P5) reporter partially covering the region did not show any reproducible reporter activity as well (injections done by Yvonne Petersen from EMBL, transgenic facility) (Figure 17) One or more of these candidate regions may be the critical element(s) for MFM and NE. Regulatory activity in MFM and NE may be a consequence of a composite activity by provided by different modules and may rely on the genomic environment of the enhancers *in* *vivo.*

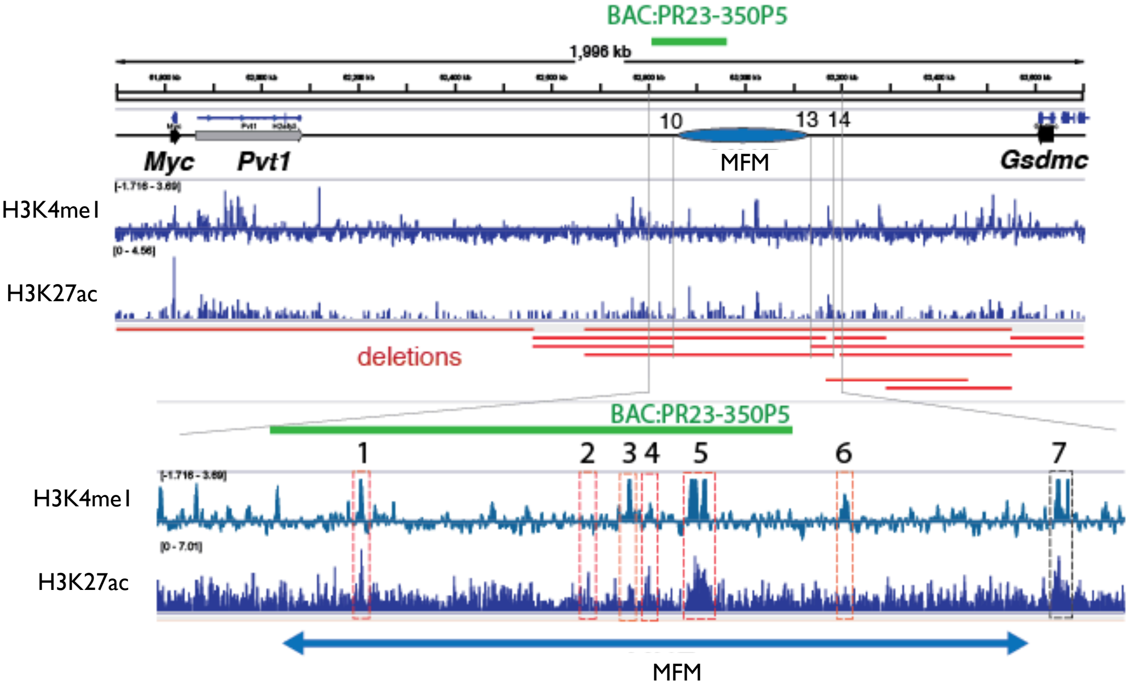


Figure17|Candidate MFM-enhancer sequences in the *c-Myc/Pvt-1* flanking locus:
The upper panel shows the normalized read-counts obtained in ChIPseq experiment using H3K4me1 and H3K27Ac specific antibodies. The red lines represent the deletions used to restrict down the regulatory region for MFM expression. The lower panel zooms in the MNE region and seven peaks, which show enhancer marks are taken as candidate regions for MFM enhancers. The green line shows the BAC that is used for reconstruction of a BAC-regulatory reporter. MNE: medial nasal enhancer. ChIPseq is performed by Massimo Petretich.

5.2.2.1 Target(s) of Embryonic-Face Specific Regulatory Regions

In order to find out whether these regions regulate endogenous gene expression, firstly, I listed the genes and lncRNA annotated in the region (from UCSC genome browser, Ensembl Genome Browser, and lncRNA database) and analyzed the expression of these in the embryonic face by quantitative real-time PCR (qPCR). Expression of all of these genes except AK040104.1 lncRNA, which is the mouse homologue of the human CCDC26 transcript, was found in e11.5 wildtype face (Figure 7B). I then investigated if and how these expression levels were affected by the deletion of the (8a-17a) interval. The samples homozygous for (8a-17a) deletion, in which the lacZ regulatory reporter expression is completely lost in the

MFM and the NE, showed a significant reduction only in the *c-Myc* gene when compared to the wildtype littermates (Figure 18A). This downregulation was not seen in samples homozygous for 17a. Therefore, it couldn't be due to a titration of transcriptional activity by the minimal promoter (Figure 18B). In addition, as the (8a-17a) deletion removed a highly expressed lncRNA, AK08920, one cannot deduce a potential regulatory effect of this deletion on the AK08920 (Figure 7A). Therefore, in e11.5 face, (8a-17a) region exclusively acts on the *c-Myc* gene more than 0.8Mb away and has no influence on the flaking genes including *Gsdmc* cluster, which is on the telomeric side and *Pvt1* gene, which is between *c-Myc* and the (8a-17a) interval.

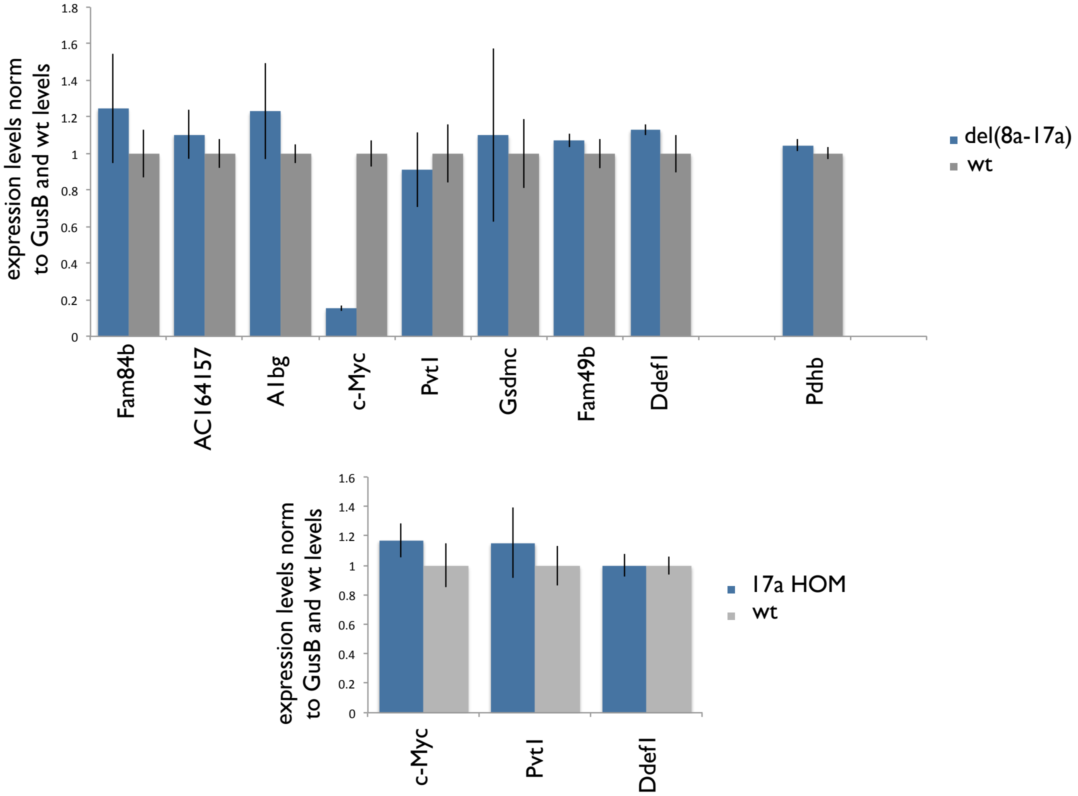


Figure18|The Target of Regulatory Elements in the *c-Myc/Pvt1* flanking locus: A) Comparison of qPCR results between del(8a-17a) and wt littermates to find out the effected genes in e11.5 face upon face enhancer deletion. The qPCR results are all normalized to *GusB* expression and expression levels relative to wildtype average is shown. B) qPCR results (normalized to *GusB)* are compared between homozygous 17a insertion and wt littermates in e11.5 embryonic face show that the regulatory sensor does not titrate the gene expression.

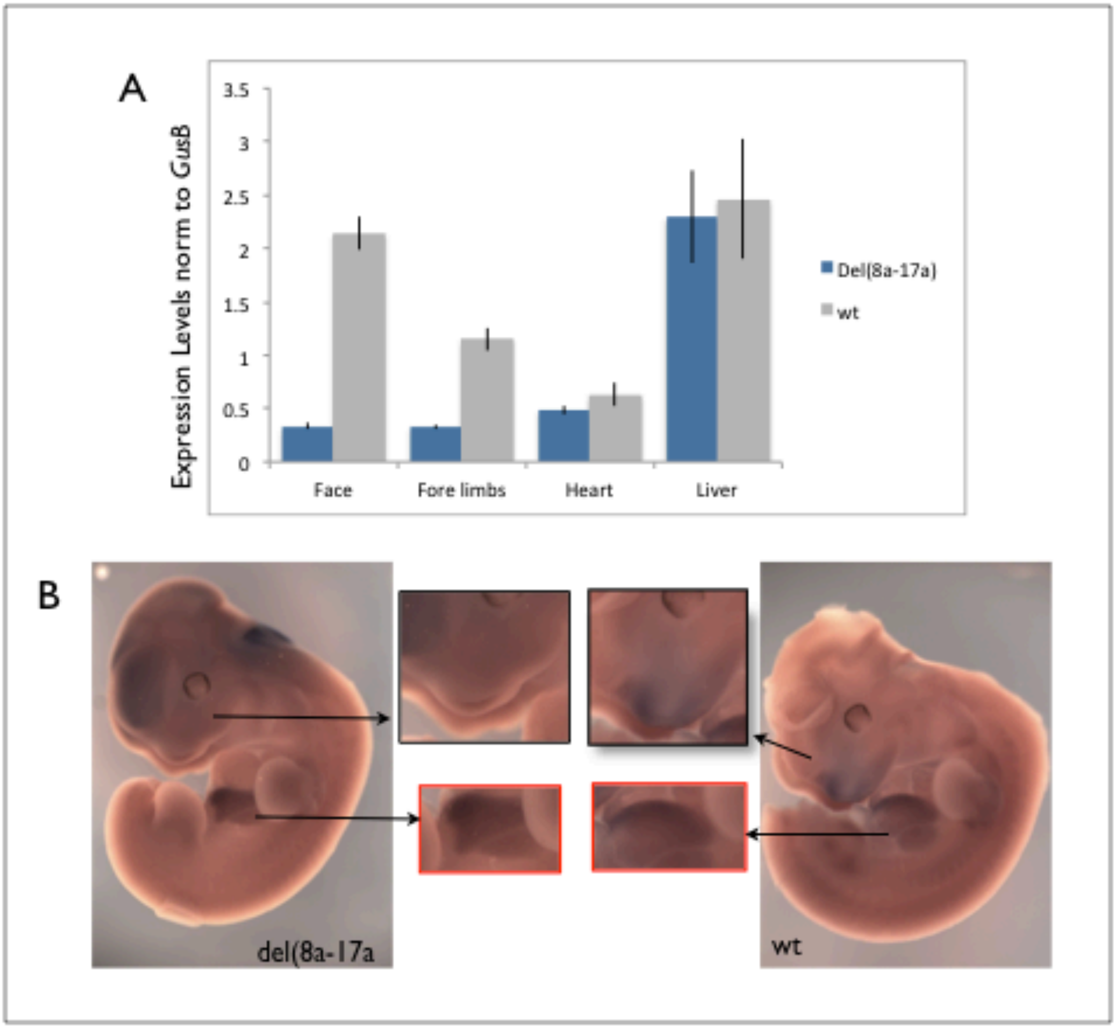


Figure19|Tissue Specific Downregulation of *c-Myc* upon (8a-17a) deletion: A) qPCRs are done on e11.5 embryonic face for *c-Myc* gene in the given tissues for del(8a-17a) and wt genotypes. The *c-Myc* expression is normalized to *GusB* expression. B) *in situ* Hybridization on e11.5 whole mount embryos using *c-Myc* antisense probe supports that the *c-Myc* downregulation is tissue specific as the liver is expressed in both genotypes but face expression is only detected in wt.

I carried out a qPCR analysis to monitor the changes in *c-Myc* expression in the face, the forelimbs, the liver, and the heart at this stage upon (8a-17a) deletion to define whether the *c-Myc* downregulation was tissue-specific or more general. In the forelimbs, where the lacZ staining is lost in the samples homozygous for (8a-17a) deletion, *c-Myc* expression also decreases (Figure 19A) On the contrary, in del(8a-17a), *c-Myc* expression did not change in the heart and in the liver, for which no regulatory elements are active in this region. These tissue specific changes were confirmed by, *in situ* hybridization using *c-Myc* antisense probe: Homozygous embryos for del(8-17) did not show the strong facial expression of *c-Myc* detected in wt control, whereas the liver expression was indistinguishable between the two genotypes (Figure 19B). This indicates that (8a-17a) region in *c-Myc/Pvt-1 flanking locus* provides tissue specific regulatory input to *c-Myc* gene in embryonic face and forelimbs.

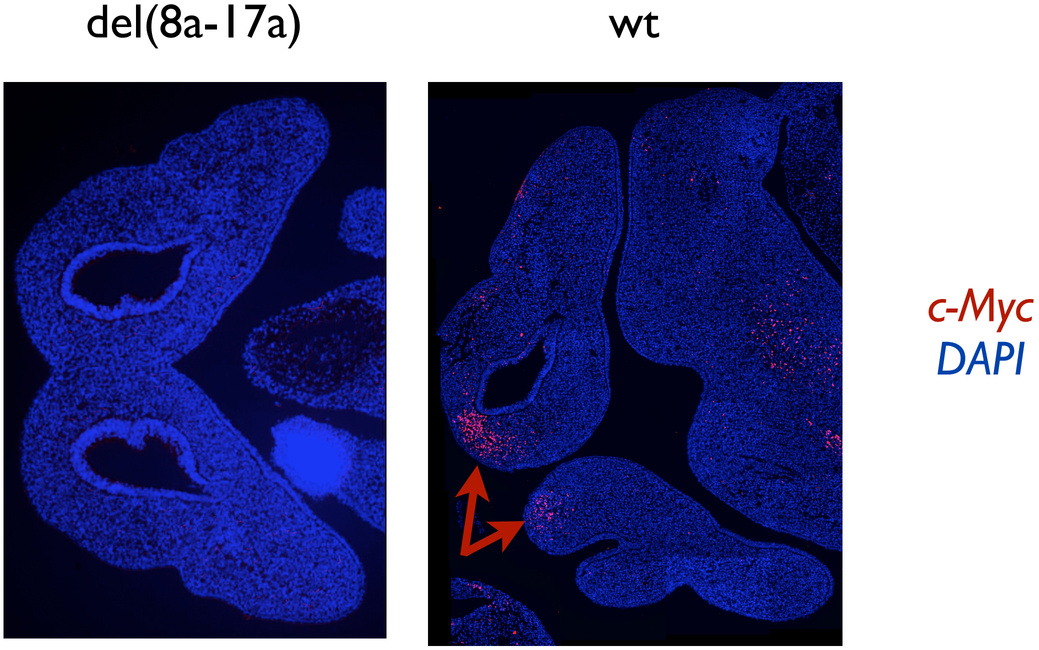


Figure20|*c-Myc* protein levels in del(8a-17a) and wt: Immunofluorescence experiment on e11.5 embryonic face detected *c-Myc* protein (shown in red) in the wt but not in the del(8a-17a). DAPI (blue) is used as counter stain and *c-Myc* is localized in the nucleus in wt allele.

In addition to the changes in the transcription level, immunofluorescence experiments using a c-Myc antibody revealed the presence of c-Myc proteins in the nucleus of cells in the frontal part of MFM in wt samples. This signal was completely missing in facial mesenchyme of e11.5 embryos homozygous for (8a-17a) deletion (Figure 20). It has already been shown that the amount of c-Myc protein responds rapidly to the changes in the transcript level due to its 30min-long half-life (reviewed in (Wiestra I and Alves J, 2009)). Most likely as a result of this, c-Myc protein decreases in the face upon (8a-17a) facial enhancer deletion.

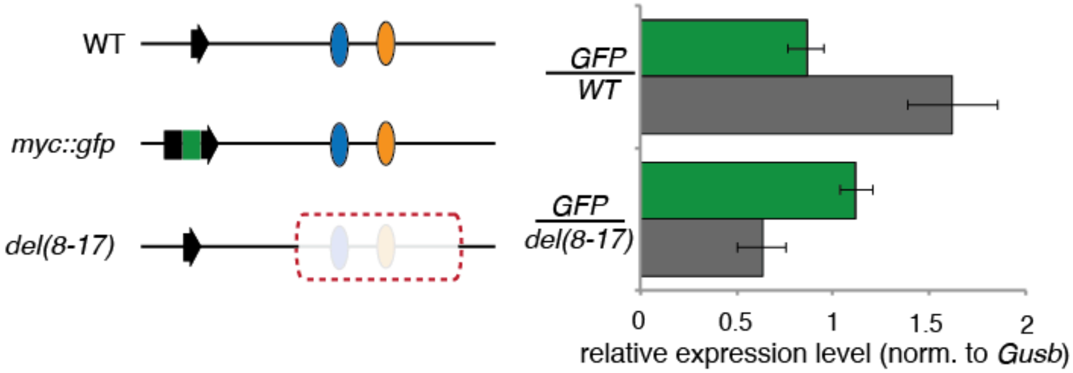


Figure21| cis activity of the regulatory elements in (8a-17a) region: qPCR is performed on e11.5 embryonic faces having wt and *myc::gfp* alleles or having del(8a-17a) and *myc::gfp* alleles. *c-Myc* and *GFP* expression is normalized to *GusB. c-Myc* expression in wt and del(8a-17a) alleles are compared using *myc::gfp* allele as a reference (Figure designed by Francois Spitz)

In the embryonic faces, which are heterozygous for the deletion of (8a-17a) region, the reduction in the expression of *c-Myc* is about half of the reduction observed in embryonic faces, homozygous for del(8a-17a). This observation was coherent with a *cis*-regulatory effect. Yet, to prove it directly, I carried out an additional experiment, using alleles where I can distinguish the two *c-Myc* alleles (Figure 21, left panel). For this, I used a *c-Myc* allele, modified by insertion in frame to a *GFP* (Huang CY *et al* 2008) to identify the contribution of the (8a-17a) region to *c-Myc* expression in cis. Males homozygous for c-Myc-GFP are crossed with heterozygous (8a-17a) deletion females and c-Myc-GFP allele is taken as reference. The qPCR results indicate that (8a-17a) deletion decreases the level of the endogenous *c-Myc*, which is in *cis* with the del(8a-17a) but does not decrease the level of *c-Myc-GFP*, which is in trans. Namely, the regulation of *c-Myc* in the embryonic face via distant regulatory sequences in (8a-17a) interval takes place in *cis* not in *trans (*Figure 21, right panel).

I further investigated the influence of smaller deletions in (8a-17a) region on *c-Myc* expression in embryonic face by qPCR analysis (Figure 22). Firstly, the deletion of the (8a-14a) region containing the regulatory element(s) for MFM decreased *c-Myc* expression nearly as strong as (8a-17a) deletion. Del(7a-10a) did not cause any significant reduction in the embryonic face expression of *c-Myc.* Therefore, the source of regulatory reduction in (8a-14a) is restricted to (10a-14a) region. The deletion of regulatory sequences for NE in (14c-15a) region and (15a-17a) deletion caused a very mild reduction in the *c-Myc* expression. This suggests that the contribution of NE enhancer in *c-Myc* expression in embryonic face is very limited either due to the number of the cells that it is active or due to its weak regulatory strength. Consequently, in concordance with the lacZ expression analysis of *c-Myc/Pvt-1* flanking locus deletions, functional regulatory elements that act on *c-Myc* in MFM at e11.5 reside in (10a-14a) region (Figure 22).

5.2.2.2 Consequences of Embryonic Face Enhancer Deletions

The (10a-14a) region, which has functional enhancers, is orthologous to the human 640kb-long LD block, which is associated with NSCLP (Birnbaum S *et al* 2009). In addition, the variations in the very same haplotype block in the Northern European population are also associated with changes in the facial morphology (Liu F *et al* 2012) (Figure 14 A-B). Therefore, I investigated the consequences of the deletion of the (8a-17a) region on facial development.

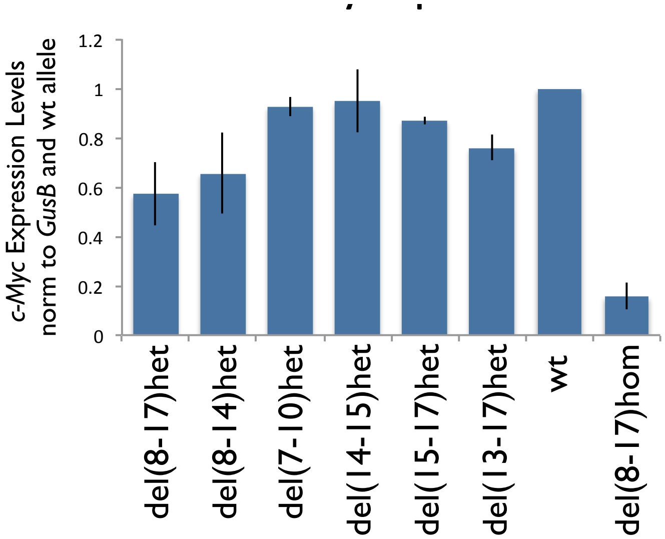
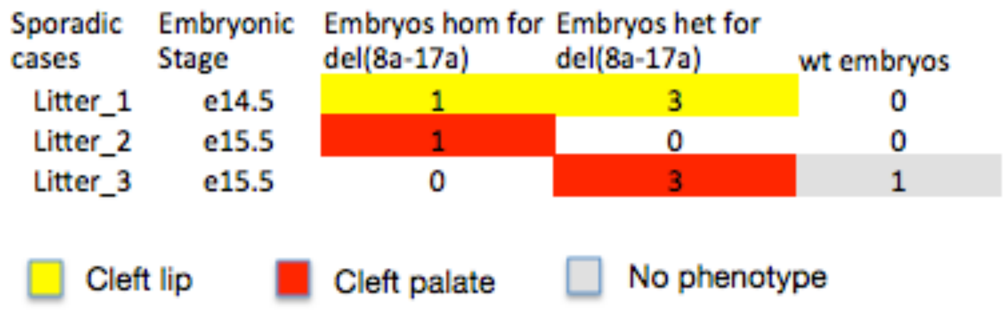


Figure22|Dissecting the (8a-17a) to narrow down the regulatory region: qPCR is performed on e11.5 face samples heterozygous for the given deletions. *c-Myc* expression level is normalized to *GusB* and *c-Myc* expression levels relative to average wt expression level is given in this figure. In addition, *c-Myc* qPCR on samples homozygous for del(8a-17a) is shown for comparison. wt error bar is added to the individual error bars by using the "error propagation" formula.

5.2.2.2.1 Sporadic Cleft Lip and Palate in del(8a-17a) embryos

I screened the heterozygous and homozygous animals for CLP incidence. None of the 60 adult (at least at three weeks old) mice, which are homozygous (8a-17a) deletion and none of 200 adult mice, which are heterozygous for (8a-17a) deletion showed a Cleft Lip Palate Phenotype.



Sporadic cases	Embryonic Stage	Embryos hom for del(8a-17a)	Embryos het for del(8a-17a)	wt embryos
Litter_1	e14.5	🟨 1	🟨 3	0
Litter_2	e15.5	🟥 1	0	0
Litter_3	e15.5	0	🟥 3	⬜ 1

🟨 Cleft lip 🟥 Cleft palate ⬜ No phenotype

Table1| Sporadic Cleft Lip and/or Palate Cases: Three litters with in total 8 embryos showed sporadic Cleft Lip and/or Palate phenotype. Cleft Lip phenotype is highlighted in yellow and cleft palate phenotype in red. Each column shows a phenotype and each raw represents a different litter.

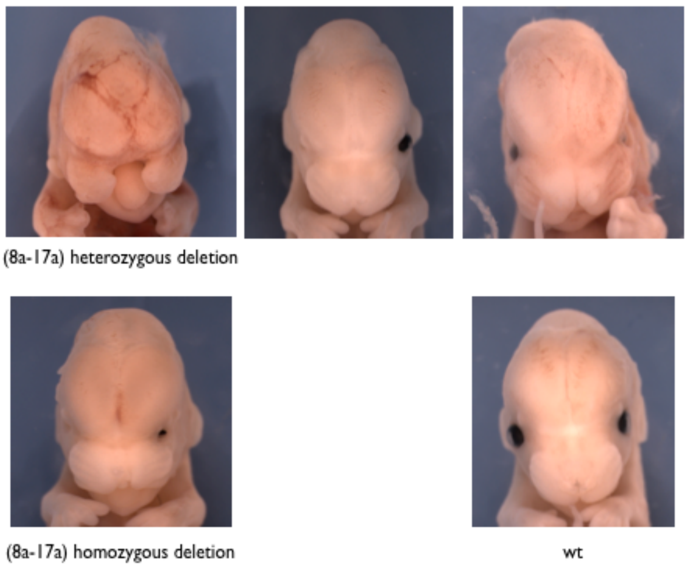


Figure23| Sporadic CLP cases: All four embryos at e14.5 are shown from Litter_1(Table1). wt sample (bottom right) comes from a different litter. The top raw shows the embryos heterozygous for del(8a-17a) and the bottom left sample is homozygous for del(8a-17a). The phenotype appears in a range from complete loss of upper lip to a subtle cheiloschisis.

Nevertheless, I have encountered a sporadic case in a litter (Litter_1) with 4 e14.5 embryos, which are either homozygous or heterozygous for (8a-17a) deletion and these embryos showed cleft lip phenotype within a large range of severity (Table 1, Figure 23). Among 104 embryos at this stage and none of the others showed cleft lip phenotype. Since palate formation is not complete at day 14.5, I carried out the rest of the phenotypic screen at e15.5 and I obtained 2 litters (Litter_2 and Litter_3) with five embryos, four of which had cleft palate phenotype. (8a-17a) region was deleted either in one allele or in both alleles in all of these sporadic cases. I have screened other 40 embryos coming from heterozygous del(8a-17a) crosses at this stage and none of the embryos showed a phenotype. Thus, whereas the large majority of del(8-17) heterozygous and homozygous failed to show facial abnormalities, sporadic cases were obtained with variable expressivity.

5.2.2.2.2 Craniofacial Morphology changes in the del(8a-17a) adults

In the adults, the measurements of characteristic distances from between 18 different landsmarks from two different angles in the skeletal preps of 9 mice (4 of them 5 weeks old, 5 of them 3 weeks old) homozygous del(8a-17a) and 11 wt (4 of them 5 weeks old, 7 of them are 3 weeks old) were performed and normalized to the wildtype average of the same age group (Figure 24 A-B). The analysis of the measurements showed that nasal bone length in caudal-rostral direction is about 20% smaller in del(8a-17a)($p<10^{-4}$) and the shape of the nasal bone is different at the caudal end. In del (8a-17a) the size reduction is 16.5% ($p<10^{-5}$) in the frontal bones. 15% ($p<10^{-3}$) in the zygomatic arches, and 12% ($p<10^{-5}$) in the mandibular bones but the parietal bone size is not significantly affected from this deletion. More interestingly, interorbital distance in del(8a-17a) mice is 4% larger when compared to wt mice ($p<0.05$). The variance of the morphology is significantly higher in del(8a-17a) for side-to-side, lateral measurements both on the dorsal and ventral side of the face, whereas, from caudal to rostral measurements do not show significant variation between homozygous (8a-17a) deletion and wt. In brief, del(8a-17a) changes the facial morphology by changing the growth itself and the range of growth in different features in different severity at post-natal stages (Figure 24-C).

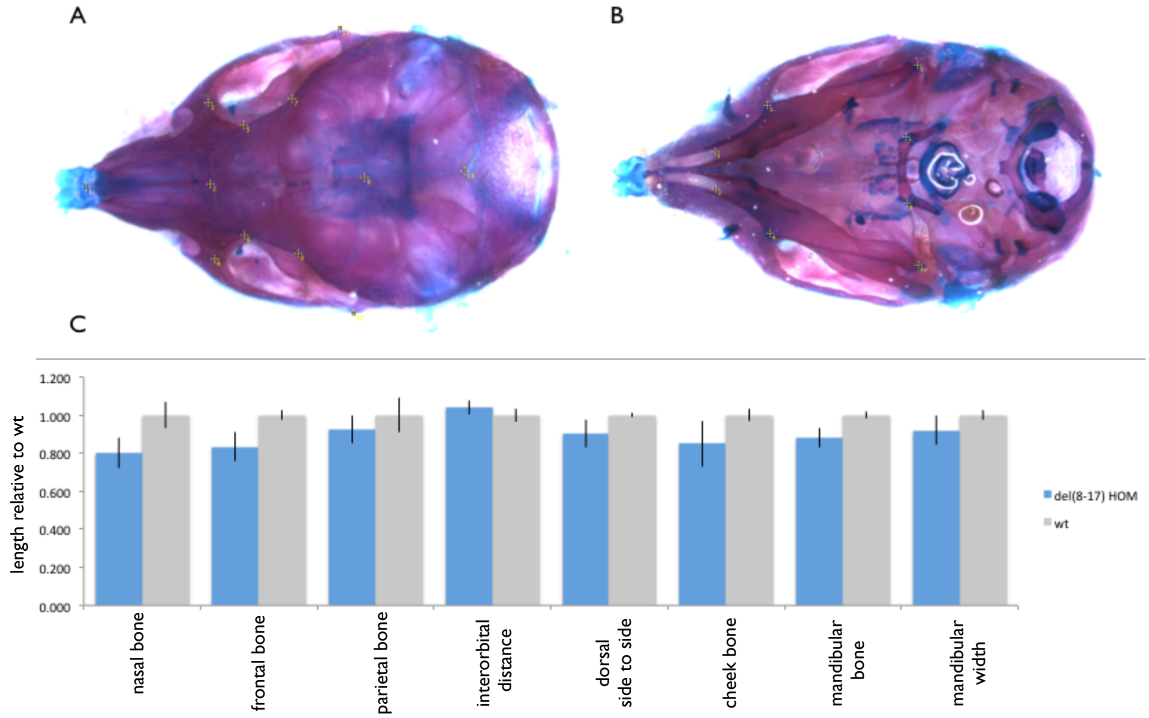


Figure24|Craniofacial Morphology Changes in del(8a-17a): Craniofacial morphology measurements are performed on skulls stained with Alizarin Red and Alcian Blue. Yellow marks indicate landmarks given in the Materials and Methods A) Top view of a skull of 3 weeks old B) Bottom view of a skull of 3 weeks old C) The measurements of the given facial features are performed on FIJI and the values are normalized to the wt average values. The data for 3 weeks old and 5 weeks old animals are pooled.

5.2.2.2.3 Facial Morphology Changes in the del(8a-17a) embryos at e11.5

Coordinated growth of facial processes at the embryonic stages is essential for the morphogenesis of the face. Therefore, I compared the morphology of the embryos homozygous for del(8a-17a) to wildtype littermates, I measured different characteristic distances between 24 landmarks on 150 sections of e11.5 embryonic face sections of 5μm thickness (Figure 25). According to these measurements, the size of the face from one side to another significantly decreased in homozygous (8a-17a) deletion embryos when compared to wildtype littermates. I found that the width of the LFM, and NE do not change in homozygous (8a-17a) deletions when compared to wt sections (p=0.97, p=0.11, respectively). In contrast, the MFM was

20% smaller in homozygous (8a-17a) deletion embryos when compared to wt littermates ($p<10^{-20}$) (Figure 25)

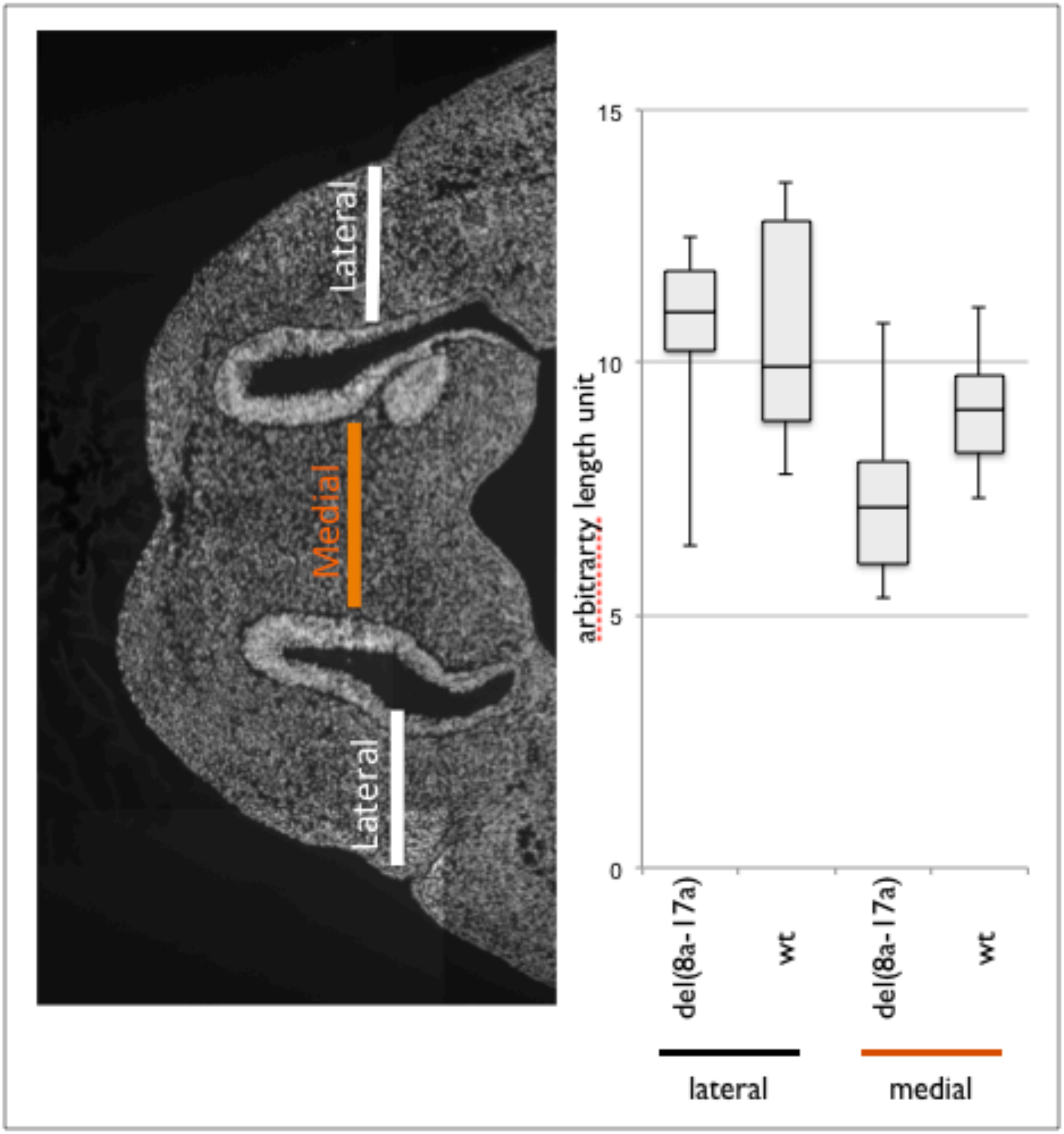


Figure25|Embryonic Face Morphology Changes in del(8a-17a): 5μm paraffin sections of e11.5 face (left) is used for morphological measurements. Lateral length is the sum of left and right lateral facial mesenchyme length. It is not different between wt and del(8a-17a)HOM (p~0.97) Medial length is the medial face mesenchyme length in the projection of lateral lines ($p<10^{-20}$).

In order to complement the measurements performed on section, I carried out the morphometric analysis on e11.5 embryos by using Optical Projection Tomography (OPT). OPT images showed that in earlier stages of e11.5 (where the forelimb distal-proximal length is less between 1mm and 1.15mm) the interorbital distance is shorter by 9% ($p<0.05$) in del(8a-17a). On the other hand, in later stages of e11.5 (where the forelimb distal-proximal length is between 1.15mm and 1.25mm) the distance between the nostrils decrease by 12% ($p<0.05$), whereas the lengths of the other features do not change significantly. Consequently, despite the very dynamic face morphogenesis taking place at this stage, statistical analysis of both 2D and 3D measurements indicate that (8a-17a) dependent *c-Myc* downregulation changes the morphology of the embryonic face (Figure 26)

5.2.2.2.4 Cellular Consequences of the del(8a-17a)

After describing the morphological changes upon del(8a-17a) in the face of e11.5 embryos and the adult crania, I examined the apoptotic and the proliferative changes in e11.5 face to find out the cellular changes that lead to these morphological changes. Immunostaining with a Cleaved Caspase 3 (CC3) antibody, which marks apoptotic cells, on paraffin sections of e11.5 embryonic faces showed that apoptosis in both wildtype and homozygous (8a-17a) deletion do not exceed 2 cells per section, which in concordance with the reported apoptotic cell level in the early stages of the face development (Beverdam A *et al* 2001) (Figure 27). In order to assess the cellular proliferation, I used an antibody against phosphorylated histone 3 (PH3) on the paraffin sections of e11.5 face coming from wt and homozygous (8a-17a) deletions. Meticulous image analysis of over three hundred sections coming from 5 wt and 7 homozygous del (8a-17a) showed mild (7%) but very significant ($p<10^{-5}$) downregulation in proliferation rate in homozygous (8a-17a) deletions when compared to wt littermates (Figure 28). Therefore, *c-Myc* downregulation in the embryonic face, which is caused by the loss of its MFM specific regulatory region, leads to a significant reduction in the number of the proliferating cells in the embryonic face.

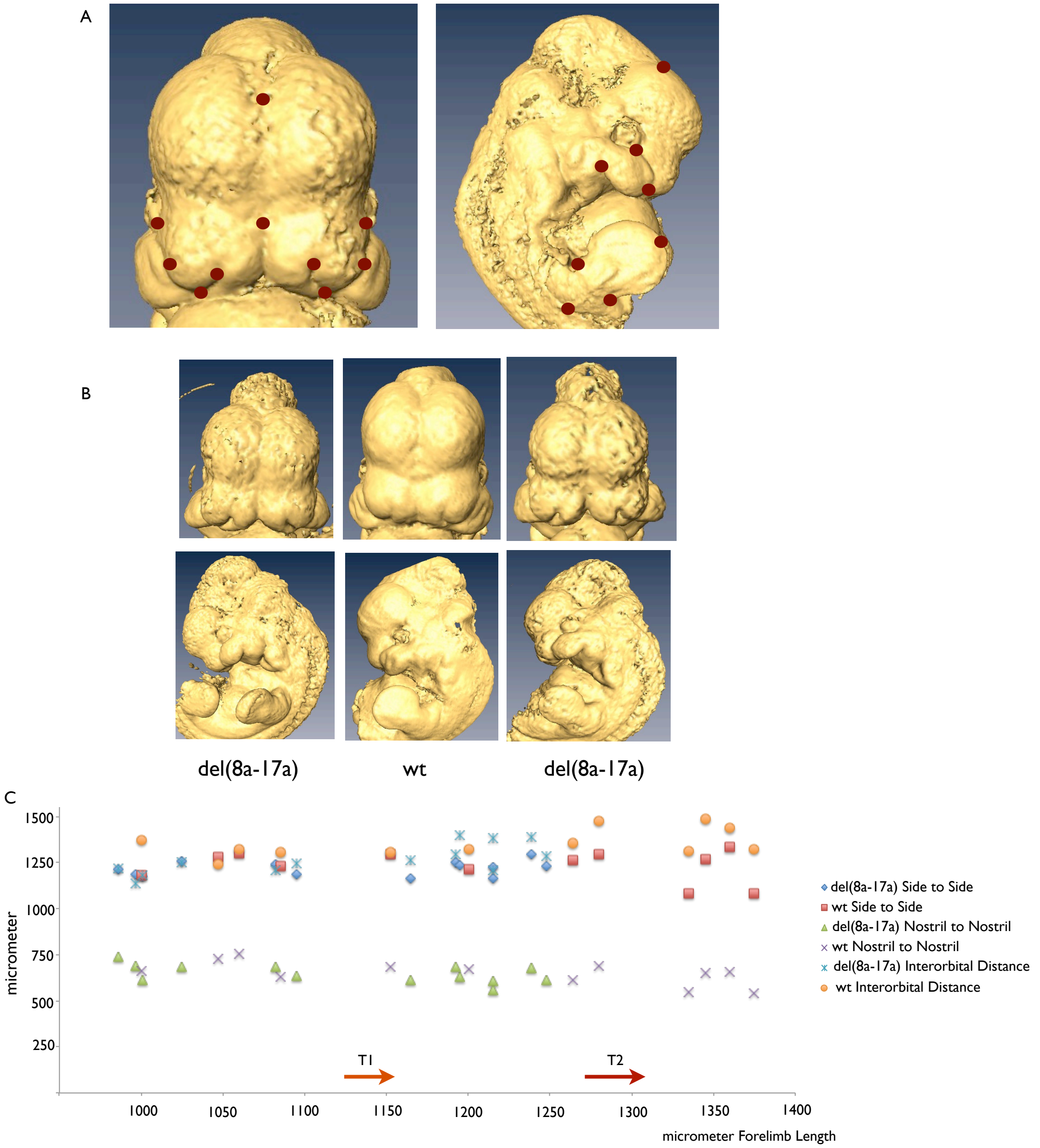


Figure26| Optical Projection Tomography on e11.5 embryos: A) The frontal and the side view of e11.5 wt embryo is given. The red dots indicate the landmarks used for measurements in 3D. B) A set of embryos at different stages, which the landmarks can be spotted, is shown. C) Measurements done on the embryos are sorted according to their distal-proximal limb size as an indication of stage (x axis) and the other facial features are plotted according to the limb size (y-axis). T1 transition indicates the fusion of the nasal processes and T2 transition indicates the appearance of slits on the branchial arches. The comparisons between the genotypes are done for the embryos younger than T1 stage or between T1 and T2 stages.

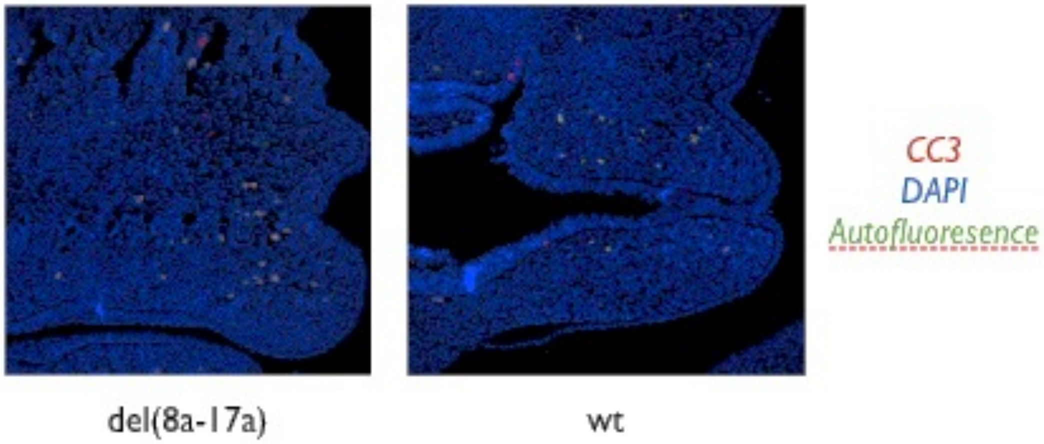


Figure27|Apoptosis in the embryonic face: Immunofluorescence done by Cleaved Caspase 3 antibody on 5μm sagittal sections of e11.5 embryonic face shows that very few cells are positive for this apoptosis marker (in red) both in samples homozygous for del(8a-17a) and wt. DAPI is used as counterstaining (blue) and GFP emission channel is used detect auto fluorescence originating from hematopoietic cells.

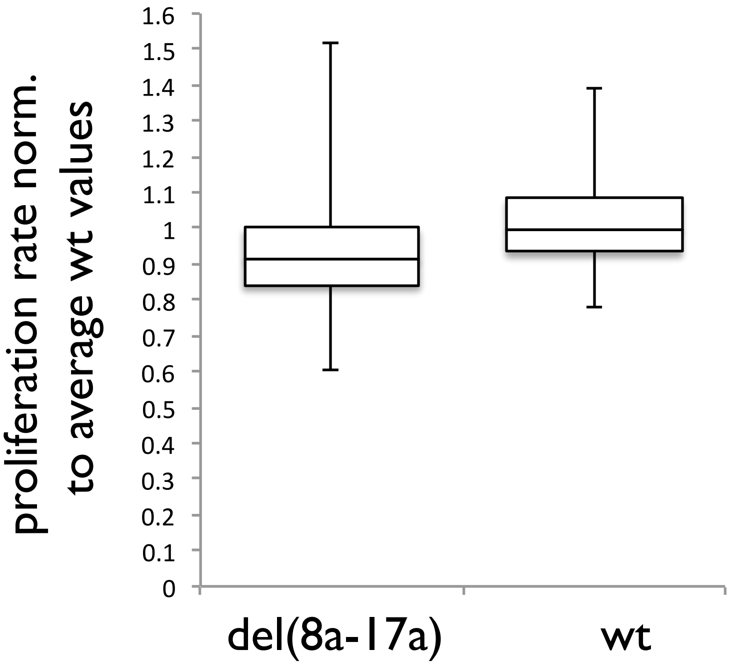


Figure28|Change in the proliferation rate upon the (8a-17a) deletion: Proliferation rate is described as the number of cells showing Phospho Histone 3, mitosis specific staining staining over the total number of cells in 5μm coronal sections of e11.5 embryonic faces. Calculated proliferation rates are normalized to average wt values. del(8a-17a) shows a mild but significant decrease when compared to wt samples.

To sum up, the results indicate that *c-Myc/Pvt-1* flanking locus has two critical regions that have regulatory activity in the embryonic face. The (10a-14a) interval is a functional region that regulates *c-Myc* gene. Besides the (14a-15a) region is critical for NE expression domain across *c-Myc/Pvt-1* flanking locus. The (8a-17a) region contributes to facial development by modulating *c-Myc* levels, which influences the proliferation rate in the embryonic face. Furthermore, I have shown that (8a-17a) region involves in the facial morphogenesis both the embryonic stage and in the adult mice.

5.2.2.3 Downstream Effects of *c-Myc* downregulation in the embryonic face

In order to find out the downstream pathways deregulated by the *c-Myc* downregulation, we performed an RNAseq experiment in collaboration with the Genomics Core Facility at EMBL. Transcriptome of 4 homozygous del(8a-17a) embryonic faces at e11.5 are compared with 4 samples from wildtype littermates. Data analysis done by John Marioni and Nuno Fonseca found that 101 genes were deregulated (FDR=0.05) (Figure 29A). Among all the genes, *c-Myc* is the most strongly downregulated gene and the level of downregulation obtained in RNAseq experiment for the *c-Myc* gene is consistent to what I have obtained in the qPCR. The rest of the deregulated genes can be separated into three groups:

The first group of genes is known to involve in hematopoiesis. Most of these are expressed exclusively in blood cells. The presence of two large arteries and extensive capillary web in the embryonic face explains why they are detected in the sample analyzed. Their overall low expression levels (<100 rpmk) reflect the limited number of blood cells in the dissected samples. In addition, qPCR verification of the RNAseq results indicated that the downregulation of hematopoietic genes does not take place in embryonic face homozygous for (8a-14a) deletion, where the MFM enhancers reside (Figure 29 A-D). On the other hand qPCR on embryonic faces for homozygous for (14c-17a) deletion and (8a-17a) deletion recapitulate the downregulation of the hematopoietic genes, suggesting the presence of hematopoietic system enhancers in the (14c-17a) region (Figure 29 D)

The second set of is composed of genes, which show moderate level of gene expression and their expression fold-change is in 20% range (Figure 29 A-C). These genes include several transcription factors and a signaling protein *Bmp7*. qPCR analysis of these genes verified that samples homozygous for (8a-14a) deletion and (8a-17a) deletion lead to mild but significant change of these genes, whereas, del(14c-17a) samples had no effect on these genes. Among these genes Bmp7, Etv5 and Sox11 levels significantly increase, whereas, Nr2f1 expression significantly decreases in del(8a-17a) samples. Sox11 and Bmp7 are known to contribute to the etiology of CLP in mouse as shown by knock-out phenotypes (Sock E *et al* 2004, Kouskoura T *et al* 2013) (Figure 29C)

The third group of genes that show significant transcriptional change upon *c-Myc* downregulation consists of highly expressed genes, which show mild but significant downregulation (Figure 29 A-B). This group includes genes involved in ribosome biogenesis and translation. qPCR verifies that both del(8a-14a) and del(8a-17a) recapitulate ribosomal protein genes are downregulated in the range of 10% to 20%. In the literature, ribosomal biogenesis genes and translational machinery genes show reproducible decrease in expression upon *c-Myc* downregulation in many cell lines, which show high level of *c-Myc* expression (reviewed in (van Riggelen J *et al* 2010)). The decrease in *c-Myc* expression from its physiological levels can still lead to the downregulation of ribosome biogenesis and translation related (Figure 29B)

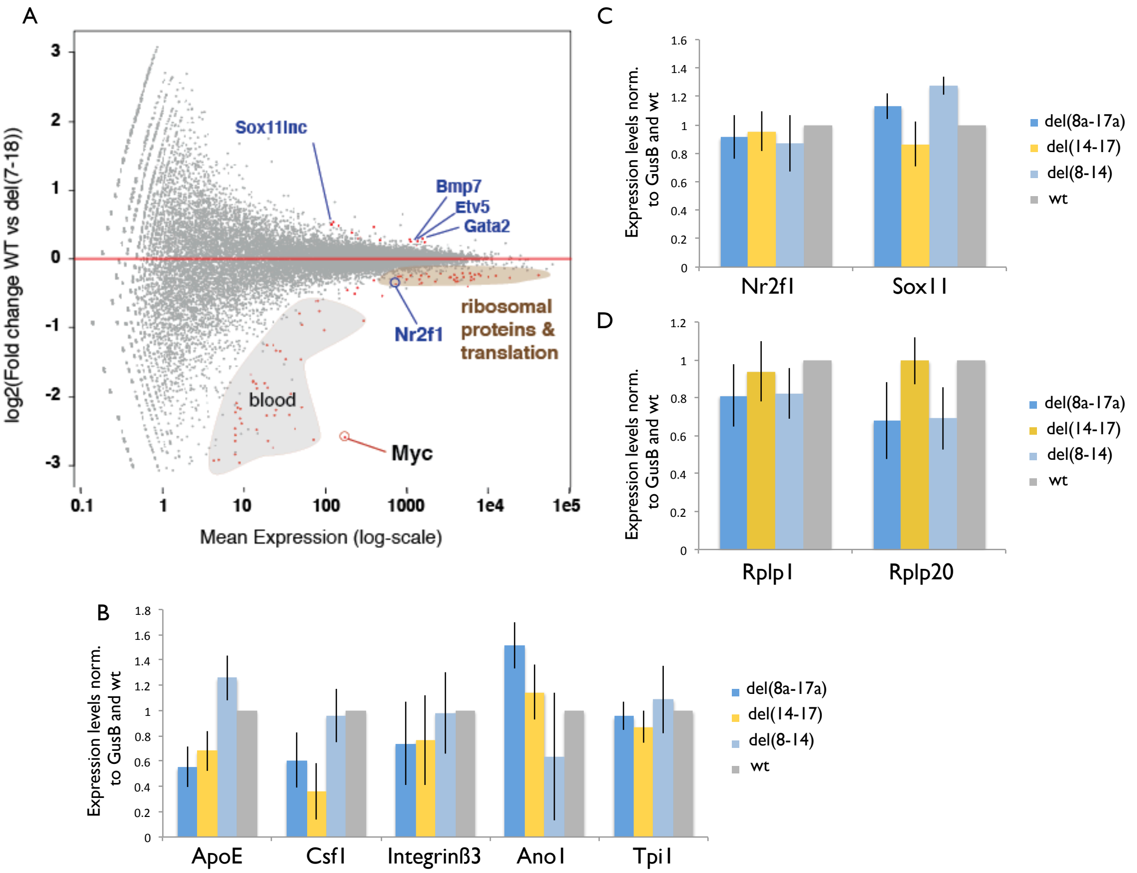


Figure29|RNAseq on e11.5 face of del(8a-17a) and wt: A) A global view of RNAseq experiment: In the y axis, fold change expression in del(8a-17a) when compared to wt samples is given in logarithmic scale. On the x axis, mean expression level of the genes are indicated. The genes differentially expressed (FDR:0.05) are marked by red dots. B) qPCR verification of the downregulation of ribosomal proteins. C) qPCR verification of deregulation of *Nr2f1* and *Sox11* transcription factors. D) qPCR verification of the hematopoietic lineage specific genes. All genes are normalized to *GusB* and in the y axis, expression levels relative to average wildtype expression is shown. Wt error bars are added up to the other genotypes by using the "error propagation" formula.

To have a global view on the cellular processes that the *c-Myc* downregulation influences, I used GOrilla (Eden E *et al* 2009) as a tool to find out enriched Gene Ontology(GO) terms among the genes differentially expressed between del(8a-17a) and wt at p<0.05. Biological process related "GO" terms were strongly enriched for hematopoiesis, metabolic processes and the regulation of metabolic processes. In terms of function related GO terms the only significant enrichment was the "structural constituent of ribosome". Ribosomal structures in the cytosol and in the non-membrane bound organelles were the only cellular component related GO terms that were significantly enriched (Supplementary Table2).

Two recent papers suggested that *c-Myc* acts as a transcriptional amplifier when it is overexpressed (Nie Z *et al* 2012, Lovén J *et al* 2012). Therefore, in order to make the RNAseq results quantitatively comparable, the samples were spiked with 72000 drosophila S2 cells. The RNAseq experiment showed that the proportion of drosophila mRNA to the total read counts was similar for each sample. This implies that *c-Myc* gene does not work as an amplifier in this context.

To sum up, RNAseq results show that *c-Myc* downregulation upon the deletions in the (8a-14a) region and (8a-17a) region, which contain MFM regulatory elements, lead to the deregulation of three major groups. First one is the deregulation of a morphogen and some transcription factors, which involve in craniofacial development. The second group is deregulation of the cellular metabolism, in which chemical stress response genes and metabolic enzymes take part. The third process is the downregulation of ribosome biogenesis and translational machinery pathways, which regulate cellular events including cell cycle progression, proliferation and metabolic stress response.

5.2.3 Locating Hematopoietic Lineage Specific Regulatory Elements in the *c-Myc/Pvt1* Flanking Locus

5.2.3.1 Phenotypic Consequences of the telomeric deletions in the *c-Myc/Pvt1* flanking locus

The genotype distribution of the embryos coming from heterozygous del(8a-17a) breedings show Mendelian distribution at e11.5, e14.5 and also at birth. However, the distribution of the pups coming from heterozygous del(8a-17a) breedings deviated from Mendelian ratios at the stage of weaning. The homozygous del(8a-17a) mice showed some post-natal mortality between the 1st week and the 3rd week after birth. The survivors with homozygous (8a-17a) deletion show clear size reduction when compared to wt littermates (Figure 30A). Weight measurements performed on e14.5 embryos and newborn pups do not indicate any difference. However, starting from the first post-natal day there is a clear reduction of growth rate in homozygous del(8a-17a) individuals when compared to wt littermates (Figure 30B). This indicates that the size difference that is observed in adults is not due to developmental delay, which is caught up in later stages, but it is a developmental retardation (Figure 30C). Size reduction is observed in heterozygous *c-Myc* deleted mice (Trumpp A *et al* 2001) but this does not cause post-natal mortality. Considering that the homozygous del(8a-17a) mice show greater *c-Myc* downregulation than heterozygous *c-Myc* gene deletion in certain tissues, one or more of these tissues that *c-Myc* involves in the development can account for post-natal mortality.

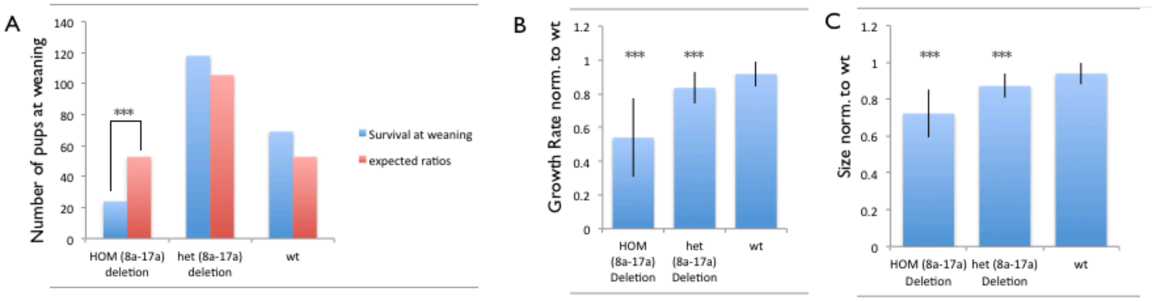


Figure30| Post-natal lethality, growth defect and size in the del(8a-17a): A)The homozygous mice for del(8a-17a) are significantly less then expected numbers. B) Growth rate refers to the weigh gained per day and the measurements are done between post-natal day one and day seven. Growth rate is significantly less in the first week of the new born pups homozygous or heterozygous for del(8a-17a) when compared to wt littermates. C) Size refers to the weights of the new born pups and it is measured from post-natal day one to day seven. The size is significantly smaller in new born pups homozygous or heterozygous for del(8a-17a) when compared to wt littermates.

Among other systems, we investigated the hematopoietic system as *c-Myc* has a vital role in hematopoiesis (Wilson A *et al* 2004) and RNAseq result indicate the presence of distant hematopoietic enhancers in the (8a-17a) region. Therefore in collaboration with Lisa von Paleske, a PhD student from Andreas Trumpp's lab, we investigated the influence of the *c-Myc/Pvt-1* flanking locus deletions on hematopoietic system. The contribution of each person to the experiments is explained in Materials and Methods section.

5.2.3.2 Investigating the Influence of the *c-Myc/Pvt1* Flanking Locus in Hematopoiesis

Due to the post-natal lethality of homozygous del(8a-17a) mice, we looked at the distribution of the hematopoietic cells in the bone marrows (BMs) of 10 day-old animals, where the distribution of the genotypes still exhibits Mendelian ratio. Bone marrows are extracted from a pool of bones, from limbs, hips, backbone and breastbone (femur, tibia, humerus, radius, ulna, ilium, columna vertebralis, sternum). The bone marrow cellularity decreased linearly with the body size reduction upon loss of (8a-17a) regions. Namely, there is no significant change in the ratio of total bone marrow cells when normalized to the body weight.

5.2.3.2.1 Effects of the deletions in the telomeric *c-Myc/Pvt1* flanking locus on the Hematopoietic Cell Distribution

Bone marrow cells were stained with the cell surface markers of hematopoietic lineage to analyze the distribution of the cells in the hematopoietic system via Fluorescence-Activated Cell Sorting (FACS). The cells were analyzed in three different categories: stem cell population, progenitor cell population, and differentiated cell population (Figure 4, Figure 31). The frequency of the cells with lineage markers (CD4, CD8, B220, Gr-1, CD11b, Ter119), which exclude stem cell (HSC) and progenitor cell population, did not change significantly between the wt mice and mice with del(8a-17a). However, the frequency of the Sca-1 positive, c-Kit positive LSK cells, which contain HSCs and multipotent progenitor cells (MPPs) cells (Wilson A *et al* 2008), were 6 times higher in homozygous del(8a-17a) when compared to wt littermates. Usage of two markers of the Slam family (positive for CD150 and negative for CD48) define a subpopulation of LSK cells, which contains short term hematopoietic stem cells (MPP1, positive for CD34) and long term hematopoietic stem cell population (LT-HSC, negative for CD34) (Kiel MJ *et al* 2005, Kim I *et al* 2006, Yilmaz OH *et al* 2006, Mikkola HKA and Orkin SH 2006, Osawa M *et al* 1996). The frequency of LSK Slam cells within the LSK population is lower in the del(8a-17a) mice when compared to wt littermates. However, the frequency of LSK Slam is slightly higher in the total bone marrow of the del(8a-17a) mice when compared to wildtype due to the abundance of LSK cells in del(8a-17a) mice (Figure

32A). Together with the changes in the frequency of LSK Slam cells, cell cycle stage of these cells moves from G_0 to G_1,, which indicates decrease in the self-renewal capacity (Figure 32B). LT-HSCs and MPP1 go through MPP2, MPP3/4 and committed progenitor cell stages, respectively, before they differentiate into certain lineages (Wilson A *et al* 2008). Strikingly, in del (8a-17a) mice, CD48+ MPP3/4 population extensively increases and the CD48- Multi-Potent Progenitors (MPPs) are completely depleted (Figure 32A). Noteworthy, heterozygous del(8a-17a) phenotype is mild when compared to wt. In brief, del(8a-17a) leads to accumulation of undifferentiated cells at different stages, particularly at the multipotent progenitor stage.

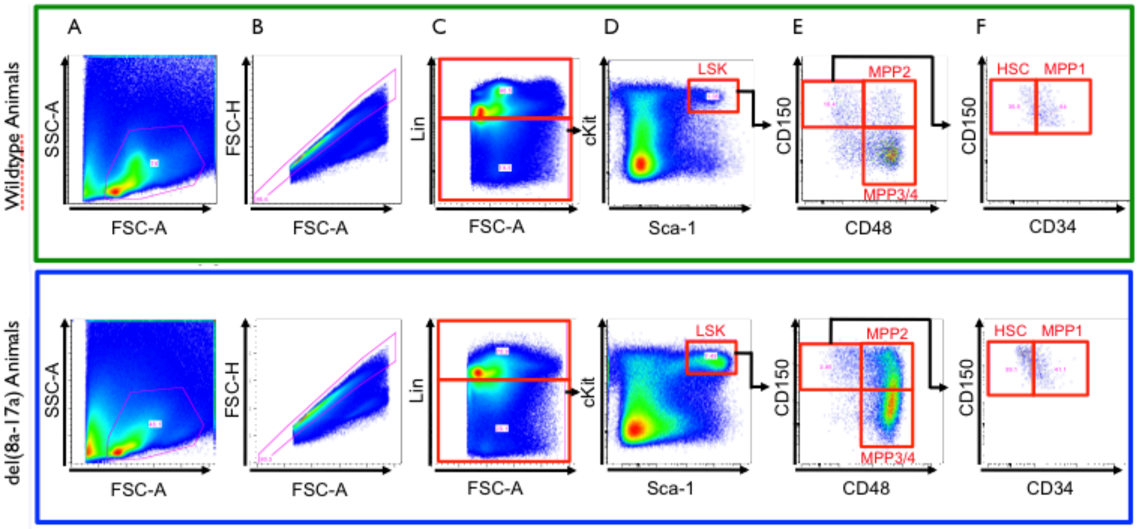


Figure31| Gating Scheme for FACS analysis of Undifferentiated Hematopoietic Cells: Total bone marrow cells are fluorescently labeled for the markers given in the axis and the given gating scheme routinely used in the Trumpp lab is applied. The top panel is an example from a 10 day-old wildtype sample and the bottom panel is an example from a 10 days old del(8a-17a) sample A) SSC-A stands for side scattering and it is affected by the inner complexity of the particle. FSC-A stands for forward scattering and it is affected by the particle volume. B) FSC-H (Height) and FSC-A(Area) are used to detect singlets. C) On the y axis Lin stands for lineage. The main differentiated cells are labeled with the same fluorophore (CD4, CD8, B220,Gr-1, CD11b, Ter119). Therefore, Lineage negative cells contain undifferentiated cell populations. D) LSK cells are negative for Lin and positive for Sca-1 and c-Kit. E) LSK-Slam population, MPP2 and MPP3/4 are gated from LSK. F) Long-Term HSCs and MPP1 population are gated from LSK-Slam population.

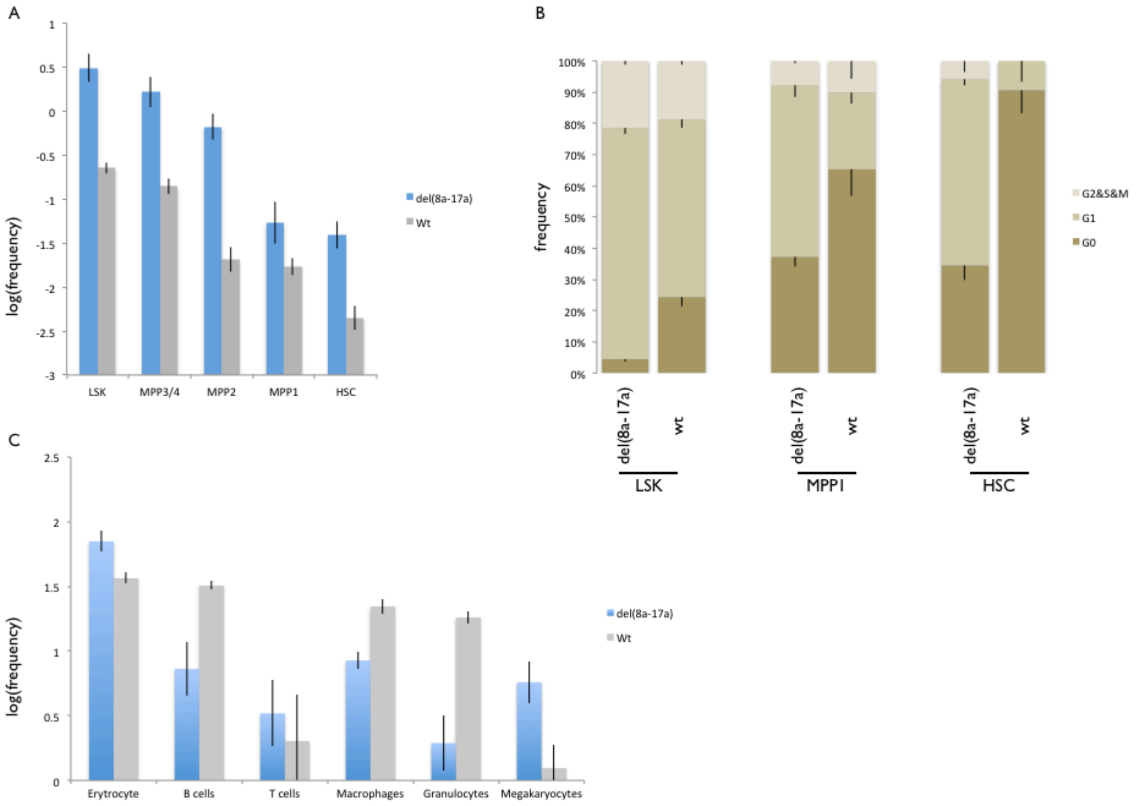


Figure32|FACS Analysis of Hematopoietic Lineage in Bone Marrow of del(8a-17a) mice: A) Frequency of each undifferentiated cell population in the bone marrow is given in logarithmic scale for del(8a-17a) and wt littermates. B) Cell cycle analysis by Ki67-Hoechst staining has been done for LSK, MPP1 and HSC population for del(8a-17a) and wt littermates. Upon deletion of the (8a-17a) interval, quiescent cells (G0) become activated (G1). C) Frequency of the differentiated hematopoietic cell population in the bone marrow is given in logarithmic scale for del(8a-17a) and wt.

The accumulation of HSCs and MPPs implies changes in the differentiation dynamics of the hematopoietic stem cells. Therefore, we performed FACS analysis on terminally differentiated cells of the hematopoietic system. In del (8a-17a) mice, the frequency of the megakaryocyte and erythrocyte population expands more than two times compared to the wildtype levels, whereas there is a clear reduction in B cell, macrophage, and granulocyte population (Figure 32C). There were not enough CD4+ or CD8+ T cells to compare T-cell levels at day 10 mice. The “LS minus K” (LS-K) population, which is negative for Sca-1 and lineage markers but positive for c-Kit, contains committed myeloid progenitors that originate from MPP2, MPP3/4 subpopulation of LSK cells. LS-K cells give rise to terminally differentiated hematopoietic cells. Unlike MPP2 and MPP3/4 populations and the differentiated hematopoietic cell population, the size of LS-K population did not change in del(8a-17a) animals. By analyzing the expression of CD34, CD127 (IL7Rα), and CD16/32, we subdivided LS-K into Common Myeloid Progenitor (CMP), Granulocyte-Macrophage Progenitors (GMP), Megakaryocyte-Erythrocyte Progenitors (MEP) (Akashi K *et al* 2000). Furthermore, Common Lymphoid Progenitors (CLyP) were identified immunophenotypically by low expression of Sca-1 and c-Kit and high expression of CD127 (Kondo M *et al* 1997). Among these progenitors only CLyP showed a significant increase in del(8a-17a), whereas the other progenitors, which originate from LS-K showed identical distribution in del(8a-17a) as their wt littermates.

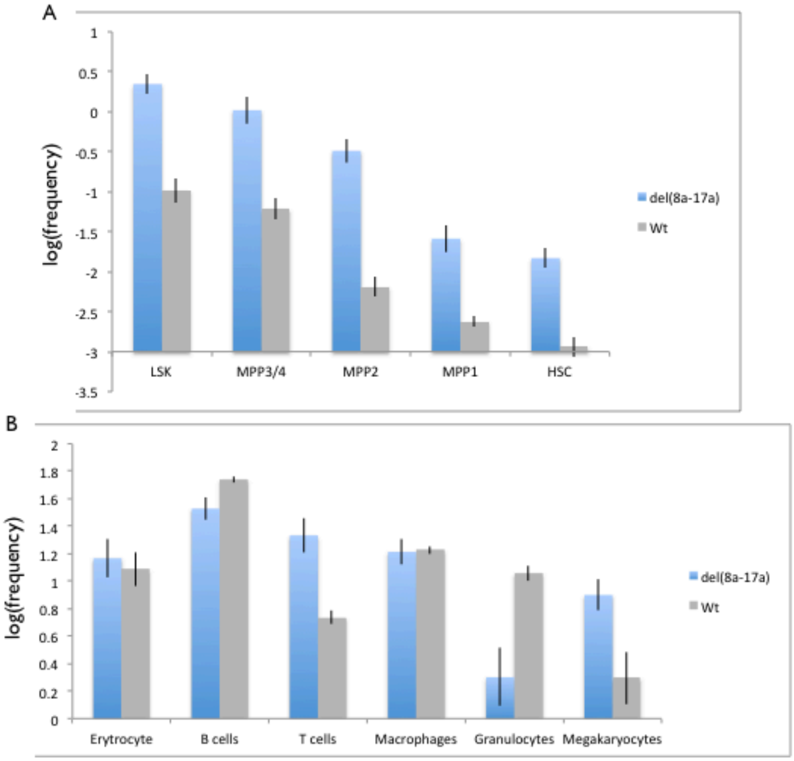


Figure33| FACS Analysis of the Hematopoietic System in Liver of del(8a-17a) mice: A) Frequency of each undifferentiated cell population in the liver is given in logarithmic scale for del(8a-17a) and wt
B) Frequency of the differentiated hematopoietic cell population in the liver is given in logarithmic scale for del(8a-17a) and wt.

The liver is the main organ for hematopoiesis in embryos, whereas, in the adults bone marrow takes over the maintenance of hematopoietic system (reviewed in Dzierzak E and Medvinsky A, 1995). In order to see, this transition from fetal liver to bone marrow is affected in the del(8a-17a) mice, we analyzed hematopoietic lineage in the liver of 10 day-old mice. The frequency of LSK cells in homozygous del(8a-17a) liver was 20 times higher than in wt littermates (Figure 33A). Just like in bone marrow, the frequency of B cell and granulocytes in the liver also shrank and the frequency of megakaryocytes population increased in del(8a-17a) liver (Figure 33B). On the other hand, macrophage and erythrocyte population did not significantly change between del(8a-17a) liver and wt littermates. Surprisingly, a statistically significant increase is observed in the T-cell population upon (8a-17a) deletion (Figure 33B). In brief, the strength of change in B-cell, granulocyte, and megakaryocyte population is different between bone marrow and liver but the trend of the deviation is exactly the same. Additionally, the HSC phenotype in the del(8a-17a) mice is similar but stronger in liver when compared to bone marrow (Figure 33A).

In order to narrow down the interval that influences the hematopoietic system in the (8a-17a), we used (14a-17a) deletion. Since this deletion did not show any post-natal lethality and growth defect, we did the analysis on 22 week-old adults. FACS analysis of (14a-17a) deletion animals showed an accumulation of LSK cells, somewhat in LT-HSC stage and extensively in the multipotent progenitors (MPP2 and MPP3/4) (Figure 34A). In terms of differentiated cells, erythrocyte and megakaryocyte populations increased in the del(14a-17a) bone marrows, whereas, B cell, macrophage, and granulocytes populations shrank (Figure 34C). When compared to del(8a-17a) mice, there is a slight increase in the frequency of the LS-K population but the LS-K derived progenitor cells(CMP, GMP, MEP) did not change in del(14a-17a) mice. In addition, a more pronounced increase is observed in Common Lymphoid Progenitors (CLyP), which give rise to B cells and T cells (Figure 34B). Intriguingly, the ratio of Lin+/Lin- cells changed dramatically from 70%/30% to 30%/70% upon the deletion of the (14a-17a) interval.

The overall changes in the hematopoietic system in terms of HSC and differentiated population upon (14a-17a) deletion shows that the defects in the hematopoietic lineage are not linked to the post-natal mortality and body size.

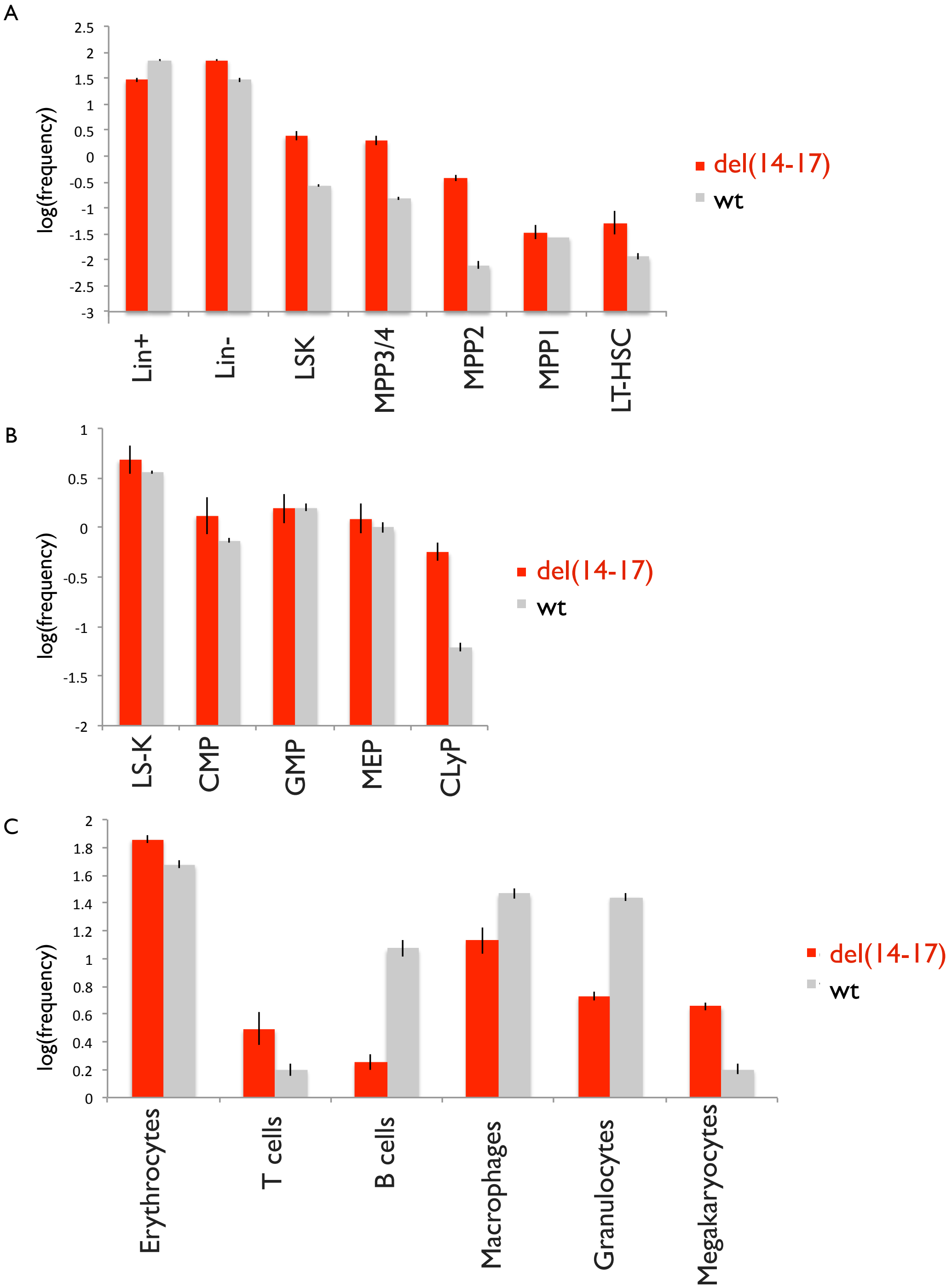


Figure34|FACS Analysis of Hematopoietic Lineage in Bone Marrow of del(14-17) mice: A) Frequency of each undifferentiated cell population in the bone marrow is given in logarithmic scale for del(14-17) mice and wt littermates. B) Frequency of each committed progenitor cell population in the bone marrow is given in logarithmic scale for del(14-17) mice and wt littermates. C) Frequency of the differentiated hematopoietic cell population in the bone marrow is given in logarithmic scale for del(8a-17a) mice and wt.

5.2.3.2.2 Targets of Hematopoietic Lineage Specific Enhancers

FACS analysis showed a clear link between the (14a-17a) region and hematopoiesis. First of all, we checked whether there is regulatory input for the hematopoietic cells in *c-Myc/Pvt1* flanking locus. Undifferentiated hematopoietic cells, particularly LT-HSC and MPP1,2,3/4 population from the mice homozygous or heterozygous for 17a insertion, showed lacZ staining. However, in del(8a-17a) mice, the size of the these cell populations, which expressed lacZ, severely dropped (Figure 35). In order to find out the target gene(s) of the (8a-17a) regulatory region, which causes hematopoietic deregulation upon deletion, we sorted LSK and LS-K cells from del(8a-17a) and wt animals. Among the other genes in the *c-Myc/Pvt1* flanking locus, *c-Myc* is the only gene downregulated upon deletion of (8a-17a) in LSK cells and its expression drop to 3% of wt levels in LSK cells. In contrast, *c-Myc* expression in LS-K cells, which are not affected in terms of frequency upon deletion, does not change (Figure 36). Therefore, extremely distant cell-type specific regulatory element(s) in the (8a-17a) region at the telomeric end of *c-Myc/Pvt1* flanking locus act on *c-Myc.*

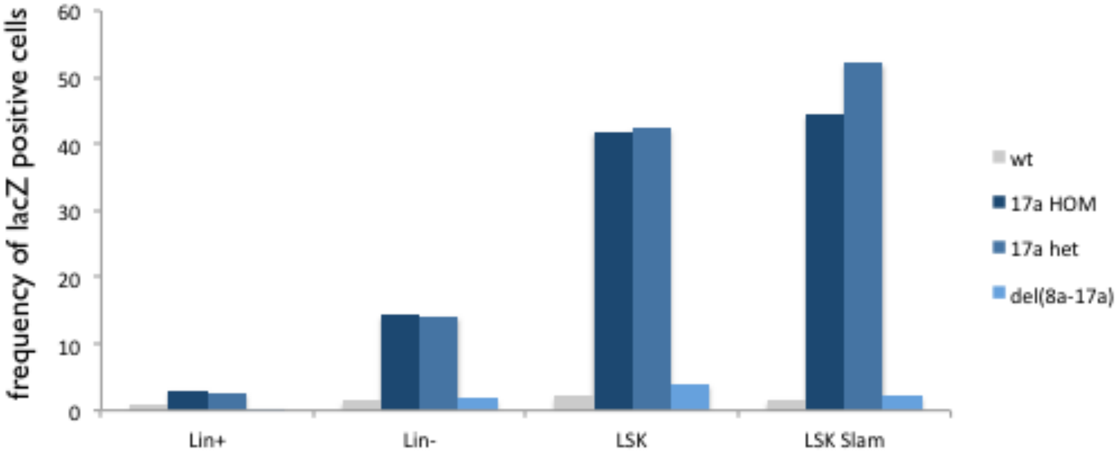


Figure35|Regulatory sensor activity in the *c-Myc/Pvt1* flanking locus in the hematopoietic cells: The graph shows the frequency of lacZ positive cells in the given population by using FDG substrate. LSK-Slam is gated from LSK and LSK population is gated from Lin-. The experiment is done on single adults homozygous for 17a insertion, heterozygous for 17a insertion, het for del(8a-17a) and wildtype control.

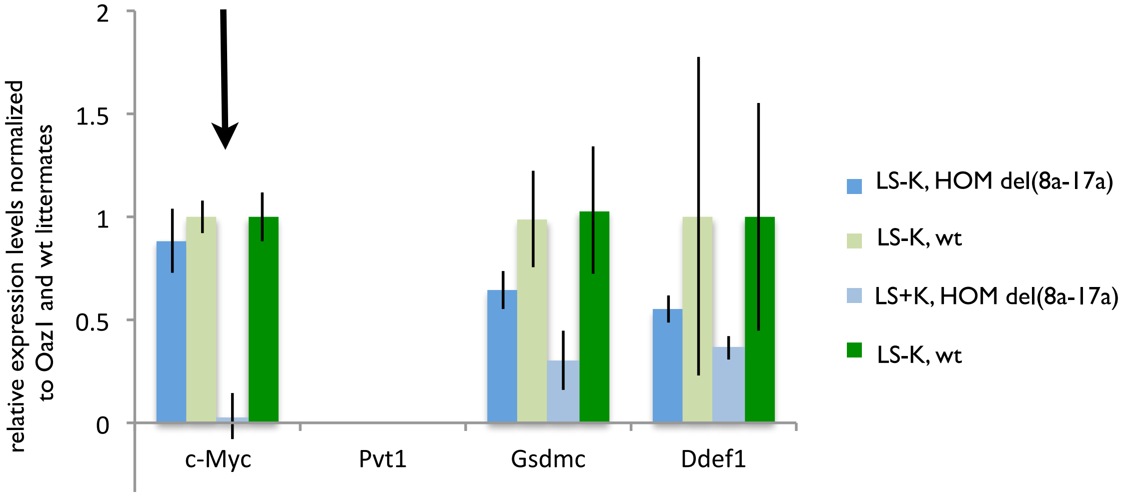


Figure36|Real-Time qPCR on the *c-Myc/Pvt1* flanking locus: LS-K and LSK population are sorted from wt and homozygous del(8a-17a) samples. qPCR is performed for the given genes in the locus. All the genes are normalized to Oaz1 and relative expression values are shown in the chart with respect to the average wildtype values. Pvt1 is not detected. Gsdmc level is very low. *c-Myc* expression in LSK population in del(8a-17a) drops down to 3% of the wildtype levels (arrow).

Cell-type specificity of the regulatory elements in this 400kb window indicates the involvement of hematopoietic lineage specific transcription factors in enhancer activity. We looked at the transcription factor binding profiles in hematopoietic cells from Bertie Göttgens' lab and histone profiles from Bing Ren's lab (Wilson NK *et al* 2010; Shen Y *et al* 2012). In the (8a-17a) region, Göttgens' data indicate two regions defined by clustered occupancy of six hematopoietic lineage specific transcription factors. These sites overlap with the active histone marks in Bing Ren's data set obtained from adult bone marrows. Besides, there are 4 additional conserved sites, which are associated with active histone marks (Figure 37, upper panel). Massimo Petretich showed that these regions are also associated with active marks in LSK cells. Therefore, these sites at the very telomeric end of *c-Myc/Pvt-1* flanking locus are candidates for regulatory elements that act on *c-Myc* in LSK cells (Figure 37, qPCR panel).

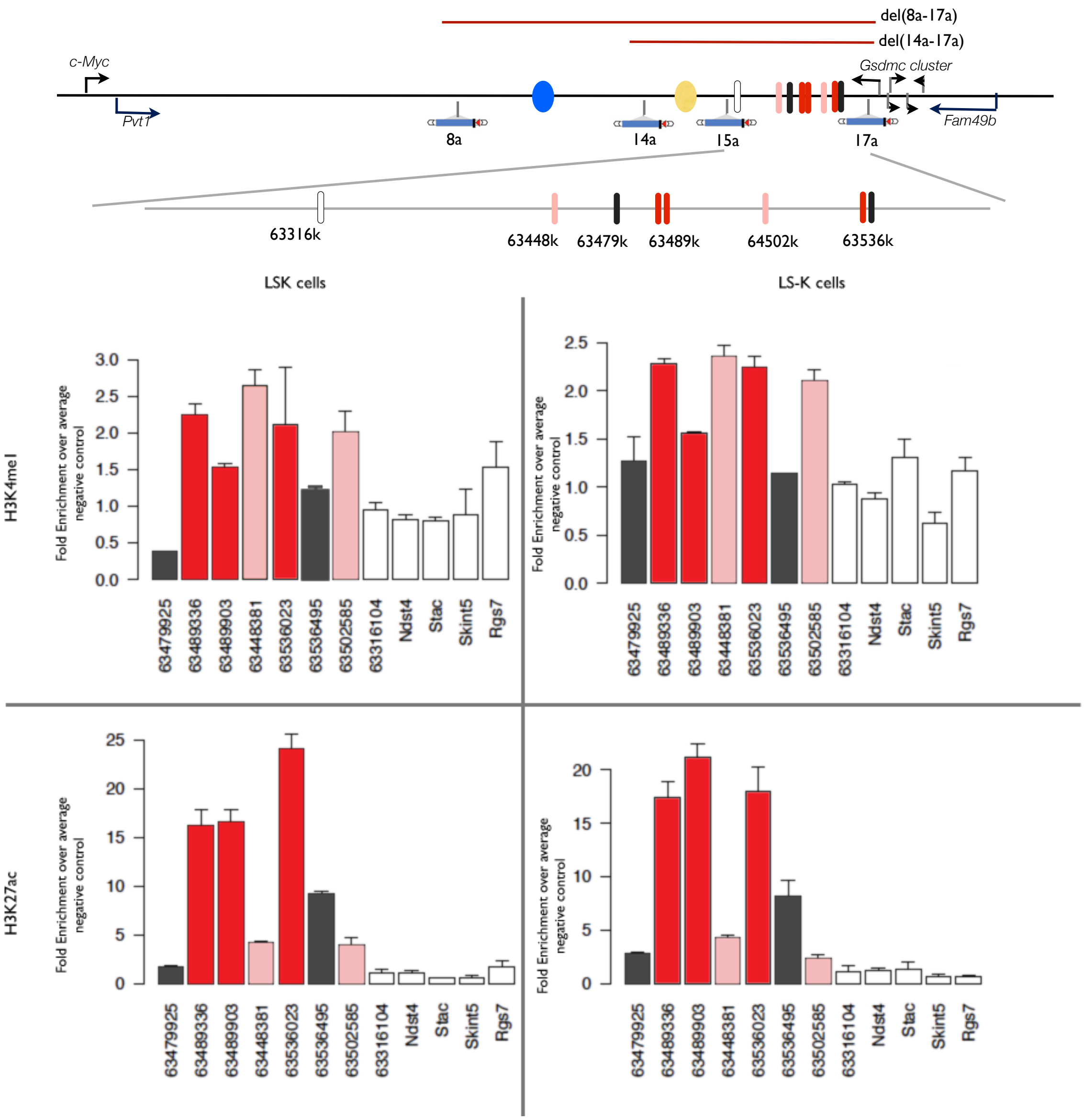


Figure37|ChIP for enhancer marks in LSK and LS-K cells: 7 candidate regions are tested together with 5 negative controls for H3K4me1 and H3K27ac marks (raws), in LSK and LS-K cells of wildtype mice (columns). The positions of the 7 candidate regions are shown in the *c-Myc/Pvt1* locus above. Fold enrichment over input values are normalized to average negative control enrichment in the y-axis. 3 of the candidate regions (red) show strong enrichment for both H3K4me1 and H3K27ac. 2 of the regions (pink) have strong H3K4me1 enrichment and mild H3K27ac enrichment. 2 of the candidate regions (grey) did not show enrichment for H3K4me1. qPCRs are performed by Massimo Petretich.

5.2.4 Consequences of the *c-Myc/Pvt1* Flanking Locus Duplications in Hematopoietic System

The rearrangements in the telomeric side of the *c-Myc/Pvt-1* flanking locus cause hematopoietic defects in mouse. In humans, the translocations and the duplications in the telomeric side of *c-Myc/Pvt-1* flanking locus are associated with deregulation of the hematopoietic system. Therefore, we investigated the effects of these human duplications in mouse models. The first mouse model had a duplication of (15a-17a) region, which is orthologous to the duplicated regions in pediatric Acute Myeloid Leukemia (AML) patients (Radtke I *et al* 2009). The second mouse model had a large duplication, which contains *c-Myc* gene and its 3Mb-long genomic surrounding (MycDup3MB) (Figure 38, upper panel). FACS analysis of hematopoietic cells from adult mice showed no major defect except a slightly elevated macrophage and CD8+ T-cell population in MycDup3MB. I have performed qPCR analysis on the CD150+ CD48- (HSC and MPP1) and CD150+ CD48+ (MPP2) populations, which are sorted from MycDup3MB adult mice (Figure 38A-B). Intriguingly, *c-Myc* levels increased in the MPP population but not in the LSK-Slam population (Figure 38C).

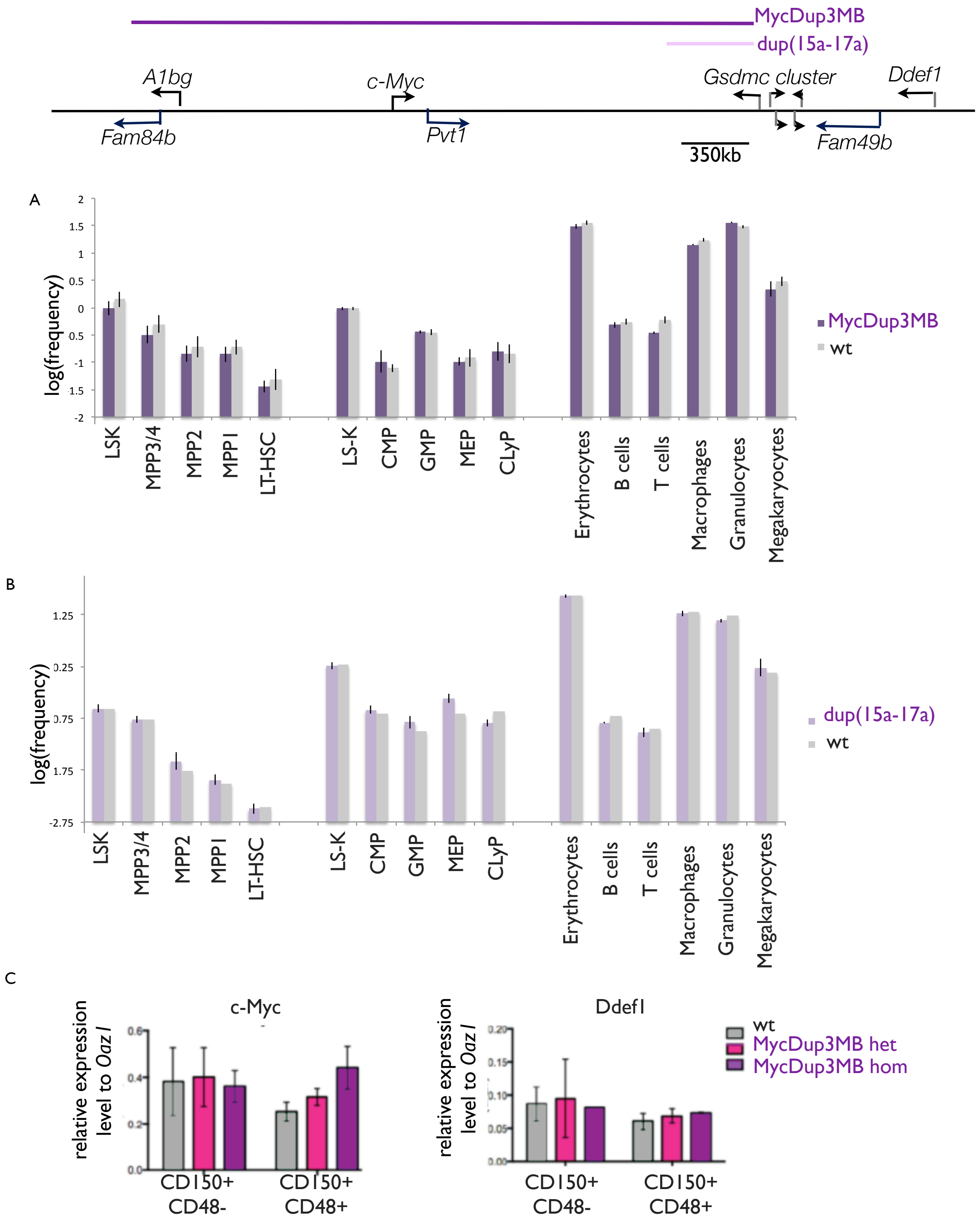


Figure38|The analysis of the duplication in the *c-Myc/Pvt1* flanking locus: The duplications are shown in the sketch above A) FACS analysis of the hematopoietic lineage in MycDup3MB and wt littermates. In the y-axis the frequency of the cell types are given in logarithmic scale. B) FACS analysis of the hematopoietic lineage in dup(15a-17a) and wt littermates. In the y-axis the frequency of the cell types are given in logarithmic scale. C) qPCR analysis of the hematopoietic samples for *c-Myc* (in the duplicated interval) and Ddef1 (outside of the duplicated interval). Expression is normalized to *Oaz1* and fold changes are represented in the y-axis.

To sum up, I have identified extremely distant regulatory regions, which exclusively act on *c-Myc* gene in embryonic face and hematopoietic cells. In these regulatory regions, multiple sites with active biochemical marks are present. In the development of face, deletions of the embryonic-face specific regulatory elements cause mild but significant reduction in the face morphology, whereas severe deregulation of hematopoietic system is observed in HSy-specific regulatory regions upon these deletions.

5.3 Role of Genome Organization in Spatiotemporal Regulation of *c-Myc*

In the *c-Myc/Pvt-1* flanking locus, I have described distant regions, which are important for face development or hematopoiesis. These regions enhance the transcription from their target promoters in a tissue specific manner from exceptionally long distance. For example the embryonic face enhancer acts on *c-Myc* gene, which is 1mb far on the centromeric side and this enhancer shows a very strong regulatory activity on the telomeric end of the locus, which is also 1mb far. However, it does not act on the 10a insertion, which is one of the two closest insertions to this regulatory region. The shape of the *c-Myc* regulatory domain for the embryonic face expression indicates that spatiotemporal regulation of gene expression is heavily dependent on the range of distant regulatory elements. In order to understand the molecular mechanisms that regulate "long-distance" communication between regulatory sequences and their target promoters *c-Myc*/Pvt-1 locus provides us an experimental model system to investigate how this is achieved. The rearrangements in this locus allow us to interfere or alter distances, synteny or boundaries and analyze the impact of the redistribution of enhancer activities on target gene expression.

5.3.1 The influence of the distance and genomic context on the genome organization

Although the regulatory landscape clearly demonstrates that the distribution of the regulatory activity is non-linear and discontinuous, it is possible that the enhancer activity is down only at some critical distances. In order to understand the contribution of "distance" as a physical parameter to the distribution of enhancer activity, I used four deletions and duplications by using three breakpoints at 10a, 13a, 20a and 21a (Figure 39). 10a is completely silent at day 11.5, but 13a insertion shows strong MFM expression and NE expression. 20a insertion shows characteristic wide-spread expression pattern and 21a insertion shows very weak staining in the tail ectoderm. *c-Myc* enhancers are located between 10a and 20a insertions and these enhancers cannot act on the regions more telomeric than 17a and *Fam49b* region enhancers cannot act more centromeric sequences than 17a. Upon (10a-20a) duplication, the distance between *c-Myc* enhancers and 10a insertion does not change. However, the regulatory input starts acting on duplicated 10a position, despite being at the same distance. On the other hand, upon (10a-20a) deletion, the widespread expression is maintained at 10a position. The deletion rules out possible local silencing due to centromeric side of 10a insertion (Figure 39, lower panel). This example shows that regulatory activity at an insertion site is independent of its distance to the enhancer and it indicates that the genomic context dictates the distribution of regulatory information.

Upon duplication of the (10a-21a) region, the sequence composition between facial enhancer and 10a insertion is exactly the same with the (10a-20a) duplication. However, unlike (10a-20a) duplication, the (10a-21a) duplication is completely blank. A different combination of telomeric and centromeric sequences blocked the regulatory input from the same distance. 13a insertion lies in between MFM regulatory region and NE regulatory region. (13a-20a) deletion results in superimposition of MFM expression and 20a-like widespread expression. On the other hand, (13a-20a) duplication gives a very clear NE expression (Figure 39, upper panel). These examples suggest that the distribution of regulatory activity can be different at the same distance from the enhancer depending on the genomic context.

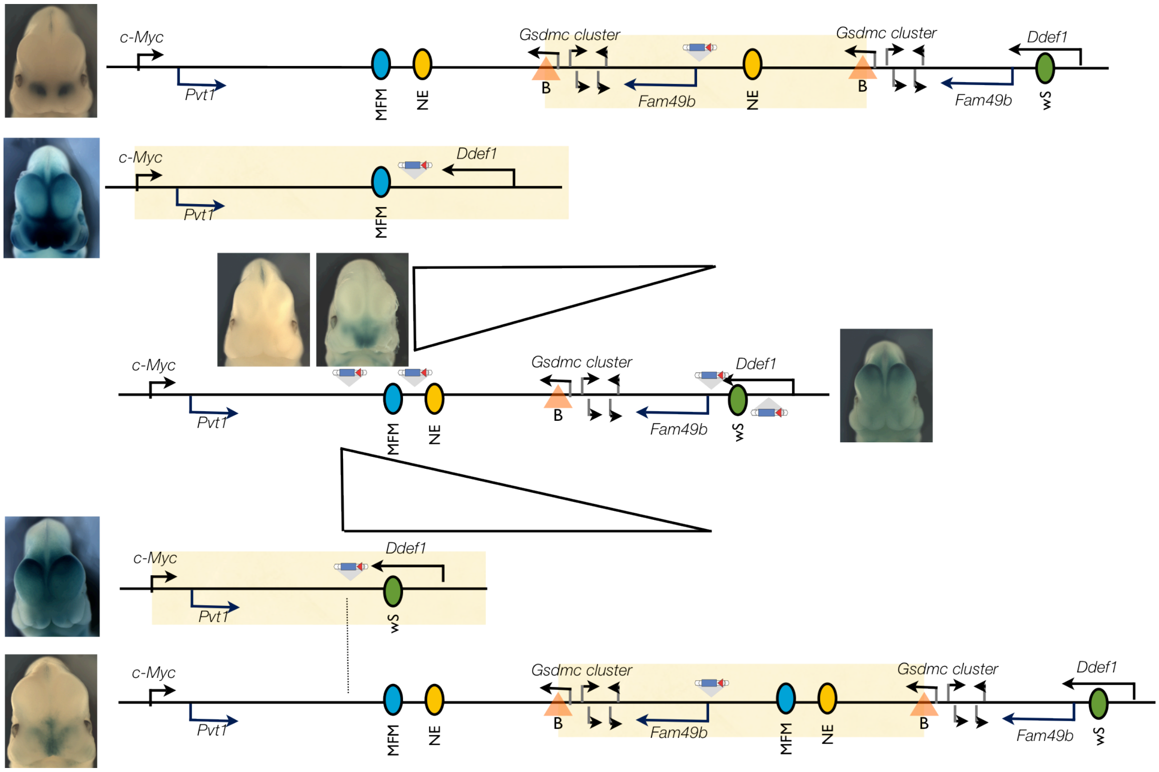


Figure39|The influence of genomic context on enhancer-promoter communication: The drawing in the middle represents the insertions and corresponding lacZ staining. MFM in blue circle represents the medial facial enhancer, NE in yellow circle represents the nasal epithelia enhancer, B in orange triangle represents boundary the interval between two different regulatory landscapes. Green circle is the enhancer active around *Fam49b* locus. Highlighted intervals mark the regions between two known boundaries.
The top two panels show the lacZ staining and the genomic landscape of the duplication(very top) and the deletion (second raw) between the (10a-20a) interval. The bottom two panels show the lacZ staining and the genomic landscape of the duplication (Last raw) and the deletion (4th raw) between the (10a-20a) interval.

5.3.2 Impact of Boundary Regions in Enhancer-Promoter Communication

There is a transition zone in between the *c-Myc/Pvt-1* flanking locus regulatory landscapes and the *Fam49b* regulatory landscape, which contains the *Gsdmc* cluster. None of the enhancers in the *c-Myc/Pvt-1* flanking locus or in the *Fam49b* locus extend over this transition region. The molecular factors that determine the extent of the enhancers are not known. Classical insulator activity fits with enhancer blocking effect of the transition region. However, it is not known whether there is a specialized boundary element that separates regulatory landscapes or whether transition region is a consequence of sum of interaction on its telomeric and centromeric side. With the help of summer student Alicia Lardennois, we have tested the function of boundaries on determination of regulatory landscapes:

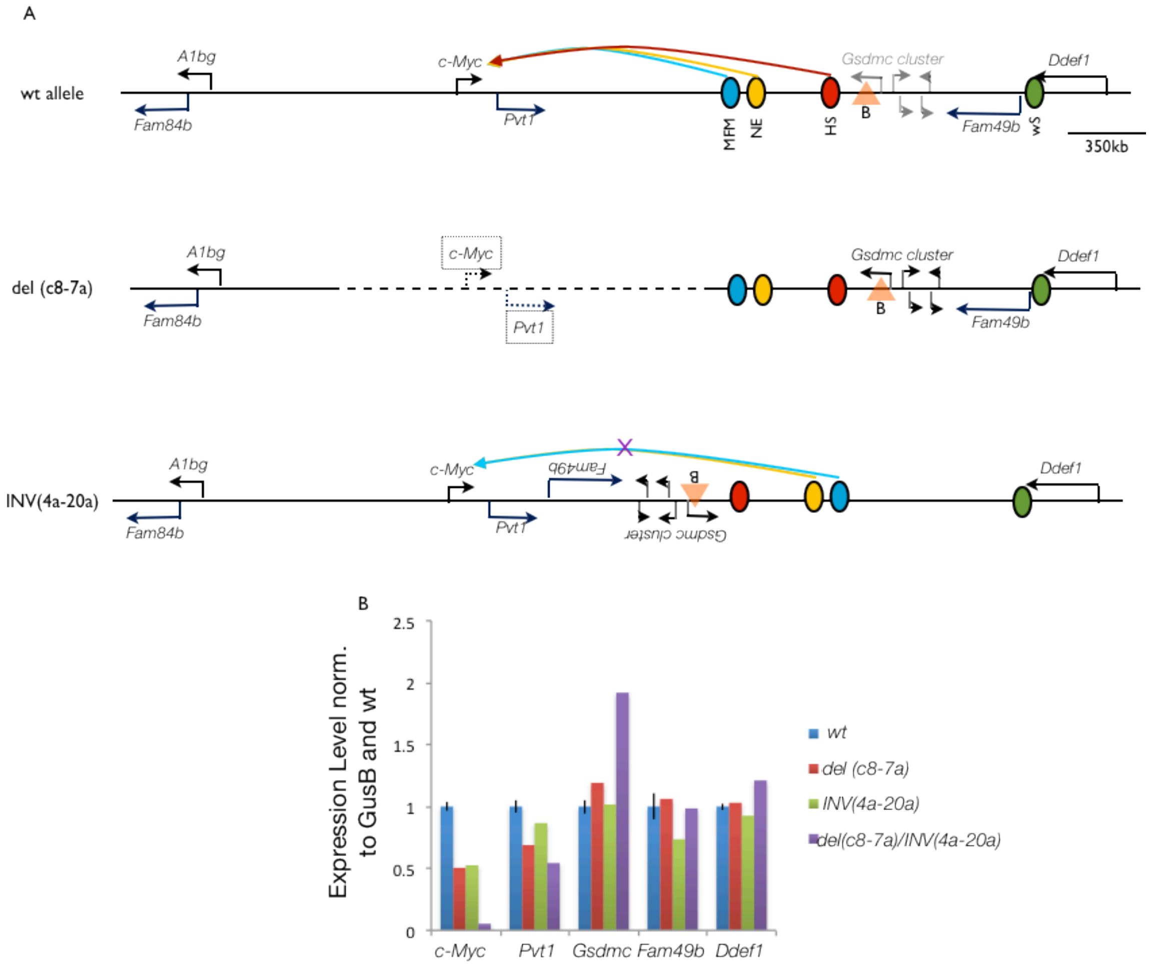


Figure40|Inversion in the *c-Myc/Pvt1* flanking locus: A) wt allele has three regulatory regions identified (MFM in blue, NE in yellow, HSy in red) in the locus one outside of the locus (wS-widespread in green). In del(c8-7a), all of the regulatory regions are intact but the *c-Myc* gene is deleted. In the inversion, boundary region around the *Gsdmc* cluster is brought between the enhancer regions and the *c-Myc* gene. The distance change between the MFM or NE enhancer and *c-Myc* gene upon inversion is less than 100kb. B) Expression levels obtained by qPCR are normalized to *GusB* and relative expression values with respect to wt average is shown in the y-axis.

I generated an inversion allele with breakpoints in *c-Myc/Pvt-1* flanking locus and in *Fam49b* locus (Figure 40A). Upon this inversion {INV(4a-20a)}, the boundary region between *c-Myc* and *Fam49b* locus regulatory landscapes is brought in between MFM enhancer and it's target gene, *c-Myc* without causing any major change in the distance between the region and the *c-Myc* promoter (<100kb). In this inverted allele, *c-Myc* expression dropped down to the same level where MFM enhancer was deleted (Figure 40B). Despite the presence of an intact MFM enhancer, it cannot communicate with *c-Myc* anymore due to the presence of this boundary region in between. LacZ staining of the inverted allele suggests that the MFM enhancer is still active, but the distribution of the enhancer activity is different. Considering that this inversion only changes the order of the regulatory elements and the organization of the locus, it shows that synteny of the locus and the relative position of boundary region with respect to the enhancer and target promoter changes the enhancer-promoter interaction.

5.3.3 Influence of Regulatory Landscapes on Endogenous Gene Expression

5.3.3.1. Imprinting in the *c-Myc/Pvt1* Flanking Locus

Our analysis of the landscape showed that the enhancers were acting very strongly on the reporters inserted at the telomeric end of the *c-Myc/Pvt-1* flanking locus. For the landscape analysis, all of the embryos used for lacZ staining came from the cross between a male with an insertion and a wt CD1 or C57/BL6 female (Figure 41A).

However, when the insertions around the telomeric 17a insertion site are inherited from the mother, the lacZ expression completely disappears (Figure 41B). This difference between the expression of lacZ in the paternal copy and the maternal copy suggested an imprinting phenomenon. As in the classical imprinted loci, the imprinting is not carried over the males, which inherited the imprinted reporter from the mothers. Therefore, *c-Myc/Pvt-1* flanking locus shows indications of classical imprinting and is a good model to investigate the influence of regulatory input distribution on endogenous gene activity.

5.3.3.1.1 The Extent of Imprinting

The lacZ staining on the maternal copies of 8a, 14c and 15a insertions was identical to the lacZ expression pattern from the paternal copy. On the other hand, none of the insertion lines closer than 50kb to 17a insertion showed lacZ staining in the maternal allele (Figure 42A). The lack of expression in the maternal allele around

17a insertion site cannot be a consequence of enhancer inactivation in the maternal allele, as the regulatory input does not change in the 8a, 14c, and 15a insertion sites between two alleles (Figure 42B). Therefore, the imprinting is only affecting a short-range region around the 17a insertion site and the range of the MFM enhancer is shortened on the telomeric end of the locus.

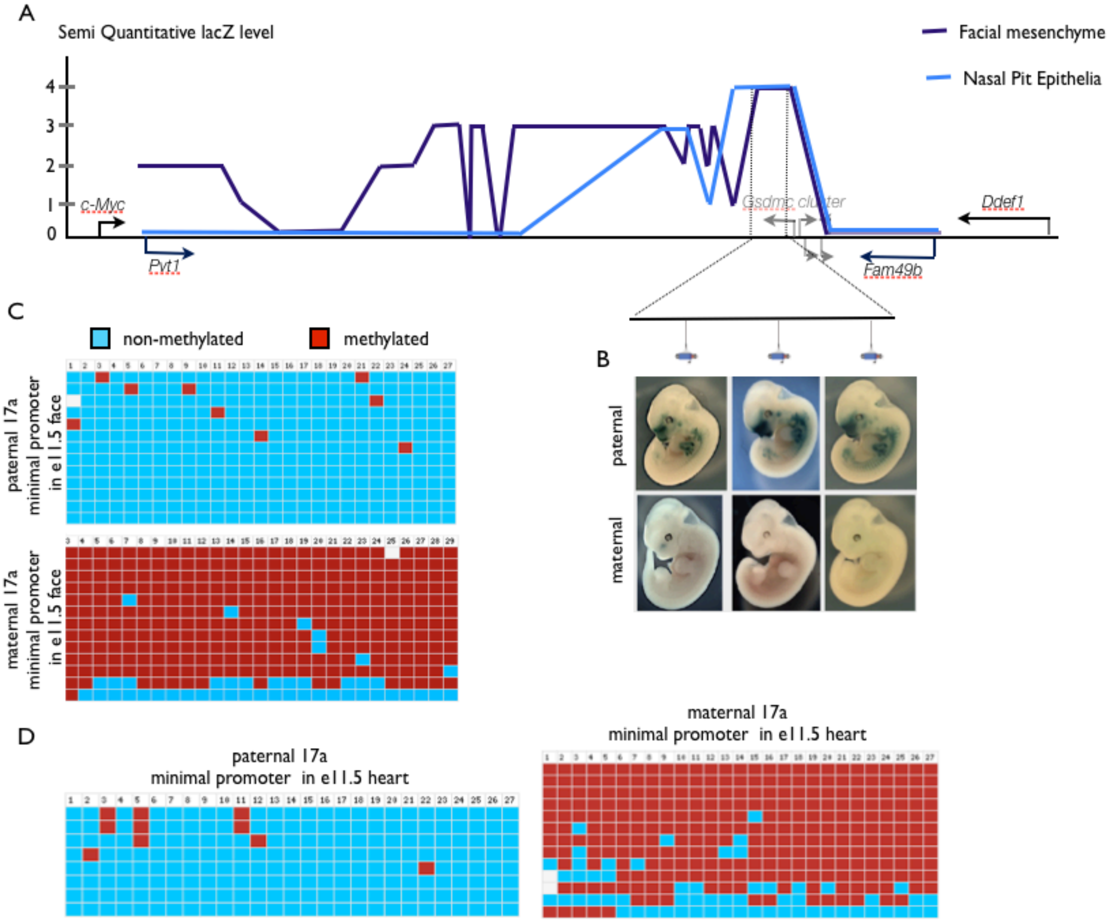


Figure41| Imprinting in the *c-Myc/Pvt1* flanking locus: A) The shape of the regulatory landscapes in the medial face mesenchyme and the nasal epithelia is given for paternal allele. B) The lacZ staining of the maternal and paternal copies of the same insertions in the telomeric end of the locus. C) Bisulfite sequencing of the minimal promoter in the transposon in e11.5 face when it is paternally and maternally inherited D) Bisulfite sequencing of the minimal promoter in the transposon in e11.5 heart when it is paternally and maternally inherited.

5.3.3.1.2 Allele Specific Methylation of Minimal Promoter

One of the landmarks of imprinted loci is methylation pattern. Bisulfite assay, which reflects the CpG methylation of region of interest, showed that the minimal promoter in the maternal allele is heavily methylated, whereas, the methylation event is scarce in the minimal promoter coming from the father (Figure 41C). In addition, I checked the methylation in the promoter of lacZ insertion in the 17a insertion site in e11.5 heart tissue, where regulatory sensor is not expressed neither in the paternal nor in the maternal allele: The paternal copy of minimal promoter was completely unmethylated, whereas the maternal copy was completely methylated in heart (Figure 41D). It shows that the methylation pattern is independent from the expression state of the minimal promoter but depends on the paternal origin of the allele.

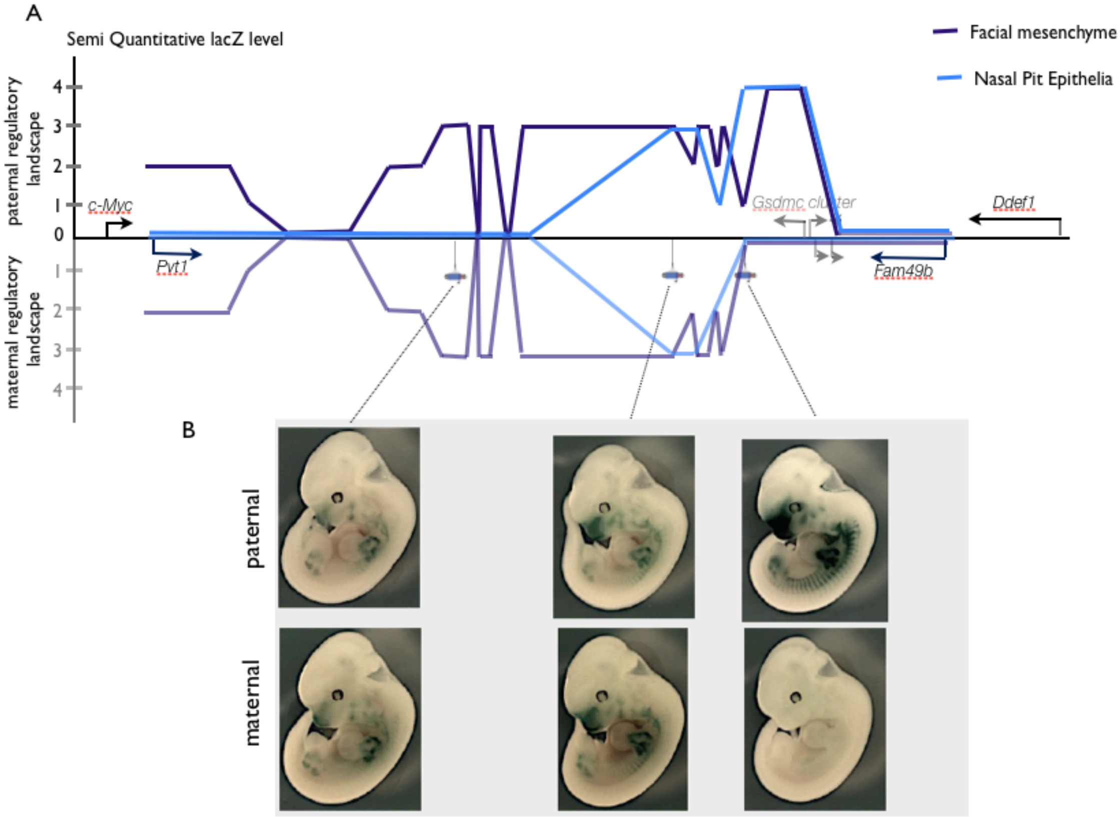


Figure42|Regulatory Landscape in the maternal allele: A) The shapes of the regulatory landscapes in the MFM and NE are given for paternal allele(top) and maternal allele(bottom). The only difference detected so far is at the telomeric end of the locus. B) Paternal and maternal lacZ staining of two insertions (first and second column) outside of the telomeric end of the locus and one insertion at the telomeric end of the locus (last column).

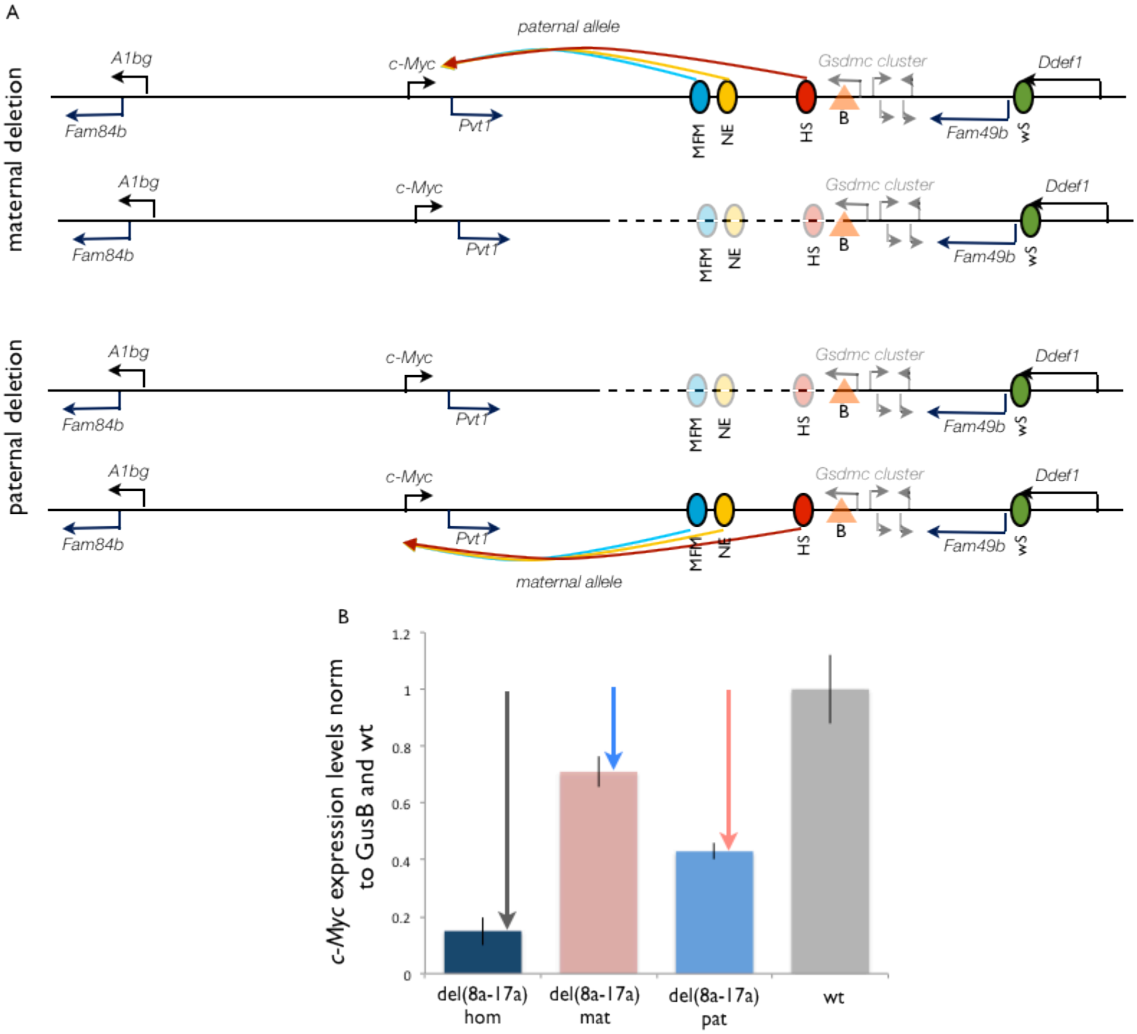


Figure43| Allele Specific Real-Time qPCR on del(8a-17a) e11.5 face samples: Same elements in Figure41 are used to show the genomic landscape of the locus. [B] indicates the boundary region. A) Maternal deletion shows that the enhancers in the paternal allele are active and the paternal deletion shows that the enhancers in the maternal allele are active. B) qPCR results for *c-Myc* expression normalized to the *GusB* and the fold change is indicated in the y-axis with respect to average wt *c-Myc* expression in e11.5 face for the given genotypes. The black arrow indicates the missing expression in the absence of regulatory region in both alleles. Pink bar indicates the present *c-Myc* expression and the blue arrow indicates the missing *c-Myc* expression in the absence of the maternal allele. Blue bar indicates the level *c-Myc* expression and the pink arrow indicates the missing *c-Myc* expression in the absence of the paternal allele.

5.3.3.1.3 Effect of Local Imprinting on the *c-Myc* expression

The maternal and paternal alleles of the *c-Myc/Pvt-1* flanking locus are subject to the very same cellular environment and protein composition but the regulatory landscape of the embryonic face is different between two alleles. The local imprinting at the telomeric end of *c-Myc/Pvt-1* flanking locus provided an experimental platform to understand if the organization of the regulatory landscapes influences endogenous gene expression. To distinguish parental contribution of MFM enhancer to endogenous *c-Myc* gene, I used heterozygous (8a-17a) deletions. The cross between heterozygous (8a-17a) deletion males and wt females are used to address the paternal contribution, whereas the cross between wt males and heterozygous (8a-17a) deletion females indicated the maternal contribution of MFM enhancer to *c-Myc* expression (Figure 43A). qPCR experiment performed on e11.5 embryonic faces showed that the maternal contribution of MFM enhancer to *c-Myc* expression is 30%, whereas, the paternal contribution is 50% (Figure 43B). Thus there is a correlation between landscapes and their impact on the endogenous gene expression. Intriguingly, the change in the expression of *c-Myc* is not ON/OFF unlike other imprinted loci and 17a insertion; rather, here the imprinting phenomenon modulates the contribution of the maternal and the paternal regulatory regions to *c-Myc* expression by 20%.

5.3.3.1.4 Influence of Imprinting in the architecture of the *c-Myc/Pvt1* Flanking Locus

The imprinting of the 17a insertion promoter (and neighboring insertions) correlates with the differential modulation of communication between MFM enhancer and *c-Myc*. Next, in order to find out the interaction profiles of maternal and paternal 17a insertion site we collaborated with Peter Krijger from Wouter de Laat lab for allele specific 4C experiment. I have prepared e11.5 embryonic face samples obtained from maternal and paternal 17a insertion. Peter Krijger conducted the 4C experiment with frequent cutters. 4C profile of the 17a minimal promoter viewpoint from the paternal allele is qualitatively different from the 4C profile of the 17a minimal promoter on the maternal allele (Figure 44 A-B). Unexpectedly, we could not find any obvious physical interaction between the (10a-

13a) region and the minimal promoter of 17a, although it requires further quantitation (Figure 44B).

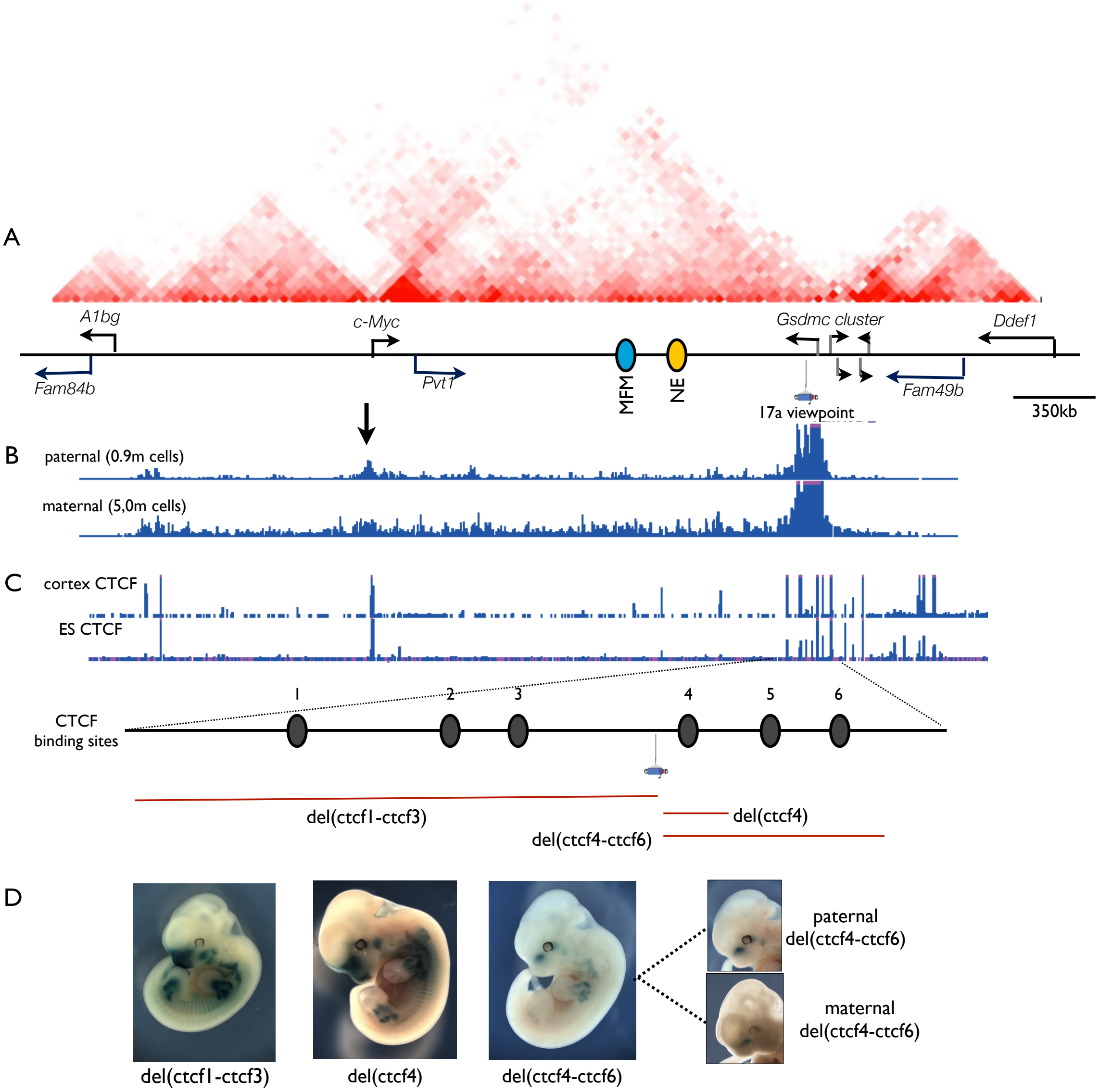


Figure44| Telomeric end of the *c-Myc/Pvt1* flanking interval: A) TAD structure of the locus (Dixon J *et al* 2012). B) 4C profile of the locus from 17a viewpoint. Read counts are shown in blue bars and they are not normalized to overall read count. The arrow points out *c-Myc* promoter and the qualitative enrichment of read counts only in the paternal allele. C) CTCF binding sites at the locus are given from two tissues. 6 CTCF sites at the telomeric end are deleted combinatorially D) lacZ staining of the deletions in the paternal alleles and the maternal allele (for del(ctcf4-ctcf6)

5.3.3.1.5 Locating Imprinting Control Center (ICR)

5.3.3.1.5.1 Systematic Deletions in the Imprinted Locus

I further examined the telomeric imprinted region of *c-Myc/Pvt-1* flanking locus. It contains clusters of CTCF sites. CTCF has been implicated in imprinting (Bell AC and Felsenfeld G, 2000). CTCF bound regions are associated with architectural changes in *ß-globin* locus and they are enriched at topological boundaries (Palstra RJ *et al* 2003; Dixon JR *et al* 2012). *c-Myc* promoter and a cluster of 6 sites at the very telomeric side of the locus shows CTCF occupancy in the whole locus (Figure 44C). I used deletions to test the involvement of these CTCF sites in the establishment of imprinting. The deletion of the (15a-17a) region that includes the first three CTCF binding sites and the (17a-17b) deletion, which removes the 4th CTCF site do not change this regulatory input in the face in the paternal allele. However, the removal of the last three CTCF binding sites by the two overlapping (17a-17d) and (17a-19a) deletions resulted in the loss of lacZ expression in MFM (Figure 44D). Therefore, the region, which includes the 5th and the 6th CTCF binding site appeared to be responsible for the allele specific communication of MFM enhancer with 17a promoter in paternal allele.

In addition, despite being missing in 17a maternal insertions, the expression of LacZ in the NE appeared upon both maternal and paternal deletions of overlapping (17a-17d) and (17a-19a) regions. This suggests in the absence of the (17a-17d) region, imprinting disappears. However, the methylation pattern of the promoter will give a more definitive answer.

5.3.3.1.5.2 Allele Specific CTCF occupancy in the Imprinted Locus

In order to see whether allele specific occupancy of CTCF site correlates with imprinting activity I did allele specific ChIP with CTCF antibody. For allele specific ChIP, I used heterozygous (17a-21a) deletion, which removes 4th, 5th, and 6th CTCF sites (Figure 45A). By setting up reciprocal het (17a-21a) deletion and wt animals, I could obtain samples where 4th, 5th and 6th CTCF sites were on the paternal allele or on the maternal allele. I performed ChIP-qPCR experiment on e11.5 embryonic face samples where these CTCF sites are present only on one allele (Figure 45B). In this experiment, CTCF occupancy of these sites was compared to the CTCF occupancy in

Ndn promoter for normalization. The results indicated that only 6th CTCF binding site is differentially occupied in the paternal allele (Figure 45C). However, this experiment has been performed only once and it requires further confirmation to show reproducible correlation with the imprinting activity and the region, which contains a CTCF occupied site.

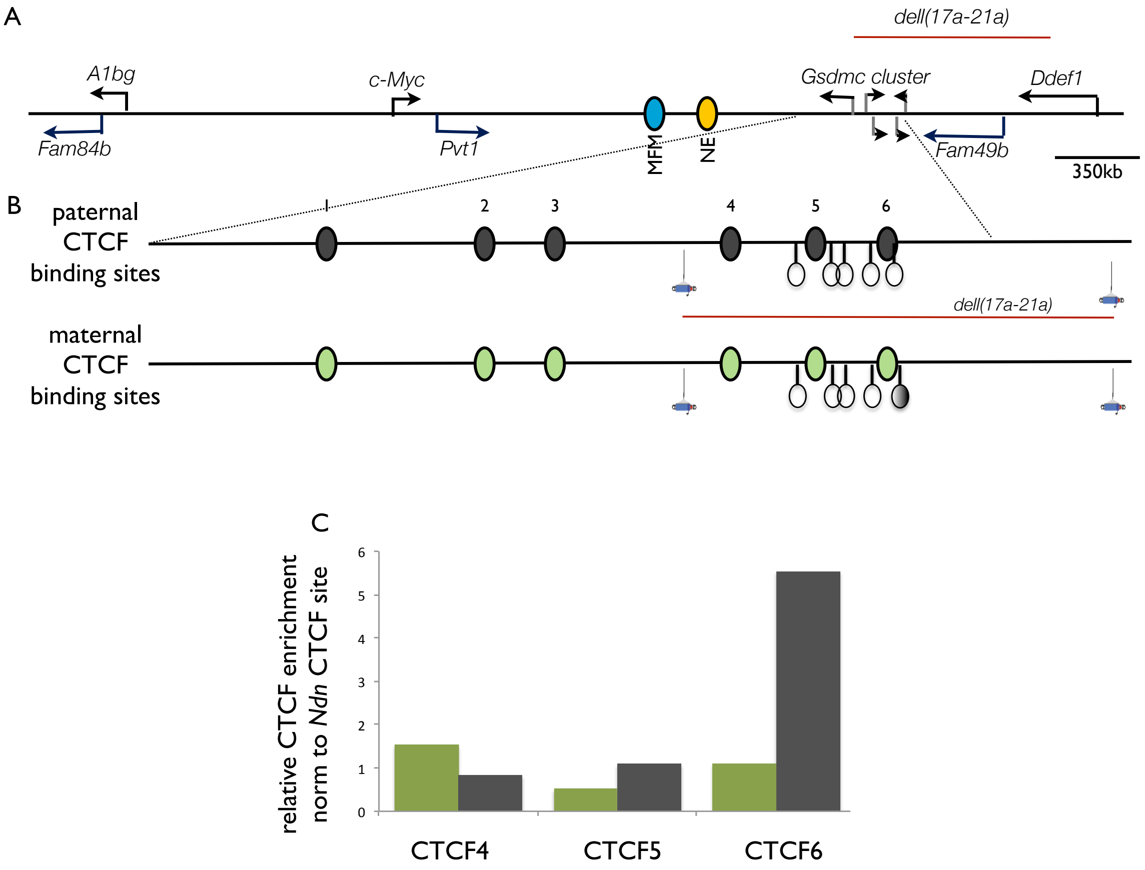


Figure45| Allele Specific Methylation and CTCF Occupancy: A) The genomic landscape is represented and the red line indicates the deletion line that is used for allele specific ChIP and bisulfite assay. B) CTCF sites are represented in circles in paternal (gray) and maternal (green) allele. The lollipops indicate the CpG sites close to the CTCF binding sites. The gray shade in the lollipops represents the prominence of methylation at these CpG sites. C) Allele specific ChIP with CTCF antibody for the 4th, 5th and 6th CTCF sites given in B. The CTCF occupancy is normalized to the CTCF occupancy in *Ndn* promoter.

5.3.3.1.5.3 Allele Specific methylation in the Imprinted Locus

In *H19/Igf2* locus, differential methylation of ICR underlies the differential occupancy of CTCF sites (Bell AC and Felsenfeld G, 2000). Therefore, in order to see if the mechanism applies to *c-Myc/Pvt-1* flanking locus, I performed allele specific bisulfite assay by using the (17a-21a) deletion like in allele specific ChIP experiment. Unlike minimal promoter, the CpG content of genomic locus that surrounds 5th and 6th CTCF sites are extremely low (2 CpGs in approximately 100bp). One of the few CpGs resides in the core-binding motif of 6th CTCF site. Bisulfite sequence analysis showed that the overall methylation of the region is also very low. Particularly the methylation in 6th CTCF site is not distinguishable between the maternal and paternal allele. However, the methylation level in the surrounding CpGs is about 10% in maternal allele and 2% in paternal allele (Figure 45B).

5.3.4 The Telomeric End of the *c-Myc/Pvt1* Flanking Locus

The findings above indicate the importance of the telomeric end of the *c-Myc/Pvt-1* locus in enhancer-promoter communication. This telomeric end contains a TAD boundary, which overlaps with the MFM regulatory domain boundary. Inversion in this locus disrupts two adjacent TAD structures and brings the boundary region between the MFM regulatory region and the *c-Myc* promoter. This rearrangement does not interfere with the activity of the enhancer(s) (as minimal promoter still shows regulatory activity in the MFM) but it blocks the communication between *c-Myc* promoter and the MFM enhancer(s). In addition, the interaction profile of the minimal promoter at the telomeric end reveals differences between maternal and paternal allele. Furthermore, tethering of the enhancer(s) to the telomeric end of the locus in the paternal allele is necessary to fully implement long-range regulation of the *c-Myc* expression.

6. DISCUSSION

6.1 Regulatory Elements in the *c-Myc/Pvt1* Flanking Locus

Functional dissection of the *c-Myc/Pvt-1* flanking locus in mouse pointed out to the presence of MFM regulatory elements in the (10a-13a) region and NE regulatory elements in the (14a-15a) region. The deletion of these elements caused morphological changes in the adult skull and in the embryonic MFM. Variations in the human 8q24 locus, which is orthologous to mouse (10a-13a) region, act as genetic risk factors for non-syndromic cleft lip and palate (NSCLP) (Birnbaum S et al 2009). Variations in this region are associated with changes in facial morphology (Liu F et al 2012). Our data shed light on the actors and the molecular nature of the risk. Both the RNAseq experiment and the follow-up qPCR verification suggested that *c-Myc* is the target gene and *c-Myc* downregulation has two major consequences in e11.5 face:

- Alteration of gene regulatory networks (GRNs) in the developing face
- Alteration of metabolic pathways via change in the ribosome biogenesis and translational control genes.

6.1.1 8q24 dependent susceptibility to Non-Syndromic Cleft Lip/Palate

6.1.1.1 *c-Myc* and Medial Nasal Mesenchyme Gene Regulatory Networks

Face development has a very strong genetic component. Monozygotic twins look almost identical or the facial structure of the kids looks like their parents or their relatives. Therefore, together with the environmental factors, genetic factors also contribute to the susceptibility to NSCLP. For example, coding sequence change in IRF6 (Zucchero TM *et al* 2004), SNPs near MAFB and ABCA4 (Beaty TH *et al* 2010) increases susceptibility to Cleft Lip and Palate in humans. In mouse*, Bmpr1A, Tgfß3, Sox11*, ß-catenin, and *Lrp6* null alleles also lead to CLP with different expressivity and penetrance (Brault *V et al* 2001, Taya Y *et al* 1999; Sock E *et al* 2004; Song L *et al* 2009). A/WySn mouse strain, which has a genetic predisposition to CLP, is a hypoactive *Wnt9b* mutant (Juriloff DM *et al* 2006). *Wnt9b* is expressed in the medial and lateral nasal processes and mutations in Wnt9b leads to cleft palate problem in communication with Fgf pathway (Jin YR et al 2012). In humans, Wnt9b is associated with NSCLP (Chiquet BT *et al* 2008). In the course of facial development, *c-Myc* interacts with many components of the gene regulatory networks such as *Tgfß3* (Zhu X *et al* 2012) and BMP receptor 1A (*Bmpr1A) (*Saito H et al 2012).

In del (8a-17a) mice, the genes that showed a mild but statistically significant change in the expression of morphogens like *Bmp7* and transcription factors like *Etv5, Sox11* and *Nr2f1,* which are involved in the facial development. Heterozygous *BMP7* mutations are associated with facial malformations in humans (Wyatt AW *et al* 2010). Moreover, conditional *Bmp7* deficient mice are shown to have cleft palate (Kouskoura T *et al* 2013). Deficiencies of *Bmp* receptors like Alk2 (Dudas M *et al* 2004) and *Bmp* signaling pathway elements like *Msx1* (Satokata I and Maas R, 1994) also lead to cleft lip and palate phenotype. *Bmp7* is expressed in the facial epithelia and the mesenchyme. Nevertheless, conditional deletion in the facial epithelia did not give rise to the cleft palate, suggesting that the *Bmp7* deficiency in the mesenchyme underlie the *Bmp7* dependent Cleft Palate formation (Kouskoura T *et al* 2013).

Etv5 is a transcription factor mediating response to *Fgf* signaling and it is strongly expressed in the palatal shelves at e13.5 (Welsh IC *et al* 2007). Variations in human ETV5 are associated with facial clefts in northern European population (Jugessur A *et al* 2009).

Sox11 deficient mice show highly penetrant CLP phenotype with a wide range of expressivity (Sock E *et al* 2004). A deletion in human 2p25, which contains SOX11 was linked with Opitz "C" trigonocephaly-like syndrome that exhibits cleft palate, together with other symptoms (Czako M *et al* 2004).

Nr2f1 is an important transcription factor for neural crest cells (NCC) and *Nr2f1* haploinsufficiency causes facial malformations in humans (Brown KK *et al* 2009). Moreover, Nr2f1 binds to a large number of NCC-specific enhancers (Rada-Iglesias A *et al* 2012) and NCCs extensively contribute to craniofacial development (reviewed in (Cordero DR *et al* 2011). Interestingly, *Nr2f1* and *TFAP2a* are known to induce permissive chromatin states when they co-occupy an enhancer sequence. Intriguingly, Jeffrey Murray's lab associates NSCLP in humans with a SNP, which alters TFAP2α binding site in IRF6 locus (Rahimov F *et al* 2008). Furthermore, in humans the genomic rearrangements next to TFAP2α are associated with orofacial clefting (Davies SJ *et al* 2004).

Altogether, deregulation of these genes in the gene regulatory networks in the embryonic facial mesenchyme may directly act on the tissue by disrupting migration, fusion, and differentiation. Besides, it may create a delay or defect in the growth of facial structures. Particularly in C57BL/6J strain *c-Myc* downregulation itself may not be enough to induce CLP but may only change the robustness of the GRN and make it susceptible to CLP.

6.1.1.2 *c-Myc* and Ribosomal Biogenesis

One of the most-reproducible downstream targets of *c-Myc* is the ribosome biogenesis genes like *Rplp0, Rplp1, Rps20* and translational control genes like *Pabc1* and *Eef2* (Figure29)*.* Except the involvement of ribosomal biogenesis genes in Diamond-Blackfan anemia, which shows craniofacial defects, developmental consequences of translation-related genes are not very well known (Doherty L *et al* 2010). On the other hand, both translational control genes and ribosomal biogenesis genes involve in metabolic pathways. *Rplp1* expression increases upon TCDD administration, which is known to induce cleft palate in mice (Yamada T *et al* 2013, Jin K-S *et al* 2012). In addition, *Eef2K*, which inhibits translation elongation factor (*Eef2*), is responsible for cell survival under starvation conditions (Lepriver G *et al* 2013). In addition, ribosomal synthesis is associated with cellular response to cytotoxicity, drug sensitivity, cellular growth, proliferation, and apoptosis (Donati G *et al* 2011, Bordeleau ME *et al* 2008, Rudra D and Warner JR, 2004). Consequently, *c-Myc* downregulation may drop the fitness of the organism or the tissue for metabolic stress, which may be caused by environmental factors.

6.1.1.3 Sporadic Cleft Lip And Palate Cases in del(8a-17a) mice.

c-Myc face enhancer deletion in (8a-17a) and (8a-14a) resulted in mild but significant changes both at the molecular, the cellular level. Strikingly, these sporadic cases were all observed from three crosses where male heterozygous del(8a-17a) were crossed with het del(8a-17a) females, which are older than 55 weeks (Table1, Figure 23). This coincidence suggests that the fetal environment may play a role and that it could be different depending on the age of the pregnant mother. The mothers' genotype (del(8-17) het) can also contribute as a factor, since del(8-17) embryos obtained by mating a del(8-17) het male to old but wild-type females did not show facial malformations. However, the sample size (eight in both cases) is too small to reach statistical significance, and further studies would be needed to precise the maternal factor that may possibly enhance or trigger the CL/P in del(8-17) animals.

6.1.1.4 Maternal Age and Cleft Lip and Palate Susceptibility in Humans

For humans the association between maternal age and CLP risk is very controversial. Hospital data from India, the USA, France, Hungary, China and some other populations indicated association between increased maternal age and occurrence of CLP, whereas, data from Italy, Canada, Australia, Seattle did not show any significant association between maternal age and CLP (Vieira AR *et al* 2002). In Asian population, the allele frequency or the strength of susceptibility upon 8q24 variation is too low to indicate any association with NSCLP (Murray T *et* al 2012) However, it is important to note that the average maternal age is higher in northern European and Northern American Population – which showed the strongest association between 8q24 risk allele and CLP - than in the South East Asian population, where the 8q24 risk allele is not associated with CLP. Thus, 8q24 risk allele may manifests its influence on the etiology of the CLP more strongly when the maternal age is high. Kerstin Ludwig from Uni-Bonn, compared the CLP risk between young and old maternal age individuals with risk allele. The individuals with CLP are sorted into two groups: One with maternal age higher than 35 and the other one lower than 25 (Table2). The ratio of the low maternal age and high maternal age among the individuals with risk allele appeared to be very similar in between these two age groups. She also estimated the CLP risk in old maternal age group between the risk allele and the common allele. For this reason, only CLP individuals with a maternal age higher than 35 are analyzed. In this analysis, the ratio of the individuals with risk allele and common allele appeared to be very similar as well. However, since the maternal age of the non-CLP individuals, which are used as controls are not recorded it is not possible to eliminate the possibility that the risk of CLP occurrence increases in patients when compared to normal individuals, in higher maternal age (Table 2).

Extreme Maternal Ages				
Frequency	AA	AC	CC	Total
age<25	20	63	46	129
age>35	4	35	19	58
Total Number of CLP cases	24	98	65	187
Genotype Age Distribution	min age	max age	mean age	Std dev
AA	18	43	27.76	4.58
AC	16	42	28.92	5.2
CC	17	45	28.76	4.9

Table2: Maternal age and the Cleft Lip Palate Occurrences: The Table at the top shows the distribution of the 8q24 alleles among the CLP patients in extreme maternal age groups. The table at the bottom shows the maternal age distribution of 8q24 alleles. None of the distributions are significantly different from each other.

6.1.1.5 Environmental Factors that increase Cleft Lip and Palate Risk

It is known that metabolic problems related to alcohol consumption, smoking, and folic acid stress during pregnancy increase NSCLP susceptibility (Dixon MJ *et al* 2011). Among many other teratogens, 6-AN is the only teratogen that is shown to increase CP incidence in BL/6 background in 1980s (Karolyi J *et al* 1988). In order to test whether del(8a-17a) mice acquire Cleft Palate (CP) in environmentally sensitized background, I used 6-aminonicotinamide (6-AN) administration. I performed intra-peritoneal 6-AN injections to the females heterozygous for the (8a-17a) deletion females at the 11th day of pregnancy and I collected the embryos three to four days after the injections. I have observed that the administration of 6-AN affects the entire body of the embryo. Upon these injections 10 percent of the embryos regardless of their genotype showed Cleft Palate phenotype together with very dramatic edema and developmental delay is observed. Therefore, the use of 6-AN is not a suitable to test the effect of del(8a-17a) on the CLP in a sensitized background.

6.1.1.6 The Link Between Facial Morphology and NSCLP

Before NSCLP was associated with any genes, facial morphology of the people who have NSCLP in the family history are associated with the shortening of nasion-anterior nasal spine (upper facial height) and increase in the width of the face (Weinberg SM *et al* 2009; Liu F *et al* 2012). This is supported by the morphometric analysis of mice strains with high incidence of CLP. For example, A/WySn strain or transgenic mice overexpression *dnBmpr1a*, which show predisposition to CLP, have shorter facial protrusion and wider face when compared to wt C57BL/6 mice (Saito H *et al* 2012). These independent cases suggest the presence of a common mechanism at the embryonic stages that gives rise to differences in facial morphology and CLP susceptibility.

In a study conducted at the beginning of 20th century, Sheldon Reed reported a correlation between the relative size of embryonic frontonasal process (with respect to the surrounding facial structures) and CLP in house mice and based on his observations he claimed imbalanced growth of frontonasal process in embryos and surrounding tissues causes CLP (Reed SC, 1933). In 1968 Daphne Trasler reported morphological differences in the MFM of A/J mice, which shows susceptibility to CLP when compared to C57BL/6 mice, which has never CLP. In addition, when the A/J mice are further sensitized for CLP by salicylic acid administration, the occurrence of the phenotype is correlates with the strength of growth imbalance between facial structures (Trasler DG, 1968). Therefore, in concordance with Sheldon Reed's claims, Daphne Trasler proposed a "developmental threshold" model for Cleft Lip and Palate. According to this model, environmental and genetic factors contribute to differential growth of adjacent facial structures, which underlies the differences of face morphology and when the growth difference between the nasal processes and adjacent maxillary process exceeds the "threshold" it leads to cleft lip and palate (Trasler DG, 1968).

6.1.1.7 Comparison of Facial Morphology of the del(8a-17a) strain with other susceptible mouse strains

More sophisticated measurements done by Micro-Ct on embryos of the A/WySn strain, which show susceptibility to CLP, pointed out that the relative size of the anterior maxillary part of the upper jaw is larger when compared to C57BL/6. Noteworthy, A/WySn mice showed much greater morphological variability when compared to C57/BL6 strain. In del(8a-17a) embryos, the major facial feature, which is significantly different from wildtype littermates was the decrease in the size of MFM and nasal processes. However, the growth difference is not over the “developmental threshold’ in del(8a-17a) animals, thus, the penetrance of CLP in del(8a-17a) is very low. Nevertheless, the measurements on 3 weeks old and 5 weeks old animals homozygous for del(8a-17a) demonstrate that facial dismorphology in del(8a-17a) is similar to what have been described in other strains with CLP risk. For example, the nasion-snout length, which contains nasal bone and protrudes the face, was smaller and interorbital distance was wider in del(8a-17a) when compared to wildtype animals. The very same morphological measurements were obtained in dnBmpr1a transgenic mice, which showed strong susceptibility to facial clefts (Saito H *et al* 2012). Furthermore, corresponding morphological features such as shorter nasion to anterior nasal spine distance (upper face distance) and wider interorbital length are associated with people, who have family history of NSCLP (Weinberg SM *et al* 2009). Noteworthy, there are also studies, which did not find any association between these morphological features and CLP susceptibility (Ward RE *et al* 1989). Nevertheless, morphological traits of del(8a-17a) strain both at the embryonic stage and the post-natal stage similar facial features with the mice strains, which shows susceptibility to NSCLP.

6.1.1.8 How can del(8a-17a) make the mice susceptible to Cleft Lip and Palate

First of all, 8q24 risk allele is present in 20% of the Northern European and American population, while the prevalence of NSCLP is 0.14%. It is therefore not a causal Mendelian mutation but a risk allele. Yet the odd ratio is 3 for heterozygous and 6 for homozygous genotype. Moreover, the contribution of other genetic factors and environmental factors such as alcohol usage, smoking, and metabolic problems during pregnancy increases the risk of NSCLP (reviewed in (Dixon MJ *et al* 2011)). Here, I elaborate on the downstream effects on *c-Myc* downregulation in the embryonic face that may underlie the 8q24 dependent susceptibility to NSCLP.

As it appeared in RNAseq experiment, del(8a-17a) downstream effects suggests two mechanisms for 8q24 dependent NSCLP susceptibility: developmental and metabolic. The developmental mechanism is due to the imbalanced growth between the MFM and the neighboring tissues in del(8a-17a) mice. In the presence of other genetic and environmental factors, this imbalance may exceed the "developmental threshold", in which Cleft Lip and Palate appears. One aspect of the developmental mechanism is the growth defect in del(8a-17a) due to the direct involvement of *c-Myc* in the cell cycle progression by activation of cell cycle genes like *Cdk4*. Besides, *c-Myc* downregulation interrupts the GRNs, which are involved in face development. For example, Nr2f1 downregulation may disrupt the neural crest cells differentiation, proliferation and migration as it is reported to cause craniofacial abnormalities (Dixon J *et al* 2006). The downregulation of other transcription factors like Sox11 and Etv5 does not only impair the proliferation and fusion of the palatal shelves but also reduces the self-renewal capacity of mesenchymal stem cells (Kubo H et al 2009). Moreover, reduction in the ribosomal biogenesis and translational control genes may locally impede cell cycle progression and tissue growth. The metabolic mechanism that can lead to susceptibility to NSCLP is the reduction in the response to metabolic stress due to downregulation of ribosomal biogenesis, translation related genes and metabolic enzymes. So that, the metabolic consequences of smoking, alcohol or folic acid metabolism problems cannot be compensated in the individuals particularly in the face. In humans, the variations in the metabolic genes like NAT1, NAT2, MTHFR and TCN2 increases susceptibility to CLP (Song T et al 2013, Bufalino A et al 2010, Martinelli M *et al* 2006).

Altogether, changes in the GRNs via *c–Myc* downregulation make the face susceptible by creating an imbalance in the facial growth and reducing the robustness of the GRN by altering the endogenous level of critical transcription factors and morphogens in these GRNs. Furthermore, in the presence of metabolic stress, which cannot be tolerated due to face specific *c-Myc* downregulation may lead to Non-Syndromic Cleft Lip and Palate (NSCLP). Dependency of the phenotype to the given genetic and environmental factors explains why NSCLP is associated with variations in 8q24 rather than being a Mendelian disorder.

6.1.2 Regulation of the Hematopoietic System by Distant Enhancers

Telomeric side deletions in the *c-Myc* locus changed the hematopoietic stem cell and lineage distribution profile. qPCR analysis demonstrated that the telomeric side of *c-Myc* locus contains hematopoietic system (HSy) enhancer(s), which act on *c-Myc*.

6.1.2.1 Effects of Distant Enhancers on undifferentiated Hematopoietic Cells

c-Myc was known to be involved in the differentiation of hematopoietic system. However, conditional *c-Myc* deletion with HSC specific *Mx-Cre* mice resulted in complete blockage of the hematopoiesis in stem cell stage. In contrast, the (14-17a) and the (8a-17a) deletions lead to a differential distribution of hematopoietic lineages.

First of all, *Mx-Cre* driven *c-Myc* conditional knockout causes a very strong increase in the LSK, which contains HSC population, whereas LS-K cells, which contains committed progenitor cells, are completely depleted (Wilson A *et al* 2004). In both del(14a-17a) and del(8a-17a), we found a significant and very strong increase in the frequency of LSK cells in the bone marrow. However, in contrast to *Mx-Cre* driven *c-Myc* deletion, neither del(8a-17a), nor del(14a-17a) showed a decrease in the frequency of LS-K cells (Lineage negative, Sca-1 negative, c-Kit positive) (Figure 32, Figure 34).

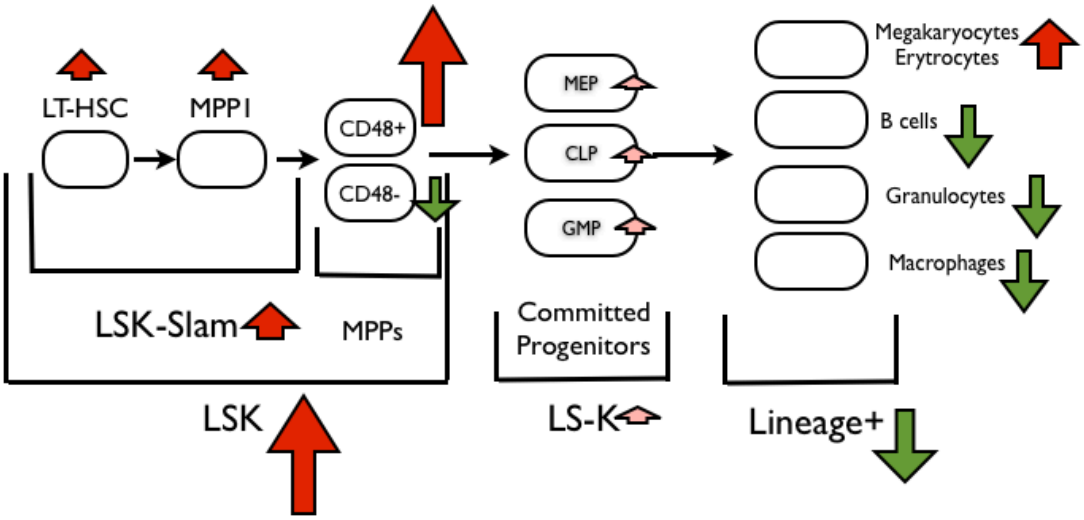


Figure46| Hematopoietic Lineage Regulation by *c-Myc* downregulation: Here is a summary of the changes in the hematopoietic lineage distribution upon telomeric deletions in the *c-Myc/Pvt1* flaking locus. The size of the arrows indicates the strength of the change. Red color indicates an increase in the frequency of the given cell type, whereas green color indicates a decrease in the frequency.

In del (8a-17a) and (14a-17a) mice, frequency of LT-HSC and MPP1 population slightly increases. Although their relative proliferation rate does not change, most of the HSC and MPP1 gets activated and enter G1 phase. Since they are activated they might lost their self-renewal capacity and they cannot re-populate hematopoietic system in a competitive HSy replenishment assay (data not shown). Alternatively, the differentiation capacity of the HSCs is independently affected from their self-renewal capacity thus they cannot re-populate the hematopoietic system. The most striking phenotype upon del(8a-17a) is the accumulation of CD48+ Multi Potent Progenitors 2/3/4 (MPP2/3/4). Although MPP2/3/4 give rise to LS-K cells, the frequency of LS-K population is very slightly increased in del(14a-17a) and not affected at all in del(8a-17a). Considering that *Mx-Cre* driven *c-Myc* deletion completely abolishes LS-K cells but (8a-17a) does not have dramatic effects on this population, it suggests that LS-K cells use other *c-Myc* enhancers outside of the deleted interval. Accordingly, although *c-Myc* expression is equally high in LSK and LS-K population, deletion of the (8a-17a) interval decreases *c-Myc* expression down to 3% of the wildtype levels in LSK cells, but only to 85% in LS-K cells. Therefore, among undifferentiated cells the strongest effect of distant hematopoietic regulatory region deletion via (8a-17a) and (14a-17a) is accumulation of CD48+ MPPs (Figure 46)

A very recent analysis of a compound deletion of the (15a-17a) interval and *c-Myc* gene recapitulated the hematopoietic phenotype of del(8a-17a) and del(14a-17a). Although further analysis and experiments are going to be done to confirm the finding, this suggest that the elements required for hematopoietic system is in the (15a-17a) interval and it is genetically linked to *c-Myc* gene through a cis-interaction.

6.1.2.2. Effects of Distant Enhancers on terminally differentiated Hematopoietic Cells

In *Mx-Cre* induced *c-Myc* deletion all of the differentiated cell populations are reduced. In contrast, the effect of del(8a-17a) and del(14a-17a) changes from one cell type to another. For example, the granulocyte lineage is the most severely affected lineage as macrophage lineage is almost completely blocked and granulocyte population shrank by more than half upon both (8a-17a) and (14a-17a) deletion. It is already shown that granulocyte-monocyte progenitors require *c-Myc* for proper differentiation and it fits to our observations in granulocyte lineage in vivo (Guo Y et al 2009). In lymphoid lineage, there is a very strong reduction in B-cell lineage in del(8a-17a) and (14a-17a), whereas, in T cell population increases in the peripheral tissues of the (14-17a) animals, which are analyzed at a later age than del(8a-17a) mice. *c-Myc* is shown to be critical for lymphoid lineage differentiation into B-cell-like cells in the cell culture, whereas, the role of *c-Myc* in T-cell differentiation and maturation is not clear in vivo. (Guo Y et al 2009, Douglas NC *et al* 2001). Interestingly, both megakaryocytes and erythrocytes population, which come from the same common progenitor (MEP) are depleted in *Mx-Cre* driven *c-Myc* conditional knock out and it causes anemia. On the contrary, neither in del(8a-17a) and nor in del(14a-17a) the number of megakaryocyte and erythrocyte decreases. This is in concordance with the studies that indicate an inhibitory role for *c-Myc* in differentiation of MEP-like cell-lines (Delgado MD and Léon J, 2010). In brief, the phenotypes that is observed in differentiated hematopoietic lineage distribution upon distant enhancer deletions in the *c-Myc/Pvt-1* flanking region recapitulates the differentiation defects observed upon *c-Myc* deficiency both in the cells lines and *in vivo.*

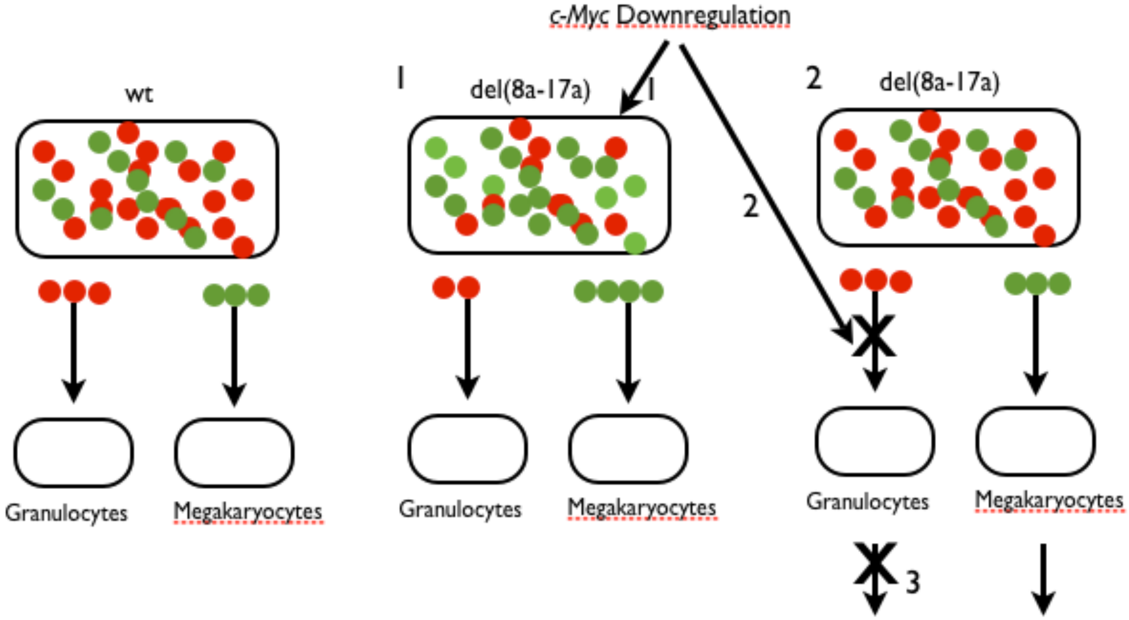


Figure47| Mode of Action of *c-Myc* in the Hematopoietic Lineage: The big pool of cells represent the committed progenitor population that give rise to the terminally differentiated hematopoietic cells. Cells with strong *c-Myc* expression are represented in red circles and cells with weak *c-Myc* expression is represented in green circles. The numbers correspond to the possible intervention points of enhancer deletions to the hematopoietic lineage.

There may be three alternative mechanisms, which may lead to the drastic changes in differentiated hematopoietic lineage distribution (Figure 47):

1) Differentiation is determined by heterogeneity of the *c-Myc* expression in the progenitor population.
2) Differentiation is based on the use of different enhancers
3) Differentiation is based on different requirements and complex regulation.

First mechanism suggests that the level of *c-Myc* can be the determinant of the direction, which committed progenitor cells differentiate through. For example, committed progenitor cells may be a pool of cells, which exhibits heterogeneous *c-Myc* expression. Among this heterogeneous population low *c-Myc* expressing cells are more likely to go through megakaryocyte/erythrocyte differentiation pathway, whereas, high *c-Myc* expressing cells go through granulocyte differentiation pathway. *c-Myc* downregulation changes the heterogeneity balance in the committed cells and the differentiation favors megakaryocyte/erythrocyte pathway. However, the size of the LS-K population and the *c-Myc* level in the LS-K progenitor population did not change dramatically in del(8a-17a), arguing against this mechanism.

The second mechanism suggests that there are cell type specific *c-Myc* regulatory elements that regulate *c-Myc* expression for each differentiation pathways. Therefore, the only lineages that have regulatory elements in the (8a-17a) interval are blocked or reduced in del(8a-17a).

The third mechanism suggests that *c-Myc* expression is regulated by the same enhancers but the response of each cell type to the *c-Myc* level change is different. For example it is shown that *c-Myc* overexpression can induce apoptosis but also proliferation depending on the cell line via activating different cellular pathways.

6.1.3 Copy Number Variation in the *c-Myc/Pvt1* Flanking Locus

Pediatric Acute Myeloid Leukemia is associated with duplications in the telomeric end of the *c-Myc/Pvt*-1 flanking locus. The hematopoietic phenotype obtained upon del(14a-17a) and del(8a-17a) has opposite features with Acute Myeloid Leukemia (AML). For example, AML patients, granulocytes, B and T lymphocytes are deregulated and the number of white blood cells, which do not carry the exact markers with normal B and T cell lymphocytes, increase in the bone marrow. These patients also suffer from low erythrocytes and megakaryocytes (reviewed in (Gorczyca W *et al* 2011)). In addition, it is reported that *c-Myc* overexpression causes hematopoietic deregulation and blood cancers in mouse (Adams JM *et al* 1985; Sidman CL *et al* 1993). The duplication of the (15a-17a) interval, which is orthologous to AML duplications in humans, and the duplication of the whole 3mb-long *c-Myc*/*Pvt-1* flanking locus had no effect on hematopoietic lineage distribution except slight increase in the reduction of CLP population in MycDup3MB. However, AML is clonal, such that a few aberrant cells can give rise to it. FACS analysis cannot detect these cells (Bochtler T *et al* 2013). Therefore, despite not seeing an alteration in the hematopoietic lineage distribution, we cannot rule out that these two duplication lines have predisposition to AML.

In the cell population scale, *c-Myc* duplication does not influence the hematopoietic lineage distribution. Dosage compensation for the extra copies of *c-Myc* could act at different stages, transcriptional level, post-transcriptional level or translational level. qPCR analysis indicated the *c-Myc* copy number increase can be compensated at transcriptional level in the LT-HSCs and MPP1 populations but the level of *c-Myc* increases linearly with the *c-Myc* copy number in the MPP2/3/4 populations. This finding is very important for the interpretation of molecular mechanisms of genetic disorders, in which genetic disorders may exhibit the phenotypes in a tissue specific manner not due to deregulation via tissue specific enhancers but differential dosage compensation of tissues.

6.2 Regulatory Organization of the *c-Myc/Pvt1* Flanking Locus

The regulatory reporter insertions in the *c-Myc/Pvt-1* flanking locus captures *c-Myc* expression domain almost in entire 2mb region, which suggests that the enhancer-promoter interactions at *c-Myc* locus are dynamic. Moreover, enhancers are operational not only for the regions between themselves and their target gene but also act on the regions outside. This observation stands against the classical way of model of the long-range interactions, where there are specific loops between the enhancers and the promoters (reviewed in Krivega I and Dean A, 2012). Interestingly, the presence of minimal promoter, which captures *c-Myc* regulatory activity, does not compete with *c-Myc* promoter. Namely, by introducing new promoters that enhancers act on, the communication between the enhancer and the endogenous target promoter is not disrupted. Considering that in the native state there is no minimal promoter, these interactions are most likely a non-functional noise as an outcome of dynamic chromatin structure.

The dynamic nature of the enhancer is important to understand where the enhancer can act on. For example, the interaction profile of the enhancer can selectively regulate one or more genes and excludes the others as the non-regulated genes are out of the enhancer's range. Therefore, it is crucial to understand the molecular players that regulate the range of enhancers. In the literature, there are many examples, in which the enhancer-target promoter communication is broken by changes in the DNA structure of these sequences. For example, breast cancer associated SNPs change the affinity of the *Tox3* enhancer sequence to FoxA1 protein and causes changes in the expression of the target gene (Cowper-Sal lari R *et al* 2012). Similarly, SNPs at the promoter sequence of *Pdfr-α* gene changes the expression profile of the gene (De Bustos C *et al* 2005). However, the activity of the enhancers on target genes is not only modulated by changes in the enhancer or promoter sequence but even in the presence of intact enhancer and promoter sequence, changes in the genomic structure causes diseases. For instance a balanced inversion in *HoxD* locus give rise to Ulnaless phenotype (Spitz F *et al* 2003, Herault Y *et al* 1998) or insertion of *musD* transposable element to mouse *Fgf8* locus changes the target promoter of limb enhancers of the locus and leads to Split Hand Foot Malformation (Kano H *et al* 2007). There are cases where a blood disorder associated SNP creates a new promoter-like element that interferes with the activation of blood specific globin genes (De Gobbi M *et al* 2006). Therefore, the genome organization, which is a concept that addresses the regulation of the distribution of enhancer activities, is critical to unravel to the molecular mechanisms of spatiotemporal regulation of the gene expression.

Here I will discuss the elements in *c-Myc* locus that regulates the enhancer-promoter communication.

6.2.1 Influence of Distance and Genomic Context in Genome Organization

In the results part I have shown that the MFM enhancer cannot act on 10a insertion site at the endogenous locus (Figure 39). Upon (10a-21a) duplication, the minimal promoter is still silent. However, in dup(10a-20a), MFM enhancer starts acting on the regulatory sensor and in del(10a-20a). The distance between the MFM enhancer and the minimal promoter is the same in all these three cases. The only parameter that changes is the centromeric genomic context of the insertion site and this suggests that the centromeric side can work as a silencer. However, in del(10a-20a), where the centromeric side of 10a insertion is present, *Fam49b* locus enhancers act on minimal promoter, which argues against a silencer role for the centromeric side of 10a. Considering that minimal promoter at 10a position is only activated upon rearrangements with 20a insertion, the surrounding of 20a insertion is critical to understand the genomic context. Intriguingly, 20a insertion lies in the middle of a CpG rich promoter site. Therefore, promoter sites or CpG islands can provide accessibility to a site and allows the enhancers act on these sites. Besides, LINEs and SINEs can be a major determinant of genomic context. Local differences in chromatin accessibility or ability to support transcription may contribute to the discontinuity or fluctuations in the regulatory input at very short distance. Finally, for the quantitative changes occurring at very short distances, it could be also due to the transposon insertion disrupting a specific element, which is involved in the local influence like simple repeats or SINEs. Therefore, not only the dynamic structure of the enhancer but also the local environment of the target site regulates the range of enhancer, which it can operate.

6.2.2 Why do the *Fam49b* locus enhancers act short Range?

20a and its neighboring insertions demonstrate a characteristic widespread expression pattern, which is very intense in the forebrain, midbrain-hindbrain boundary. This expression is very consistent in a few kb region around 20a. On the telomeric side 100kb far 21a insertion captures midbrain hindbrain expression pattern and 300kb far insertion recapitulates the widespread expression pattern. However, on the centromeric side, there is one insertion 100kb far, which shows tail-bud ectoderm expression and then there is the transition region between *c-Myc* locus and *Fam49b* locus. In brief, enhancers acting on 20a can reach to telomeric regions but not on centromeric regions. Our summer student, Alicia Lardennois investigated whether the presence of a transition region restricts the range of *Fam49b* region enhancers on the centromeric side as a boundary element. She remobilized the transposon from (10a-20a), which lacks the boundary region and looked at the expression patterns in the embryos. Two insertions, which are 30kb and 500kb far on the centromeric side of the (10a-20a) deletion, did not show the

20a-like expression pattern. This suggested that the even in the absence of the boundary element, the range of enhancer did not extend. Although it is possible that the insertions sites were not accessible to the *Fam49b* region enhancers, the results suggest that the telomeric side of 20a region shapes the regulatory domain and determines the boundary region on the centromeric side.

6.2.3 Influence of Regulatory Domains on the Endogenous Gene Activity

GROMIT reveals a regulatory domain, where the enhancer acts on non-promoter regions. I investigated whether the shape of the regulatory domain has an influence on enhancer-promoter communication. In other words, I addressed whether the promiscuous interaction of the regulatory region with non-promoter sites along the locus influences its interaction with the promoter site. In order to understand the role of distribution or regulatory activity on the expression of endogenous genes, an experimental system, in which the regulatory activity is different but the rest of the parameters like cellular background, protein content, transcript levels etc. are identical. The only system, in which two alleles in the same cell shows differential regulatory activity is the imprinted regions. In the imprinted loci like Igf2/H19, a set of genes is expressed only in the maternal or in the paternal allele and imprinting is primarily associated with differential DNA methylation in between two alleles. In a classical imprinted locus like this, the promoters of the genes are also imprinted. This makes it complicated to address whether in these loci the enhancer-promoter interaction is perturbed or the interaction is not perturbed but promoter cannot use the regulatory information due to its local environment. However, in *c-Myc/Pvt-1* flanking locus, imprinting exclusively takes place in the telomeric end in the maternal allele. Taking into account that *c-Myc* hasn't been shown as an imprinted gene, the allele specific contribution of the enhancer to *c-Myc* expression reflects the influence of the shape of regulatory landscape on endogenous gene expression. A similar phenomenon has been reported upon a particular rearrangement in *Hoxd* locus, where allele specific distribution of regulatory activity results in differential expression of inserted lacZ reporter gene (Lonfat N *et al* 2013).

The phenomena that we have observed may be an overlooked aspect of imprinting as the transcriptional change between maternal and paternal allele is limited to 20%, whereas, in the classical imprinted loci the difference is as clear as ON and OFF state of the gene.

6.2.4 Distribution of the Regulatory Activity and the Mechanisms of Enhancer-Promoter Interaction

According to the classical model, the distant regulatory sequences interact with the target promoter by making specific loops. However, what we saw in the *c-Myc* locus is that most positions in 2 mb long telomeric side of *c-Myc* locus appeared to be operationally permissive under the influence of long-range regulatory inputs, apart a few. Moreover, introduction of a minimal promoter to this locus do not lag *c-Myc* expression and deletion of *c-Myc* promoter does not have a clear positive effect on minimal promoter activity. Non-linearity and discontinuity of regulatory input along the locus argues against the "scanning" model (Wasylyk B *et al* 1983), in which enhancers move along the chromatin fiber (linearly, in one dimension) to communicate with the target promoter. The regulatory landscapes in *c-Myc* locus favor a model, where the enhancers operate in a prefolded confined space in 3D, such as in regulatory archipelagoes (Montavon T *et al* 2013). The boundaries of these confined spaces are defined by tissue invariant TADs. The structure within a TAD is determined in a cell-type specific manner (Philips-Cremins JE *et al* 2013). Eventually, the enhancer-promoter communication takes place via formation of dynamic loops in tissue-specifically determined 3D space with tissue invariant boundaries.

6.3 Final Conclusions and Outlook

I used *c-Myc/Pvt-1* flanking locus as a model to investigate the biological role of the distant non-coding regions. In this locus, I have identified three regulatory regions, two of which are active in the developing face and one of which is active in different stages of hematopoietic system. I showed that these regulatory sequences act on the *c-Myc* gene despite being unprecedentedly distant to its promoter.

The regulatory region active in the facial mesenchyme is orthologous to the LD block, which is strong risk allele for NSCLP in humans. In order to understand how this region contributes to the risk, I investigated the downstream effects of the rearrangements taking place in this regulatory region. I have found that MFM specific *c-Myc* downregulation alter a robust GRN and create a spatially restricted impaired translational region, which desensitize the tissue against metabolic stress.

In addition, the hematopoietic lineage specific regulatory regions that I have described, in collaboration with Lisa Von Paleske, are also associated with blood cancers. Therefore, identification of regulatory region shed light on the complex role of *c-Myc* during hematopoiesis.

My observations in the *c-Myc* locus showed the regulatory activity in the face is distributed in a non-continuous and non-linear manner all along the DNA but it is restricted in a TAD structure. Furthermore, the inversion in *c-Myc* locus, which moved a TAD boundary between the face enhancer and *c-Myc* gene and disrupted two adjacent TADs, blocked the communication of the enhancer with the target promoter. In coherence with the current literature, the regulatory landscapes suggest that TADs are as the basic blocks that confines regulatory domains. In addition to the structure of the locus, I have shown that not the distance between the enhancer and the promoter but the local environment of the promoter affects enhancer-promoter interactions.

I found that a peculiar local imprinting in the telomeric end of the *c-Myc* locus, where the shape of the regulatory landscape is different between maternal and paternal allele. Allele specific tethering of the face enhancer to the telomeric end of the locus correlated with a 20% increase of in the *c-Myc* expression in this tissue. Hence, I have shown a correlation between the local changes in the regulatory landscape correlated and mild changes in the enhancer-promoter communication.

In conclusion, my results did not only identify critical functional regulatory element and their target in *c-Myc* locus, but also shed light on the influence of the genome organization on the communication between these regulatory sequences and their target promoter. From the evolutionary point of view, the regulation of the range of enhancers allows evolutionary tinkering by providing a flexible manner to modulate spatiotemporal gene regulation.

7. REFERENCES

Adams, J. M., Harris, A. W., Pinkert, C. A., Corcoran, L. M., Alexander, W. S., Cory, S., … Brinster, R. L. (1985). The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature*, *318*(6046), 533–8.

Adey, A., Burton, J. N., Kitzman, J. O., Hiatt, J. B., Lewis, A. P., Martin, B. K., … Shendure, J. (2013). The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*, *500*(7461), 207–11.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L. A., & Rubin, E. M. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS biology*, *5*(9), e234.

Akashi, K., Traver, D., Miyamoto, T., & Weissman, I. L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, *404*(6774), 193–7.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. D. (1994). From Single Cells to Multicellular Organisms. In *Molecular Biology of the Cell* (3rd editio.). New York: Garland Science.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental cell*, *16*(1), 47–57.

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science (New York, N.Y.)*, *339*(6123), 1074–7.

Arnosti, D. N., & Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *Journal of cellular biochemistry*, *94*(5), 890–8.

Ashburner, M. (1970). Patterns of puffing activity in the salivary gland chromosomes of Drosophila. *Chromosoma*, *31*(3), 356–376.

Atchison, M. L. (1988). Enhancers: mechanisms of action and cell specificity. *Annual review of cell biology*, *4*, 127–53.

Attanasio, C., Nord, A. S., Zhu, Y., Blow, M. J., Li, Z., Liberton, D. K., … Visel, A. (2013). Fine tuning of craniofacial morphology by distant-acting enhancers. *Science (New York, N.Y.)*, *342*(6157), 1241006

Banerji, Julian, Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308.

Banerji, J, Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–40.

Bantignies, F., Roure, V., Comet, I., Leblanc, B., Schuettengruber, B., Bonnet, J., … Cavalli, G. (2011). Polycomb-dependent regulatory contacts between distant Hox loci in Drosophila. *Cell*, *144*(2), 214–26.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823–37.

Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., … Marti-Renom, M. A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, *18*(1), 107–14.

Baus, J., Liu, L., Heggestad, A. D., Sanz, S., & Fletcher, B. S. (2005). Hyperactive transposase mutants of the Sleeping Beauty transposon. *Molecular therapy : the journal of the American Society of Gene Therapy*, *12*(6), 1148–56.

Beaty, T. H., Murray, J. C., Marazita, M. L., Munger, R. G., Ruczinski, I., Hetmanski, J. B., … Fallin, M. D. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nature genetics*, *42*(6), 525–9.

Bell, A. C., & Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, *405*(6785), 482–5.

Belloni, E., Muenke, M., Roessler, E., Traverso, G., Siegel-Bartelt, J., Frumkin, A., … Scherer, S. W. (1996). Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nature genetics*, *14*(3), 353–6.

Bellosta, P., & Gallant, P. (2010). Myc Function in Drosophila. *Genes & cancer*, *1*(6), 542–546.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., … Lander, E. S. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, *125*(2), 315–26.

Bertrand, K., Squires, C., & Yanofsky, C. (1976). Transcription termination in vivo in the leader region of the tryptophan operon of Escherichia coli. *Journal of Molecular Biology*, *103*(2), 319–337.

Bettess, M. D., Dubois, N., Murphy, M. J., Dubey, C., Roger, C., Robine, S., & Trumpp, A. (2005). c-Myc is required for the formation of intestinal crypts but dispensable for homeostasis of the adult intestinal epithelium. *Molecular and cellular biology*, *25*(17), 7868–78.

Beverdam, A., Brouwer, A., Reijnen, M., Korving, J., & Meijlink, F. (2001). Severe nasal clefting and abnormal embryonic apoptosis in Alx3/Alx4 double mutant mice. *Development (Cambridge, England)*, *128*(20), 3975–86.

Birnbaum, S., Ludwig, K. U., Reutter, H., Herms, S., Steffens, M., Rubini, M., … Mangold, E. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nature genetics*, *41*(4), 473–7.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nature genetics*, *42*(9), 806–10.

Bochtler, T., Stölzel, F., Heilig, C. E., Kunz, C., Mohr, B., Jauch, A., … Krämer, A. (2013). Clonal heterogeneity as detected by metaphase karyotyping is an indicator of poor prognosis in acute myeloid leukemia. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, *31*(31), 3898–905.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., … Cremer, T. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, *3*(5), e157.

Bonal, C., Thorel, F., Ait-Lounis, A., Reith, W., Trumpp, A., & Herrera, P. L. (2009). Pancreatic inactivation of c-Myc decreases acinar mass and transdifferentiates acinar cells into adipocytes in mice. *Gastroenterology*, *136*(1), 309–319.e9.

Bonn, S., Zinzen, R. P., Girardot, C., Gustafson, E. H., Perez-Gonzalez, A., Delhomme, N., … Furlong, E. E. M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature genetics*, *44*(2), 148–56.

Bordeleau, M.-E., Robert, F., Gerard, B., Lindqvist, L., Chen, S. M. H., Wendel, H.-G., … Pelletier, J. (2008). Therapeutic suppression of translation initiation modulates chemosensitivity in a mouse lymphoma model. *The Journal of clinical investigation*, *118*(7), 2651–60.

Borok, M. J., Tran, D. A., Ho, M. C. W., & Drewell, R. A. (2010). Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. *Development (Cambridge, England)*, *137*(1), 5–13.

Bouhassira, E. E., Westerman, K., & Leboulch, P. (1997). Transcriptional Behavior of LCR Enhancer Elements Integrated at the Same Chromosomal Locus by Recombinase-Mediated Cassette Exchange. *Blood*, *90*(9), 3332–3344.

Branco, M. R., Ficz, G., & Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature reviews. Genetics*, *13*(1), 7–13.

Branco, M. R., & Pombo, A. (2006). Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, *4*(5), e138.

Brand, A. H., & Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development (Cambridge, England)*, *118*(2), 401–15.

Brault, V., Moore, R., Kutsch, S., Ishibashi, M., Rowitch, D. H., McMahon, A. P., … Kemler, R. (2001). Inactivation of the beta-catenin gene by Wnt1-Cre-mediated deletion results in dramatic brain malformation and failure of craniofacial development. *Development (Cambridge, England)*, *128*(8), 1253–64.

Brien, G. L., Gambero, G., O'Connell, D. J., Jerman, E., Turner, S. A., Egan, C. M., … Bracken, A. P. (2012). Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. *Nature structural & molecular biology*, *19*(12), 1273–81.

Brown, K. K., Alkuraya, F. S., Matos, M., Robertson, R. L., Kimonis, V. E., & Morton, C. C. (2009). NR2F1 deletion in a patient with a de novo paracentric inversion, inv(5)(q15q33.2), and syndromic deafness. *American journal of medical genetics. Part A*, *149A*(5), 931–8.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology*, *212*(4), 563–78.

Bufalino, A., Ribeiro Paranaíba, L. M., Nascimento de Aquino, S., Martelli-Júnior, H., Oliveira Swerts, M. S., & Coletta, R. D. (2010). Maternal polymorphisms in folic acid metabolic genes are associated with nonsyndromic cleft lip and/or palate in the Brazilian population. *Birth defects research. Part A, Clinical and molecular teratology*, *88*(11), 980–6.

Bungert, J., Tanimoto, K., Patel, S., Liu, Q., Fear, M., & Engel, J. D. (1999). Hypersensitive Site 2 Specifies a Unique Function within the Human beta -

Globin Locus Control Region To Stimulate Globin Gene Transcription. *Mol. Cell. Biol.*, *19*(4), 3062–3072.

Bushey, A. M., Dorman, E. R., & Corces, V. G. (2008). Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Molecular cell*, *32*(1), 1–9.

Cai, H. N., & Levine, M. (1997). The gypsy insulator can function as a promoter-specific silencer in the Drosophila embryo. *The EMBO journal*, *16*(7), 1732–41.

Cai, H. N., & Shen, P. (2001). Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. *Science (New York, N.Y.)*, *291*(5503), 493–5.

Calhoun, V. C., Stathopoulos, A., & Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(14), 9243–7.

Chambeyron, S., & Bickmore, W. A. (2004). Does looping and clustering in the nucleus regulate gene expression? *Current opinion in cell biology*, *16*(3), 256–62.

Chen, C., Morris, Q., & Mitchell, J. A. (2012). Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features. *BMC genomics*, *13*(1), 152.

Chen, C.-K., Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Smedley, D., & Spitz, F. (2013). TRACER: a resource to study the regulatory architecture of the mouse genome. *BMC genomics*, *14*, 215.

Chen, L., & Widom, J. (2005). Mechanism of transcriptional silencing in yeast. *Cell*, *120*(1), 37–48.

Chi, J.-T., Wang, Z., Nuyten, D. S. A., Rodriguez, E. H., Schaner, M. E., Salim, A., … Brown, P. O. (2006). Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. (E. T. Liu, Ed.)*PLoS medicine*, *3*(3), e47.

Chiquet, B. T., Blanton, S. H., Burt, A., Ma, D., Stal, S., Mulliken, J. B., & Hecht, J. T. (2008). Variation in WNT genes is associated with non-syndromic cleft lip with or without cleft palate. *Human molecular genetics*, *17*(14), 2212–8.

Cho, H., Kim, T.-K., Mancebo, H., Lane, W. S., Flores, O., & Reinberg, D. (1999). A protein phosphatase functions to recycle RNA polymerase II. *Genes & Development*, *13*(12), 1540–1552.

Chung, J. H., Whiteley, M., & Felsenfeld, G. (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, *74*(3), 505–14.

Cordero, D. R., Brugmann, S., Chu, Y., Bajpai, R., Jame, M., & Helms, J. A. (2011). Cranial neural crest cells on the move: their roles in craniofacial development. *American journal of medical genetics. Part A*, *155A*(2), 270–9.

Cowper-Sal lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., … Lupien, M. (2012). Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature genetics*, *44*(11), 1191–

Crossley, M., & Brownlee, G. G. (1990). Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. *Nature*, *345*(6274), 444–6.

Czakó, M., Riegel, M., Morava, E., Bajnóczky, K., & Kosztolányi, G. (2004). Opitz "C" trigonocephaly-like syndrome in a patient with terminal deletion of 2p and partial duplication of 17q. *American journal of medical genetics. Part A*, *131*(3), 310–2.

Dalla-Favera, R., Gelmann, E. P., Martinotti, S., Franchini, G., Papas, T. S., Gallo, R. C., & Wong-Staal, F. (1982). Cloning and characterization of different human sequences related to the onc gene (v-myc) of avian myelocytomatosis virus (MC29). *Proceedings of the National Academy of Sciences of the United States of America*, *79*(21), 6497–501.

Dang, C. V. (1999). c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Molecular and cellular biology*, *19*(1), 1–11.

Davies, S. J., Wise, C., Venkatesh, B., Mirza, G., Jefferson, A., Volpi, E. V, & Ragoussis, J. (2004). Mapping of three translocation breakpoints associated with orofacial clefting within 6p24 and identification of new transcripts within the region. *Cytogenetic and genome research*, *105*(1), 47–53.

Davis, A. C., Wims, M., Spotts, G. D., Hann, S. R., & Bradley, A. (1993). A null c-myc mutation causes lethality before 10.5 days of gestation in homozygotes and reduced fertility in heterozygous female mice. *Genes & development*, *7*(4), 671–82.

De Bustos, C., Smits, A., Strömberg, B., Collins, V. P., Nistér, M., & Afink, G. (2005). A PDGFRA promoter polymorphism, which disrupts the binding of ZNF148, is associated with primitive neuroectodermal tumours and ependymomas. *Journal of medical genetics*, *42*(1), 31–7

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., … Higgs, D. R. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science (New York, N.Y.)*, *312*(5777), 1215–7.

De Wit, E., & de Laat, W. (2012). A decade of 3C technologies: insights into nuclear organization. *Genes & development*, *26*(1), 11–24.

Dean, A. (2006). On a chromosome far, far away: LCRs and gene expression. *Trends in genetics : TIG*, *22*(1), 38–45.

Delgado, M. D., & León, J. (2010). Myc roles in hematopoiesis and leukemia. *Genes & cancer*, *1*(6), 605–16.

Dixon, J., Jones, N. C., Sandell, L. L., Jayasinghe, S. M., Crane, J., Rey, J.-P., … Trainor, P. A. (2006). Tcof1/Treacle is required for neural crest cell formation and proliferation deficiencies that cause craniofacial abnormalities. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(36), 13403–8.

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–80.

Dixon, M. J., Marazita, M. L., Beaty, T. H., & Murray, J. C. (2011). Cleft lip and palate: understanding genetic and environmental influences. *Nature reviews. Genetics*, *12*(3), 167–78.

Doherty, L., Sheen, M. R., Vlachos, A., Choesmel, V., O'Donohue, M.-F., Clinton, C., … Gazda, H. T. (2010). Ribosomal protein genes RPS10 and RPS26 are commonly mutated in Diamond-Blackfan anemia. *American journal of human genetics*, *86*(2), 222–8.

Donati, G., Bertoni, S., Brighenti, E., Vici, M., Treré, D., Volarevic, S., … Derenzini, M. (2011). The balance between rRNA and ribosomal protein synthesis up- and downregulates the tumour suppressor p53 in mammalian cells. *Oncogene*, *30*(29), 3274–88.

Donze, D., Adams, C. R., Rine, J., & Kamakaka, R. T. (1999). The boundaries of the silenced HMR domain in Saccharomyces cerevisiae. *Genes & development*, *13*(6), 698–708.

Dorn, A., Bollekens, J., Staub, A., Benoist, C., & Mathis, D. (1987). A multiplicity of CCAAT box-binding proteins. *Cell*, *50*(6), 863–72.

Dose, M., Khan, I., Guo, Z., Kovalovsky, D., Krueger, A., von Boehmer, H., … Gounari, F. (2006). c-Myc mediates pre-TCR-induced proliferation but not developmental progression. *Blood*, *108*(8), 2669–77.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., … Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, *16*(10), 1299–309.

Douglas, N. C., Jacobs, H., Bothwell, A. L., & Hayday, A. C. (2001). Defining the specific physiological requirements for c-Myc in T cell development. *Nature immunology*, *2*(4), 307–15.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., … Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, *465*(7296), 363–7.

Dubois, N. C., Adolphe, C., Ehninger, A., Wang, R. A., Robertson, E. J., & Trumpp, A. (2008). Placental rescue reveals a sole requirement for c-Myc in embryonic erythroblast survival and hematopoietic stem cell function. *Development (Cambridge, England)*, *135*(14), 2455–65.

Dudas, M., Sridurongrit, S., Nagy, A., Okazaki, K., & Kaartinen, V. (2004). Craniofacial defects in mice lacking BMP type I receptor Alk2 in neural crest cells. *Mechanisms of development*, *121*(2), 173–82

Dunn, T. M., Hahn, S., Ogden, S., & Schleif, R. F. (1984). An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(16), 5017–20.

Dzierzak, E., & Medvinsky, A. (1995). Mouse embryonic hematopoiesis. *Trends in genetics : TIG*, *11*(9), 359–66.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, *10*, 48.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., & Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, *411*(6836), 494–8.

Endoh, M., Endo, T. A., Endoh, T., Isono, K., Sharif, J., Ohara, O., … Koseki, H. (2012). Histone H2A mono-ubiquitination is a crucial step to mediate PRC1-dependent repression of developmental genes to maintain ES cell identity. *PLoS genetics*, *8*(7), e1002774.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, *9*(3), 215–6.

Eskeland, R., Leeb, M., Grimes, G. R., Kress, C., Boyle, S., Sproul, D., … Bickmore, W. A. (2010). Ring1B compacts chromatin structure and represses gene expression independent of histone ubiquitination. *Molecular cell*, *38*(3), 452–64.

C. **elegans** Sequencing Consortium. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science (New York, N.Y.)*, *282*(5396), 2012–8.

Falvo, J. V, Parekh, B. S., Lin, C. H., Fraenkel, E., & Maniatis, T. (2000). Assembly of a functional beta interferon enhanceosome is dependent on ATF-2-c-jun heterodimer orientation. *Molecular and cellular biology*, *20*(13), 4814–25.

Fantes, J., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., … Mannens, M. (1995). Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Human molecular genetics*, *4*(3), 415–22.

Fernandez, P. C., Frank, S. R., Wang, L., Schroeder, M., Liu, S., Greene, J., … Amati, B. (2003). Genomic targets of the human c-Myc protein. *Genes & development*, *17*(9), 1115–29.

Fiering, S., Whitelaw, E., & Martin, D. I. (2000). To be or not to be active: the stochastic nature of enhancer action. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *22*(4), 381–7.

Fire, A., Albertson, D., Harrison, S. W., & Moerman, D. G. (1991). Production of antisense RNA leads to effective and specific inhibition of gene expression in C. elegans muscle. *Development (Cambridge, England)*, *113*(2), 503–14.

Friedli, M., Barde, I., Arcangeli, M., Verp, S., Quazzola, A., Zakany, J., … Antonarakis, S. E. (2010). A systematic enhancer screen using lentivector transgenesis identifies conserved and non-conserved functional elements at the Olig1 and Olig2 locus. *PloS one*, *5*(12), e15741.

Fujihara, M., Ito, N., Pace, J. L., Watanabe, Y., Russell, S. W., & Suzuki, T. (1994). Role of endogenous interferon-beta in lipopolysaccharide-triggered activation of the inducible nitric-oxide synthase gene in a mouse macrophage cell line, J774. *J. Biol. Chem.*, *269*(17), 12773–12778.

Gause, M., Webber, H. A., Misulovin, Z., Haller, G., Rollins, R. A., Eissenberg, J. C., … Dorsett, D. (2008). Functional links between Drosophila Nipped-B and cohesin in somatic and meiotic cells. *Chromosoma*, *117*(1), 51–66.

Gerasimova, T. I., Byrd, K., & Corces, V. G. (2000). A chromatin insulator determines the nuclear localization of DNA. *Molecular cell*, *6*(5), 1025–35.

Gershenzon, N. I., & Ioshikhes, I. P. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics (Oxford, England)*, *21*(8), 1295–300.

Geyer, P. K., Spana, C., & Corces, V. G. (1986). On the molecular mechanism of gypsy-induced mutations at the yellow locus of Drosophila melanogaster. *The EMBO journal*, *5*(10), 2657–62.

Geyer, P. K., & Clark, I. (2002). Protecting against promiscuity: the regulatory role of insulators. *Cellular and molecular life sciences : CMLS*, *59*(12), 2112–27.

Gheldof, N., Smith, E. M., Tabuchi, T. M., Koch, C. M., Dunham, I., Stamatoyannopoulos, J. A., & Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic acids research*, *38*(13), 4325–36.

Ghosh, D., Gerasimova, T. I., & Corces, V. G. (2001). Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. *The EMBO journal*, *20*(10), 2518–27.

Gidoni, D., Dynan, W. S., & Tjian, R. (1984). Multiple specific contacts between a mammalian transcription factor and its cognate promoters. *Nature*, *312*(5993), 409–13.

Gierer, A. (1966). Model for DNA and protein interactions and the function of the operator. *Nature*, *212*(5069), 1480–1.

Gilmour, D. S., & Lis, J. T. (1984). Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(14), 4275–9.

Goodrich, J. A., & Tjian, R. (1994). TBP-TAF complexes: selectivity factors for eukaryotic transcription. *Current Opinion in Cell Biology*, *6*(3), 403–409.

Gorczyca, W., Sun, Z.-Y., Cronin, W., Li, X., Mau, S., & Tugulea, S. (2011). Immunophenotypic pattern of myeloid populations by flow cytometry analysis. *Methods in cell biology*, *103*, 221–66.

Goto, T., Macdonald, P., & Maniatis, T. (1989). Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell*, *57*(3), 413–22.

Grandori, C, Cowley, S. M., James, L. P., & Eisenman, R. N. (2000). The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annual review of cell and developmental biology*, *16*, 653–99.

Grandori, C, Gomez-Roman, N., Felton-Edkins, Z. A., Ngouenet, C., Galloway, D. A., Eisenman, R. N., & White, R. J. (2005). c-Myc binds to human ribosomal DNA and stimulates transcription of rRNA genes by RNA polymerase I. *Nature cell biology*, *7*(3), 311–8.

Gregory, T. R. (2005). Synergy between sequence and size in large-scale genomics. *Nature reviews. Genetics*, *6*(9), 699–708.

Grosschedl, R., & Birnstiel, M. L. (1980). Spacer DNA sequences upstream of the T-A-T-A-A-A-T-A sequence are essential for promotion of H2A histone gene transcription in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, *77*(12), 7102–6.

Grosveld, F., van Assendelft, G. B., Greaves, D. R., & Kollias, G. (1987). Position-independent, high-level expression of the human beta-globin gene in transgenic mice. *Cell*, *51*(6), 975–85.

Gu, M., & Lima, C. D. (2005). Processing the message: structural insights into capping and decapping mRNA. *Current opinion in structural biology*, *15*(1), 99–106.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., … van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, *453*(7197), 948–51.

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, *130*(1), 77–88.

Guo, Y., Niu, C., Breslin, P., Tang, M., Zhang, S., Wei, W., … Zhang, J. (2009). c-Myc-mediated control of cell fate in megakaryocyte-erythrocyte progenitors. *Blood*, *114*(10), 2097–106.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nature structural & molecular biology*, *11*(5), 394–403.

Hamilton, A. J., & Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, N.Y.)*, *286*(5441), 950–2.

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., … Wei, C.-L. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature genetics*, *43*(7), 630–8.

Harris, M. B., Mostecki, J., & Rothman, P. B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. *The Journal of biological chemistry*, *280*(13), 13114–21.

Hatch, N., & Sarid, J. (1994). Glial fibrillary acidic protein transcriptional regulation is independent of a TFIID-binding downstream initiator sequence. *Journal of neurochemistry*, *63*(6), 2003–9.

He, S., Sun, J.-M., Li, L., & Davie, J. R. (2005). Differential intranuclear organization of transcription factors Sp1 and Sp3. *Molecular biology of the cell*, *16*(9), 4073–83.

Hecht, A., Strahl-Bolsinger, S., & Grunstein, M. (1996). Spreading of transcriptional repressor SIR3 from telomeric heterochromatin. *Nature*, *383*(6595), 92–6.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, *39*(3), 311–8.

Hérault, Y., Rassoulzadegan, M., Cuzin, F., & Duboule, D. (1998). Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nature genetics*, *20*(4), 381–4.

Hill, R. E., Favor, J., Hogan, B. L., Ton, C. C., Saunders, G. F., Hanson, I. M., … van Heyningen, V. (1991). Mouse small eye results from mutations in a paired-like homeobox-containing gene. *Nature*, *354*(6354), 522–5.

Ho, J. W. K., Bishop, E., Karchenko, P. V, Nègre, N., White, K. P., & Park, P. J. (2011). ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. *BMC genomics*, *12*, 134.

Hon, G., Wang, W., & Ren, B. (2009). Discovery and annotation of functional chromatin signatures in the human genome. *PLoS computational biology*, *5*(11), e1000566.

Horie, K., Yusa, K., Yae, K., Odajima, J., Fischer, S. E. J., Keng, V. W., … Takeda, J. (2003). Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Molecular and cellular biology*, *23*(24), 9189–207.

Hsiang, M. W., & Cole, R. D. (1977). Structure of histone H1-DNA complex: effect of histone H1 on DNA condensation. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 4852–6.

Huang, C.-Y., Bredemeyer, A. L., Walker, L. M., Bassing, C. H., & Sleckman, B. P. (2008). Dynamic regulation of c-Myc proto-oncogene expression during lymphocyte development revealed by a GFP-c-Myc knock-in mouse. *European journal of immunology*, *38*(2), 342–9.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–45.

Imsiecke, G., Custodio, M., Borojevic, R., Steffen, R., Moustafa, M. A., & Müller, W. E. (1995). Genome size and chromosomes in marine sponges [Suberites Domuncula, Geodia Cydonium]. *Cell biology international*, *19*(12), 995–1000.

Ioshikhes, I. P., & Zhang, M. Q. (2000). Large-scale human promoter mapping using CpG islands. *Nature genetics*, *26*(1), 61–3.

Ivics, Z., Hackett, P. B., Plasterk, R. H., & Izsvák, Z. (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*, *91*(4), 501–10.

JACOB, F., & MONOD, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, *3*, 318–56.

Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M., & Epstein, D. J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development (Cambridge, England)*, *133*(4), 761–72.

Jin, H., Stojnic, R., Adryan, B., Ozdemir, A., Stathopoulos, A., & Frasch, M. (2013). Genome-wide screens for in vivo Tinman binding sites identify cardiac enhancers with diverse functional architectures. *PLoS genetics*, *9*(1), e1003195.

Jin, K.-S., Park, C. M., & Lee, Y.-W. (2012). Identification of differentially expressed genes by 2,3,7,8-tetrachlorodibenzo-p-dioxin in human bronchial epithelial cells. *Human & experimental toxicology*, *31*(1), 107–12.

Jin, Y.-R., Han, X. H., Taketo, M. M., & Yoon, J. K. (2012). Wnt9b-dependent FGF signaling is crucial for outgrowth of the nasal and maxillary processes during upper jaw and lip development. *Development (Cambridge, England)*, *139*(10), 1821–30.

Johnson, P., & Friedmann, T. (1990). Limited bidirectional activity of two housekeeping gene promoters: human HPRT and PGK. *Gene*, *88*(2), 207–13.

Jones, K. A., Yamamoto, K. R., & Tjian, R. (1985). Two distinct transcription factors bind to the HSV thymidine kinase promoter in vitro. *Cell*, *42*(2), 559–572.

Joos, S., Falk, M. H., Lichter, P., Haluska, F. G., Henglein, B., Lenoir, G. M., & Bornkamm, G. W. (1992). Variable breakpoints in Burkitt lymphoma cells with chromosomal t(8;14) translocation separate c-myc and the IgH locus up to several hundred kb. *Human molecular genetics*, *1*(8), 625–32.

Jugessur, A., Shi, M., Gjessing, H. K., Lie, R. T., Wilcox, A. J., Weinberg, C. R., … Murray, J. C. (2009). Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PloS one*, *4*(4), e5385.

Juriloff, D. M., Harris, M. J., McMahon, A. P., Carroll, T. J., & Lidral, A. C. (2006). Wnt9b is the mutated gene involved in multifactorial nonsyndromic cleft lip with or without cleft palate in A/WySn mice, as confirmed by a genetic complementation test. *Birth defects research. Part A, Clinical and molecular teratology*, *76*(8), 574–9.

Juven-Gershon, T., & Kadonaga, J. T. (2010). Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology*, *339*(2), 225–9.

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley interdisciplinary reviews. Developmental biology*, *1*(1), 40–51.

Kaneko, S., Li, G., Son, J., Xu, C.-F., Margueron, R., Neubert, T. A., & Reinberg, D. (2010). Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes & development*, *24*(23), 2615–20.

Kano, H., Kurahashi, H., & Toda, T. (2007). Genetically regulated epigenetic transcriptional activation of retrotransposon insertion confers mouse dactylaplasia phenotype. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19034–9.

Karolyi, J., Erickson, R. P., & Liu, S. (1988). Genetics of susceptibility to 6-aminonicotinamide-induced cleft palate in the mouse: studies in congenic and recombinant inbred strains. *Teratology*, *37*(3), 283–7.

Kavanagh, D. H., Dwyer, S., O'Donovan, M. C., & Owen, M. J. (2013). The ENCODE project: implications for psychiatric genetics. *Molecular psychiatry*, *18*(5), 540–2.

Keng, V. W., Yae, K., Hayakawa, T., Mizuno, S., Uno, Y., Yusa, K., … Takeda, J. (2005). Region-specific saturation germline mutagenesis in mice using the Sleeping Beauty transposon system. *Nature methods*, *2*(10), 763–9.

Kiel, M. J., Yilmaz, O. H., Iwashita, T., Yilmaz, O. H., Terhorst, C., & Morrison, S. J. (2005). SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell*, *121*(7), 1109–21.

Kim, I., He, S., Yilmaz, O. H., Kiel, M. J., & Morrison, S. J. (2006). Enhanced purification of fetal liver hematopoietic stem cells using SLAM family receptors. *Blood*, *108*(2), 737–44.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., … Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, *128*(6), 1231–45.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–7.

Kimura-Yoshida, C., Kitajima, K., Oda-Ishii, I., Tian, E., Suzuki, M., Yamamoto, M., … Matsuo, I. (2004). Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development (Cambridge, England)*, *131*(1), 57–71.

Kioussis, D., Vanin, E., DeLange, T., Flavell, R. A., & Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*, *306*(5944), 662–6.

Kirov, N. C., Lieberman, P. M., & Rushlow, C. (1996). The transcriptional corepressor DSP1 inhibits activated transcription by disrupting TFIIA-TBP complex formation. *The EMBO journal*, *15*(24), 7079–87.

Kleinjan, D. A., Seawright, A., Schedl, A., Quinlan, R. A., Danes, S., & van Heyningen, V. (2001). Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. *Human molecular genetics*, *10*(19), 2049–59.

Kleinjan, D. A, & Lettice, L. A. (2008). Long-range gene control and genetic disease. *Advances in genetics*, *61*, 339–88.

Kogenaru, S., Qing, Y., Guo, Y., & Wang, N. (2012). RNA-seq and microarray complement each other in transcriptome profiling. *BMC genomics*, *13*, 629.

Kondo, M., Weissman, I. L., & Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*, *91*(5), 661–72.

Kothary, R., Perry, M. D., Moran, L. A., & Rossant, J. (1987). Cell-lineage-specific expression of the mouse hsp68 gene during embryogenesis. *Developmental biology*, *121*(2), 342–8.

Kouskoura, T., Kozlova, A., Alexiou, M., Blumer, S., Zouvelou, V., Katsaros, C., … Graf, D. (2013). The etiology of cleft palate formation in BMP7-deficient mice. *PloS one*, *8*(3), e59463.

Kovalchuk, A. L., Qi, C. F., Torrey, T. A., Taddesse-Heath, L., Feigenbaum, L., Park, S. S., … Morse, H. C. (2000). Burkitt lymphoma in the mouse. *The Journal of experimental medicine*, *192*(8), 1183–90.

Kozmik, Z. (2005). Pax genes in eye development and evolution. *Current opinion in genetics & development*, *15*(4), 430–8.

Krivega, I., & Dean, A. (2012). Enhancer and promoter interactions-long distance calls. *Current opinion in genetics & development*, *22*(2), 79–85.

Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., … Bernstein, B. E. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. (B. van Steensel, Ed.)*PLoS genetics*, *4*(10), e1000242.

Kubo, H., Shimizu, M., Taya, Y., Kawamoto, T., Michida, M., Kaneko, E., … Kato, Y. (2009). Identification of mesenchymal stem cell (MSC)-transcription factors by microarray and knockdown analyses, and signature molecule-marked MSC in bone marrow by immunohistochemistry. *Genes to cells : devoted to molecular & cellular mechanisms*, *14*(3), 407–24.

Kukulski, W., Schorb, M., Welsch, S., Picco, A., Kaksonen, M., & Briggs, J. A. G. (2011). Correlated fluorescence and 3D electron microscopy with high sensitivity and spatial precision. *The Journal of cell biology*, *192*(1), 111–9.

Kulkarni, M. M., & Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development (Cambridge, England)*, *130*(26), 6569–75.

Kulozik, A. E., Bellan-Koch, A., Bail, S., Kohne, E., & Kleihauer, E. (1991). Thalassemia intermedia: moderate reduction of beta globin gene transcriptional activity by a novel mutation of the proximal CACCC promoter element. *Blood*, *77*(9), 2054–8.

Kumaran, R. I., & Spector, D. L. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *The Journal of cell biology*, *180*(1), 51–65.

Kurukuti, S., Tiwari, V. K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., … Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(28), 10684–9.

Lamb, B. T., & Gearhart, J. D. (1995). YAC transgenics and the study of genetics and human disease. *Current opinion in genetics & development*, *5*(3), 342–8.

Langowski, J. (2010). Chromosome conformation by crosslinking: polymer physics matters. *Nucleus (Austin, Tex.)*, *1*(1), 37–9.

Lazo, P. A., Lee, J. S., & Tsichlis, P. N. (1990). Long-distance activation of the Myc protooncogene by provirus insertion in Mlvi-1 or Mlvi-4 in rat T-cell lymphomas. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(1), 170–3.

Lee, M. P., Howcroft, K., Kotekar, A., Yang, H. H., Buetow, K. H., & Singer, D. S. (2005). ATG deserts define a novel core promoter subclass. *Genome research*, *15*(9), 1189–97.

Leprivier, G., Remke, M., Rotblat, B., Dubuc, A., Mateo, A.-R. F., Kool, M., … Sorensen, P. H. (2013). The eEF2 kinase confers resistance to nutrient deprivation by blocking translation elongation. *Cell*, *153*(5), 1064–79.

Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., … Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(11), 7548–53.

Lettice, L. A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P. S., Gautier, P., … FitzPatrick, D. R. (2011). Enhancer-adoption as a mechanism of human developmental disease. *Human mutation*, *32*(12), 1492–9.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. *Nature*, *276*(5688), 565–70.

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–19.

Li, F., Mao, G., Tong, D., Huang, J., Gu, L., Yang, W., & Li, G.-M. (2013). The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell*, *153*(3), 590–600.

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1-2), 84–98.

Li, L. M., & Arnosti, D. N. (2011). Long- and short-range transcriptional repressors induce distinct chromatin states on repressed genes. *Current biology : CB*, *21*(5), 406–12.

Li, M. A., Pettitt, S. J., Eckert, S., Ning, Z., Rice, S., Cadiñanos, J., … Bradley, A. (2013). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Molecular and cellular biology*, *33*(7), 1317–30.

Li, Q., Peterson, K. R., Fang, X., & Stamatoyannopoulos, G. (2002). Locus control regions. *Blood*, *100*(9), 3077–86.

Li, Z., Schug, J., Tuteja, G., White, P., & Kaestner, K. H. (2011). The nucleosome map of the mammalian liver. *Nature structural & molecular biology*, *18*(6), 742–6.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–93.

Lin, J. M., Collins, P. J., Trinklein, N. D., Fu, Y., Xi, H., Myers, R. M., & Weng, Z. (2007). Transcription factor binding and modified histones in human bidirectional promoters. *Genome research*, *17*(6), 818–27.

Lin, Y. S., & Green, M. R. (1989). Similarities between prokaryotic and eukaryotic cyclic AMP-responsive promoter elements. *Nature*, *340*(6235), 656–9.

Litt, M. D., Simpson, M., Gaszner, M., Allis, C. D., & Felsenfeld, G. (2001). Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. *Science (New York, N.Y.)*, *293*(5539), 2453–5.

Liu, B., Maul, R. S., & Kaetzel, D. M. (1996). Repression of platelet-derived growth factor A-chain gene transcription by an upstream silencer element. Participation by sequence-specific single-stranded DNA-binding proteins. *The Journal of biological chemistry*, *271*(42), 26281–90.

Liu, F., van der Lijn, F., Schurmann, C., Zhu, G., Chakravarty, M. M., Hysi, P. G., … Kayser, M. (2012). A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS genetics*, *8*(9), e1002932.

Liu, J., & Krantz, I. D. (2009). Cornelia de Lange syndrome, cohesin, and beyond. *Clinical genetics*, *76*(4), 303–14.

Lonfat, N., Montavon, T., Jebb, D., Tschopp, P., Nguyen Huynh, T. H., Zakany, J., & Duboule, D. (2013). Transgene- and locus-dependent imprinting reveals allele-specific chromosome conformations. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(29), 11946–51.

Lovén, J., Orlando, D. A., Sigova, A. A., Lin, C. Y., Rahl, P. B., Burge, C. B., … Young, R. A. (2012). Revisiting global gene expression analysis. *Cell*, *151*(3), 476–82.

Luco, R. F., Pan, Q., Tominaga, K., Blencowe, B. J., Pereira-Smith, O. M., & Misteli, T. (2010). Regulation of alternative splicing by histone modifications. *Science*, *327*(5968), 996–1000.

Ludwig, M., Patel, N., & Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, *125*(5), 949–958.

Lundgren, M., Chow, C. M., Sabbattini, P., Georgiou, A., Minaee, S., & Dillon, N. (2000). Transcription factor dosage affects changes in higher order chromatin structure associated with activation of a heterochromatic gene. *Cell*, *103*(5), 733–43.

Luo, G., Ivics, Z., Izsvák, Z., & Bradley, A. (1998). Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(18), 10769–73.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annual review of microbiology*, *60*, 327–49.

Malynn, B. A., de Alboran, I. M., O'Hagan, R. C., Bronson, R., Davidson, L., DePinho, R. A., & Alt, F. W. (2000). N-myc can functionally replace c-myc in murine development, cellular growth, and differentiation. *Genes & development*, *14*(11), 1390–9.

Mangold, E., Ludwig, K. U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., … Nöthen, M. M. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nature genetics*, *42*(1), 24–6.

Mantovani, R. (1998). A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Research*, *26*(5), 1135–1143.

Marandel, L., Labbe, C., Bobe, J., & Le Bail, P.-Y. (2012). Evolutionary history of c-myc in teleosts and characterization of the duplicated c-myca genes in goldfish embryos. *Molecular reproduction and development*, *79*(2), 85–96.

Margueron, R., & Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. *Nature*, *469*(7330), 343–9.

Marinić, M., Aktas, T., Ruf, S., & Spitz, F. (2013). An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Developmental cell*, *24*(5), 530–42.

Martinelli, M., Scapoli, L., Palmieri, A., Pezzetti, F., Baciliero, U., Padula, E., … Carinci, F. (2006). Study of four genes belonging to the folate pathway: transcobalamin

2 is involved in the onset of non-syndromic cleft lip with or without cleft palate. *Human mutation*, *27*(3), 294.

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annual review of genomics and human genetics*, *7*, 29–59.

Mateyak, M. K., Obaya, A. J., & Sedivy, J. M. (1999). c-Myc regulates cyclin D-Cdk4 and -Cdk6 activity but affects cell cycle progression at multiple independent points. *Molecular and cellular biology*, *19*(7), 4672–83.

May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., … Visel, A. (2012). Large-scale discovery of enhancers from human heart tissue. *Nature genetics*, *44*(1), 89–93.

McBride, D. J., Buckle, A., van Heyningen, V., & Kleinjan, D. A. (2011). DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 cis-regulatory region. *PloS one*, *6*(12), e28616.

McKnight, S. L., & Kingsbury, R. (1982). Transcriptional control signals of a eukaryotic protein-coding gene. *Science (New York, N.Y.)*, *217*(4557), 316–24.

Merika, M, Williams, A. J., Chen, G., Collins, T., & Thanos, D. (1998). Recruitment of CBP/p300 by the IFN beta enhanceosome is required for synergistic activation of transcription. *Molecular cell*, *1*(2), 277–87.

Merika, M., & Thanos, D. (2001). Enhanceosomes. *Current Opinion in Genetics & Development*, *11*(2), 205–208.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., … Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, *448*(7153), 553–60.

Mikkola, H. K. A., & Orkin, S. H. (2006). The journey of developing hematopoietic stem cells. *Development (Cambridge, England)*, *133*(19), 3733–44.

Mitchell, J. A., & Fraser, P. (2008). Transcription factories are nuclear subcompartments that remain in the absence of transcription. *Genes & development*, *22*(1), 20–5.

Mitchell, R. S., Beitzel, B. F., Schroder, A. R. W., Shinn, P., Chen, H., Berry, C. C., … Bushman, F. D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS biology*, *2*(8), E234.

Monod, J. (1966). From enzymatic adaptation to allosteric transitions. *Science (New York, N.Y.)*, *154*(3748), 475–83.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., … Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. *Cell*, *147*(5), 1132–45.

Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D. A., Adams, C. M., Azubel, M., … Kornberg, R. D. (2013). Architecture of an RNA Polymerase II Transcription Pre-Initiation Complex. *Science (New York, N.Y.)*.

Muravyova, E., Golovnin, A., Gracheva, E., Parshikov, A., Belenkaya, T., Pirrotta, V., & Georgiev, P. (2001). Loss of insulator activity by paired Su(Hw) chromatin insulators. *Science (New York, N.Y.)*, *291*(5503), 495–8.

Murray, T., Taub, M. A., Ruczinski, I., Scott, A. F., Hetmanski, J. B., Schwender, H., … Beaty, T. H. (2012). Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between Asians and Europeans. *Genetic epidemiology*, *36*(4), 392–9.

Myers, L. C., & Kornberg, R. D. (2000). Mediator of transcriptional regulation. *Annual review of biochemistry*, *69*, 729–49.

Nakabayashi, H., Hashimoto, T., Miyao, Y., Tjong, K. K., Chan, J., & Tamaoki, T. (1991). A position-dependent silencer plays a major role in repressing alpha-fetoprotein expression in human hepatoma. *Molecular and cellular biology*, *11*(12), 5885–93.

Nakashima, K., Arai, S., Suzuki, A., Nariai, Y., Urano, T., Nakayama, M., … Miyazaki, T. (2013). PAD4 regulates proliferation of multipotent haematopoietic cells by controlling c-myc expression. *Nature communications*, *4*, 1836.

Nanni, L., Ming, J. E., Bocian, M., Steinhaus, K., Bianchi, D. W., Die-Smulders, C., … Muenke, M. (1999). The mutational spectrum of the sonic hedgehog gene in holoprosencephaly: SHH mutations cause a significant proportion of autosomal dominant holoprosencephaly. *Human molecular genetics*, *8*(13), 2479–88.

Natesan, S., & Gilman, M. Z. (1993). DNA bending and orientation-dependent function of YY1 in the c-fos promoter. *Genes & development*, *7*(12B), 2497–509.

Navas, P. A., Peterson, K. R., Li, Q., Skarpidi, E., Rohde, A., Shaw, S. E., … Stamatoyannopoulos, G. (1998). Developmental specificity of the interaction between the locus control region and embryonic or fetal globin genes in transgenic mice with an HS3 core deletion. *Molecular and cellular biology*, *18*(7), 4188–96.

Németh, A., Guibert, S., Tiwari, V. K., Ohlsson, R., & Längst, G. (2008). Epigenetic regulation of TTF-I-mediated promoter-terminator interactions of rRNA genes. *The EMBO journal*, *27*(8), 1255–65.

Nie, Z., Hu, G., Wei, G., Cui, K., Yamane, A., Resch, W., … Levens, D. (2012). c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell*, *151*(1), 68–79.

Nieuwenhuis, E., & Hui, C. (2005). Hedgehog signaling and congenital malformations. *Clinical genetics*, *67*(3), 193–208.

Nobrega, M. A., Ovcharenko, I., Afzal, V., & Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science (New York, N.Y.)*, *302*(5644), 413.

Nóbrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V., & Rubin, E. M. (2004). Megabase deletions of gene deserts result in viable mice. *Nature*, *431*(7011), 988–93.

Noordermeer, D., & Duboule, D. (n.d.). Chromatin looping and organization at developmentally regulated gene loci. *Wiley interdisciplinary reviews. Developmental biology*, *2*(5), 615–30.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., … Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), 381–5.

Nourbakhsh, M., Hoffmann, K., & Hauser, H. (1993). Interferon-beta promoters contain a DNA element that acts as a position-independent silencer on the NF-kappa B site. *The EMBO journal*, *12*(2), 451–9.

Obaya, A. J., Mateyak, M. K., & Sedivy, J. M. (1999). Mysterious liaisons: the relationship between c-Myc and the cell cycle. *Oncogene*, *18*(19), 2934–41.

Odent, S., Atti-Bitach, T., Blayau, M., Mathieu, M., Aug, J., Delezo de, A. L., … Vekemans, M. (1999). Expression of the Sonic hedgehog (SHH) gene during early human development and phenotypic expression of new mutations causing holoprosencephaly. *Human molecular genetics*, *8*(9), 1683–9

Ogbourne, S., & Antalis, T. M. (1998). Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *The Biochemical journal*, *331 (Pt 1*, 1–14.

Ohno, S. (1972). So much “junk” DNA in our genome. *Brookhaven symposia in biology*, *23*, 366–70.

Orlando, V., & Paro, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell*, *75*(6), 1187–98.

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & development*, *10*(21), 2657–83.

Osawa, M., Hanada, K., Hamada, H., & Nakauchi, H. (1996). Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science (New York, N.Y.)*, *273*(5272), 242–5.

Osborne, C. S., Chakalova, L., Brown, K. E., Carter, D., Horton, A., Debrand, E., … Fraser, P. (2004). Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, *36*(10), 1065–71.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W., & Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome research*, *15*(1), 137–45.

Palomero, T., Lim, W. K., Odom, D. T., Sulis, M. L., Real, P. J., Margolin, A., … Ferrando, A. A. (2006). NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(48), 18261–6.

Palstra, R.-J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., & de Laat, W. (2003). The beta-globin nuclear compartment in development and erythroid differentiation. *Nature genetics*, *35*(2), 190–4.

Panne, D. (2008). The enhanceosome. *Current opinion in structural biology*, *18*(2), 236–42.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., … Merkenschlager, M. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, *132*(3), 422–33.

Parsons, T. E., Kristensen, E., Hornung, L., Diewert, V. M., Boyd, S. K., German, R. Z., & Hallgrímsson, B. (2008). Phenotypic variability and craniofacial dysmorphology: increased shape variance in a mouse model for cleft lip. *Journal of anatomy*, *212*(2), 135–43.

Patrushev, L. I., & Minkevich, I. G. (2008). The problem of the eukaryotic genome size. *Biochemistry. Biokhimii͡a*, *73*(13), 1519–52.

Pedersen, R. A. (1971). DNA content, ribosomal gene multiplicity, and cell size in fish. *The Journal of experimental zoology*, *177*(1), 65–78.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., … Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, *444*(7118), 499–502.

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Current biology : CB*, *20*(17), 1562–7.

Peterson, K. R., Clegg, C. H., Navas, P. A., Norton, E. J., Kimbrough, T. G., & Stamatoyannopoulos, G. (1996). Effect of deletion of 5'HS3 or 5'HS2 of the human beta-globin locus control region on the developmental regulation of globin gene expression in beta-globin locus yeast artificial chromosome transgenic mice. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(13), 6605–9.

Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., … Corces, V. G. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, *153*(6), 1281–95.

Poulin, F., Nobrega, M. A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E. M., & Pennacchio, L. A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics*, *85*(6), 774–81.

Pshenichnaya, I., Schouwey, K., Armaro, M., Larue, L., Knoepfler, P. S., Eisenman, R. N., … Beermann, F. (2012). Constitutive gray hair in mice induced by melanocyte-specific deletion of c-Myc. *Pigment cell & melanoma research*, *25*(3), 312–25.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. *Nature*, *322*(6081), 697–701.

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., … Rokhsar, D. S. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*(7198), 1064–71.

Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S. A., Swigut, T., & Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. *Cell stem cell*, *11*(5), 633–48.

Radtke, I., Mullighan, C. G., Ishii, M., Su, X., Cheng, J., Ma, J., … Downing, J. R. (2009). Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12944–9.

Rahimov, F., Marazita, M. L., Visel, A., Cooper, M. E., Hitchler, M. J., Rubini, M., … Murray, J. C. (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nature genetics*, *40*(11), 1341–7.

Ramos, E., Ghosh, D., Baxter, E., & Corces, V. G. (2006). Genomic organization of gypsy chromatin insulators in Drosophila melanogaster. *Genetics*, *172*(4), 2337–49.

Ranish, J. A., Yudkovsky, N., & Hahn, S. (1999). Intermediates in formation and activity of the RNA polymerase II preinitiation complex: holoenzyme recruitment and a postrecruitment role for the TATA box and TFIIB. *Genes & Development*, *13*(1), 49–63.

Recillas-Targa, F, Bell, A. C., & Felsenfeld, G. (1999). Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(25), 14354–9.

Recillas-Targa, Félix, Pikaart, M. J., Burgess-Beusse, B., Bell, A. C., Litt, M. D., West, A. G., … Felsenfeld, G. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(10), 6883–8.

Reddy, K. L., Zullo, J. M., Bertolino, E., & Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, *452*(7184), 243–7.

Reed, S. C. (1933). An Embryological Study Of Hairlip in Mice. *The Anatomical Record*, *56*(2), 101–117.

Rieger, M. A., & Schroeder, T. (2012). Hematopoiesis. *Cold Spring Harbor perspectives in biology*, *4*(12).

Riethoven, J.-J. M. (2010). Regulatory regions in DNA: promoters, enhancers, silencers, and insulators. *Methods in molecular biology (Clifton, N.J.)*, *674*, 33–42.

Ringrose, L., & Paro, R. (2007). Polycomb/Trithorax response elements and epigenetic memory of cell identity. *Development (Cambridge, England)*, *134*(2), 223–32.

Robledo, R. F., Rajan, L., Li, X., & Lufkin, T. (2002). The Dlx5 and Dlx6 homeobox genes are essential for craniofacial, axial, and appendicular skeletal development. *Genes & development*, *16*(9), 1089–101.

Roesler, W. J., Vandenbark, G. R., & Hanson, R. W. (1989). Identification of multiple protein binding domains in the promoter-regulatory region of the phosphoenolpyruvate carboxykinase (GTP) gene. *The Journal of biological chemistry*, *264*(16), 9657–64.

Rollins, R. A., Morcillo, P., & Dorsett, D. (1999). Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics*, *152*(2), 577–93.

Roseman, R. R., Swan, J. M., & Geyer, P. K. (1995). A Drosophila insulator protein facilitates dosage compensation of the X chromosome min-white gene located at autosomal insertion sites. *Development (Cambridge, England)*, *121*(11), 3573–82.

Rossi, F. M., Kringstein, A. M., Spicher, A., Guicherit, O. M., & Blau, H. M. (2000). Transcriptional control: rheostat converted to on/off switch. *Molecular cell*, *6*(3), 723–8.

Rudra, D., & Warner, J. R. (2004). What better measure than ribosome synthesis? *Genes & development*, *18*(20), 2431–6.

Ruf, S., Symmons, O., Uslu, V. V., Dolle, D., Hot, C., Ettwiller, L., & Spitz, F. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature genetics*, *43*(4), 379–86.

Saito, H., Yamamura, K., & Suzuki, N. (2012). Reduced bone morphogenetic protein receptor type 1A signaling in neural-crest-derived cells causes facial dysmorphism. *Disease models & mechanisms*, *5*(6), 948–55.

Sakurai, M., & Strominger, J. L. (1988). B-cell-specific enhancer activity of conserved upstream elements of the class II major histocompatibility complex DQB gene. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(18), 6909–13.

Sambrook, J., & Russell, D. W. (2006). Molecular Cloning. *CSH protocols*, *2006*(4).

Sandelin, A., Carninci, P., Lenhard, B., Ponjavic, J., Hayashizaki, Y., & Hume, D. A. (2007). Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nature reviews. Genetics*, *8*(6), 424–36.

Satokata, I., & Maas, R. (1994). Msx1 deficient mice exhibit cleft palate and abnormalities of craniofacial and tooth development. *Nature genetics*, *6*(4), 348–56.

Sawado, T., Igarashi, K., & Groudine, M. (2001). Activation of beta-major globin gene transcription is associated with recruitment of NF-E2 to the beta-globin LCR and gene promoter. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(18), 10226–31.

Scheer, N., & Campos-Ortega, J. A. (1999). Use of the Gal4-UAS technique for targeted gene expression in the zebrafish. *Mechanisms of development*, *80*(2), 153–8.

Schmitt, S., Prestel, M., & Paro, R. (2005). Intergenic transcription through a polycomb group response element counteracts silencing. *Genes & development*, *19*(6), 697–708.

Schuettengruber, B., & Cavalli, G. (2009). Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. *Development (Cambridge, England)*, *136*(21), 3531–42.

Schuster-Böckler, B., & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature*, *488*(7412), 504–7.

Seitan, V. C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K. E., … Merkenschlager, M. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature*, *476*(7361), 467–71.

Sharpe, J., Lettice, L., Hecksher-Sorensen, J., Fox, M., Hill, R., & Krumlauf, R. (1999). Identification of sonic hedgehog as a candidate gene responsible for the polydactylous mouse mutant Sasquatch. *Current biology : CB*, *9*(2), 97–100.

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., … Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*, *488*(7409), 116–20.

Shopland, L. S., Lynch, C. R., Peterson, K. A., Thornton, K., Kepper, N., Hase, J. von, … O'Brien, T. P. (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence.

Shumaker, D. K., Dechat, T., Kohlmaier, A., Adam, S. A., Bozovsky, M. R., Erdos, M. R., … Goldman, R. D. (2006). Mutant nuclear lamin A leads to progressive alterations of epigenetic control in premature aging. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8703–8.

Sidman, C. L., Denial, T. M., Marshall, J. D., & Roths, J. B. (1993). Multiple Mechanisms of Tumorigenesis in E{micro}-myc Transgenic Mice. *Cancer Res.*, *53*(7), 1665–1669.

Siebenlist, U., Simpson, R. B., & Gilbert, W. (1980). E. coli RNA polymerase interacts homologously with two different promoters. *Cell*, *20*(2), 269–81.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., … de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature genetics*, *38*(11), 1348–54.

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual review of biochemistry*, *72*, 449–79.

Smith, N. G. C., Brandström, M., & Ellegren, H. (2004). Evidence for turnover of functional noncoding DNA in mammalian genome evolution. *Genomics*, *84*(5), 806–13.

Sock, E., Rettig, S. D., Enderich, J., Bösl, M. R., Tamm, E. R., & Wegner, M. (2004). Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Molecular and cellular biology*, *24*(15), 6635–44.

Solomon, M. J., & Varshavsky, A. (1985). Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences of the United States of America*, *82*(19), 6470–4.

Solomon, M. J., Larsen, P. L., & Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell*, *53*(6), 937–47.

Song, L., Li, Y., Wang, K., Wang, Y.-Z., Molotkov, A., Gao, L., … Zhou, C. J. (2009). Lrp6-mediated canonical Wnt signaling is required for lip formation and fusion. *Development (Cambridge, England)*, *136*(18), 3161–71.

Song, T., Wu, D., Wang, Y., Li, H., Yin, N., & Zhao, Z. (2013). Association of NAT1 and NAT2 genes with nonsyndromic cleft lip and palate. *Molecular medicine reports*, *8*(1), 211–6.

Spangrude, G. J., Heimfeld, S., & Weissman, I. L. (1988). Purification and characterization of mouse hematopoietic stem cells. *Science (New York, N.Y.)*, *241*(4861), 58–62.

Spellman, P. T., & Rubin, G. M. (2002). Evidence for large domains of similarly expressed genes in the Drosophila genome. *Journal of biology*, *1*(1), 5.

Spitz, F, & Duboule, D. (2008). Global control regions and regulatory landscapes in vertebrate development and evolution. *Advances in genetics*, *61*, 175–205.

Spitz, F, Gonzalez, F., & Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*, *113*(3), 405–17.

Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., … de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, *20*(17), 2349–54.

Splinter, E., de Wit, E., Nora, E. P., Klous, P., van de Werken, H. J. G., Zhu, Y., … de Laat, W. (2011). The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on Xist RNA. *Genes & development*, *25*(13), 1371–83.

Stanojevic, D., Small, S., & Levine, M. (1991). Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science (New York, N.Y.)*, *254*(5036), 1385–7.

Stoelzle, T., Schwarb, P., Trumpp, A., & Hynes, N. E. (2009). c-Myc affects mRNA translation, cell proliferation and progenitor cell function in the mammary gland. *BMC biology*, *7*, 63.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, *98*(1), 1–4.

Struhl, K. (2001). Gene regulation. A paradigm for precision. *Science (New York, N.Y.)*, *293*(5532), 1054–5.

Sur, I. K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., … Taipale, J. (2012). Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science (New York, N.Y.)*, *338*(6112), 1360–3.

Sutherland, H. G., Martin, D. I., & Whitelaw, E. (1997). A globin enhancer acts by increasing the proportion of erythrocytes expressing a linked transgene. *Molecular and cellular biology*, *17*(3), 1607–14.

Taberlay, P. C., Kelly, T. K., Liu, C.-C., You, J. S., De Carvalho, D. D., Miranda, T. B., … Jones, P. A. (2011). Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell*, *147*(6), 1283–94.

Talbot, D., & Grosveld, F. (1991). The 5'HS2 of the globin locus control region enhances transcription through the interaction of a multimeric complex binding at two functionally distinct NF-E2 binding sites. *The EMBO journal*, *10*(6), 1391–8.

Tan-Wong, S. M., French, J. D., Proudfoot, N. J., & Brown, M. A. (2008). Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5160–5.

Taya, Y., O'Kane, S., & Ferguson, M. W. (1999). Pathogenesis of cleft palate in TGF-beta3 knockout mice. *Development (Cambridge, England)*, *126*(17), 3869–79.

Tewari, A. K., Yardimci, G. G., Shibata, Y., Sheffield, N. C., Song, L., Taylor, B. S., … Febbo, P. G. (2012). Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. *Genome biology*, *13*(10), R88.

Tewari, R., Gillemans, N., Harper, A., Wijgerde, M., Zafarana, G., Drabek, D., … Philipsen, S. (1996). The human beta-globin locus control region confers an early embryonic erythroid-specific expression pattern to a basic promoter driving the bacterial lacZ gene. *Development (Cambridge, England)*, *122*(12), 3991–9.

Thanos, D., & Maniatis, T. (1995). Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. *Cell*, *83*(7), 1091–1100.

Thein, S. L. (2005). Genetic modifiers of beta-thalassemia. *Haematologica*, *90*(5), 649–60.

Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell*, *10*(6), 1453–65.

Ton, C. C., Hirvonen, H., Miwa, H., Weil, M. M., Monaghan, P., Jordan, T., … Drechsler, M. (1991). Positional cloning and characterization of a paired box- and homeobox-containing gene from the aniridia region. *Cell*, *67*(6), 1059–74.

Towbin, B. D., González-Aguilera, C., Sack, R., Gaidatzis, D., Kalck, V., Meister, P., … Gasser, S. M. (2012). Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. *Cell*, *150*(5), 934–47.

Trasler, D. G. (1968). Pathogenesis of cleft lip and its relation to embryonic face shape in A-J and C57BL mice. *Teratology*, *1*(1), 33–49.

Traven, A., Jelicic, B., & Sopta, M. (2006). Yeast Gal4: a transcriptional paradigm revisited. *EMBO reports*, *7*(5), 496–9.

Trujillo, M. A., Sakagashira, M., & Eberhardt, N. L. (2006). The human growth hormone gene contains a silencer embedded within an Alu repeat in the 3’-flanking region. *Molecular endocrinology (Baltimore, Md.)*, *20*(10), 2559–75.

Trumpp, A., Refaeli, Y., Oskarsson, T., Gasser, S., Murphy, M., Martin, G. R., & Bishop, J. M. (2001). c-Myc regulates mammalian body size by controlling cell number but not cell size. *Nature*, *414*(6865), 768–73.

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., … Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science (New York, N.Y.)*, *329*(5992), 689–93.

Tuupanen, S., Yan, J., Turunen, M., Gylfe, A. E., Kaasinen, E., Li, L., … Aaltonen, L. A. (n.d.). Characterization of the colorectal cancer-associated enhancer MYC-335 at 8q24: the role of rs67491583. *Cancer genetics*, *205*(1-2), 25–33.

Van de Werken, H. J. G., Landan, G., Holwerda, S. J. B., Hoichman, M., Klous, P., Chachik, R., … de Laat, W. (2012). Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature methods*, *9*(10), 969–72.

Van Riggelen, J., Yetil, A., & Felsher, D. W. (2010). MYC as a regulator of ribosome biogenesis and protein synthesis. *Nature reviews. Cancer*, *10*(4), 301–9.

Vannini, A., & Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular cell*, *45*(4), 439–46.

Vieira, A. R., Orioli, I. M., & Murray, J. C. (2002). Maternal age and oral clefts: a reappraisal. *Oral surgery, oral medicine, oral pathology, oral radiology, and endodontics*, *94*(5), 530–5. Retrieved from

Visel, A., Prabhakar, S., Akiyama, J. A., Shoukry, M., Lewis, K. D., Holt, A., … Pennacchio, L. A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature genetics*, *40*(2), 158–60.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–8.

Visel, A., Rubin, E. M., & Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*, *461*(7261), 199–205.

Visel, A., Zhu, Y., May, D., Afzal, V., Gong, E., Attanasio, C., … Pennacchio, L. A. (2010). Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature*, *464*(7287), 409–12.

Vogel, M. J., Peric-Hupkes, D., & van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nature protocols*, *2*(6), 1467–78.

Walters, M. C., Fiering, S., Eidemiller, J., Magis, W., Groudine, M., & Martin, D. I. (1995). Enhancers increase the probability but not the level of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(15), 7125–9.

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., … Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, *40*(7), 897–903.

Ward, R. E., Bixler, D., & Raywood, E. R. (1989). A study of cephalometric features in cleft lip-cleft palate families. I: Phenotypic heterogeneity and genetic predisposition in parents of sporadic cases. *The Cleft palate journal*, *26*(4), 318–25; discussion 325–6.

Wasserman, N. F., Aneas, I., & Nobrega, M. A. (2010). An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer. *Genome research*, *20*(9), 1191–7.

Wasylyk, B., Wasylyk, C., Augereau, P., & Chambon, P. (1983). The SV40 72 bp repeat preferentially potentiates transcription starting from proximal natural or substitute promoter elements. *Cell*, *32*(2), 503–14.

Weiland, Y., Lemmer, P., & Cremer, C. (2011). Combining FISH with localisation microscopy: Super-resolution imaging of nuclear genome nanostructures. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, *19*(1), 5–23.

Weinberg, S. M., Naidoo, S. D., Bardi, K. M., Brandon, C. A., Neiswanger, K., Resick, J. M., … Marazita, M. L. (2009). Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. *Orthodontics & craniofacial research*, *12*(4), 271–81.

Weintraub, H. (1988). Formation of stable transcription complexes as assayed by analysis of individual templates. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(16), 5819–23.

Welsh, I. C., Hagge-Greenberg, A., & O'Brien, T. P. (2007). A dosage-dependent role for Spry2 in growth and patterning during palate development. *Mechanisms of development*, *124*(9-10), 746–61.

Wen, B., Wu, H., Shinkai, Y., Irizarry, R. A., & Feinberg, A. P. (2009). Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, *41*(2), 246–50.

West, A. G., Gaszner, M., & Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. *Genes & development*, *16*(3), 271–88.

Wierstra, I., & Alves, J. (2008). The c-myc promoter: still MysterY and challenge. *Advances in cancer research*, *99*, 113–333.

Wilson, A., Murphy, M. J., Oskarsson, T., Kaloulis, K., Bettess, M. D., Oser, G. M., … Trumpp, A. (2004). c-Myc controls the balance between hematopoietic stem cell self-renewal and differentiation. *Genes & development*, *18*(22), 2747–63.

Wilson, A., Laurenti, E., Oser, G., van der Wath, R. C., Blanco-Bose, W., Jaworski, M., … Trumpp, A. (2008). Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell*, *135*(6), 1118–29.

Wilson, N. K., Foster, S. D., Wang, X., Knezevic, K., Schütte, J., Kaimakis, P., … Göttgens, B. (2010). Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell stem cell*, *7*(4), 532–44.

Woychik, N. (2002). The RNA Polymerase II MachineryStructure Illuminates Function. *Cell*, *108*(4), 453–463.

Wu, M.-H., Dimopoulos, G., Mantalaris, A., & Varley, J. (2004). The effect of hyperosmotic pressure on antibody production and gene expression in the GS-NS0 cell line. *Biotechnology and applied biochemistry*, *40*(Pt 1), 41–6.

Wyatt, A. W., Osborne, R. J., Stewart, H., & Ragge, N. K. (2010). Bone morphogenetic protein 7 (BMP7) mutations are associated with variable ocular, brain, ear, palate, and skeletal anomalies. *Human mutation*, *31*(7), 781–7.

Xi, H., Yu, Y., Fu, Y., Foley, J., Halees, A., & Weng, Z. (2007). Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome research*, *17*(6), 798–806.

Yagi, K., Furuhashi, M., Aoki, H., Goto, D., Kuwano, H., Sugamura, K., … Kato, M. (2002). c-myc is a downstream target of the Smad pathway. *The Journal of biological chemistry*, *277*(1), 854–61.

Yamada, T., Hirata, A., Sasabe, E., Yoshimura, T., Ohno, S., Kitamura, N., & Yamamoto, T. (2013). TCDD disrupts posterior palatogenesis and causes cleft palate. *Journal of cranio-maxillo-facial surgery : official publication of the European Association for Cranio-Maxillo-Facial Surgery*.

Yant, S. R., Wu, X., Huang, Y., Garrison, B., Burgess, S. M., & Kay, M. A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Molecular and cellular biology*, *25*(6), 2085–94.

Yee, S. P., & Rigby, P. W. (1993). The regulation of myogenin gene expression during the embryonic development of the mouse. *Genes & development*, *7*(7A), 1277–89.

Yilmaz, O. H., Kiel, M. J., & Morrison, S. J. (2006). SLAM family markers are conserved among hematopoietic stem cells from old and reconstituted mice and markedly increase their purity. *Blood*, *107*(3), 924–30.

Yip, K. Y., Cheng, C., Bhardwaj, N., Brown, J. B., Leng, J., Kundaje, A., … Gerstein, M. (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome biology*, *13*(9), R48.

Yoon, Y. S., Jeong, S., Rong, Q., Park, K.-Y., Chung, J. H., & Pfeifer, K. (2007). Analysis of the H19ICR insulator. *Molecular and cellular biology*, *27*(9), 3499–510.

You, Z., Saims, D., Chen, S., Zhang, Z., Guttridge, D. C., Guan, K.-L., … Wang, C.-Y. (2002). Wnt signaling promotes oncogenic transformation by inhibiting c-Myc-induced apoptosis. *The Journal of cell biology*, *157*(3), 429–40.

Yusufzai, T. M., & Felsenfeld, G. (2004). The 5'-HS4 chicken beta-globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(23), 8620–4.

Zhao, H., & Dean, A. (2004). An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic acids research*, *32*(16), 4903–19.

Zheng-Bradley, X., Rung, J., Parkinson, H., & Brazma, A. (2010). Large scale comparison of global gene expression patterns in human and mouse. *Genome biology*, *11*(12), R124.

Zhou, V. W., Goren, A., & Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature reviews. Genetics*, *12*(1), 7–18.

Zhu, X., Ozturk, F., Liu, C., Oakley, G. G., & Nawshad, A. (2012). Transforming growth factor-β activates c-Myc to promote palatal growth. *Journal of cellular biochemistry*, *113*(10), 3069–85.

Zucchero, T. M., Cooper, M. E., Maher, B. S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., … Murray, J. C. (2004). Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *The New England journal of medicine*, *351*(8), 769–80.

8. APPENDIX

8.1 List Of Figures

8.2 List Of Tables

8.3 List of Abbreviations

$(NH_4)_2SO_4$	Ammonium sulfate
#	Number
%	Percent
°C	Degree celsius
<	Inferior
>	Superior
∞	Infinity
3C	Chromosome Conformation Capture
3D	Three dimensions
4C	Circularized 3C
HS	Hypersensitive Site
5C	3C Carbon Copy
6-AN	6-aminonicotinamide
ac	Acetylation
AML	Acute myeloid leukemia
APC	Anaphase promoting complex
AR	Androgen receptor
ATP	Adenosine triphosphate
ba	Branchial arch
BAC	Bacterial artificial chromosome
BCIP	5-Bromo-4-chloro-3-indolyl phosphate
BCL	B-cell lymphoma
BM	Bone marrow
BMP	Bone morphogenetic protein
bp	Base pair
BRCA	Breast cancer
BRE	TFIIB recognition element
BSA	Bovine serum albumin
C/EBP	CCAAT-enhancer binding protein
$CaCl_2$	Calcium chloride
cAMP	cyclic adenosine monophosphate
CC3	Cleaved caspase3
CD1	Outbred Mouse Line
CdLS	Cornelia de lange syndrome
Cen_Ins	Centromeric insertion
CEU	Caucasian european

	population
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
ChIP	Chromatine immunoprecipitation
ChIPseq	ChIP sequencing
cHLHLZ	Helix loop helix leucine zipper
chr	Chromosome
ChromHMM	Chromatin state discovery and characterization software or modeling approach
CLP	Cleft lip and palate
CLyP	Common lymphoid progenitor
cm	Centimeter
CMPs	Common Myeloid Progenitor
CO2	Carbon dioxide
Conc	Concentration
CpG	Cytosine phosphodiester bond guanine
CTCF	CCCTC-binding factor
CTD	C-Terminal domain
CTF/NF-1	CCAAT box-binding transcription factor/nuclear factor
d	Diencephalon
DamID	DNA adenine methyltransferase identification
DAPI	4',6'-diamidino-2-phenylindole
DCE	Downstream core element
Ddef1	Development and differentiation enhancing factor 1
DEL	Deletion
del(8a-17a)	Deletion of the (8a-17a)interval
dH2O	distilled water
DIG	Digoxigenin
DIG-AP	DIG antibody product
DKFZ	Deutsches Krebsforschungszentrum
DMEM	Dulbecco's Modified Eagle Medium

DMSO	Dimethyl sulfoxide
dmyc	drosphila myc
DNA	DeoxyriboNucleic Acid
DNA-FISH	DNA fluorescent in situ hybridization
Dnase	Deoxyribonuclease
dnBmpr1A	Dominant negative Bmpr1A
dNTP	Desoxyribonucleotid triphosphate
DPE	Downstream promoter element
Dup	Duplication
DV	Dorsal ventral
E-box	Enhancer box
E.Coli	Escherichia Coli
e	Embryonic day
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol tetraacetic acid
EMBL	European Molecular Biology Laboratory
eMFM	enhancer for Medial Facial Mesenchyme
ENCODE	Encyclopedia of DNA elements
eNE	enhancer for Nasal Pit Epithelia
EP300	E1A-associated cellular p300
eRNA	enhancer-associated RNA
ESC	Embryonic stem cell
f	Face
FACS	Fluorescence activated cell sorting
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements sequencing
FCS	Fetal calf serum
FDG	Fluorescein Di-β-D-Galactopyranoside
FDR	False discovery rate
FIJI	Fiji is just imageJ
FSC	Forward scatter
ftz	Fushi tarazu
g	Gram
G0	Resting phase
G1	Gap1 phase

Gal4-UAS	Gal4-upstream activation sequence
GC box	Guanine cytosine box
GFP	Green fluorescent protein
GMP	Granulocyte-macrophage progenitor
GO	Gene ontology
GRN	Gene regulatory network
GROMIT	Genome regulatory organization mapping by inserted transposon
GTFs	General transcription factor
GWAS	Genome wide association studies
H2A	Histone 2A
H2A.Z	H2A family, memberZ
H2O2	Hydrogen peroxide
H3	Histone 3
H4	Histone 4
HEK293T	Human embryonic kidney 293T
HeLa cell	Henrietta Lacks cell
HEPES	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid
het	Heterozygous
HMR	Hybrid male rescue
hom	Homozygous
HP	Heterochromatine protein
HPE	Holoprosencephaly
Hprt	Hypoxanthine phosphoribosyltransferase
HPV	Human papilloma virus
HRP	Horseradish peroxidase
HS	Hypersensitivity site
HSC	Hematopoietic stem cell
Hsp	Heat shock protein
HSy	Hematopoietic system
ICR	Imprinting control region
ID	Identity
IFN	Interferon
Ig	Immunoglobulin
Igf	Insulin-like growth factor
IL	Interleukine

ILR	Interleukine Receptor
Inr	Initiator
Ins	Insertion
INV	Inversion
IPTG	Isopropyl β-D-1-thiogalactopyranoside
K	Lysine
K3Fe3(CN)6	Potassium ferricyanide
K4Fe2(CN)6	potassium ferrocyanide
kb	Kilobase
KCl	Potassium chloride
KOH	Potassium hydroxide
L (primer)	Left primer
L	Liver
lacI	Lac repressor
LAD	Lamina associated domain
LCR	Locus control region
LD	Linkage disequilibrium
LFM	Lateral face mesenchyme
lincRNA	Long intergenic non-coding RNA
Lin	Lineage
LINE	Long interspersed nuclear element
lncRNA	Long non-coding RNA
lnp	Lateral nasal process
log	logarithm
Lrp6	Low density lipoprotein receptor-related protein 6
LS-K	LS minus K (Lin- Sca-1- c-Kit+)
LSK	Lin- Sca-1+ c-Kit+
LT-HSC	Long-term HSC
LTR	Long terminal repeat
M	Molar
MAFB	Musculoaponeurotic fibrosarcoma oncogene homolog B
MAR	Matrix attachment region
mb	Megabase
md	Mandibular
me	Monomethylation
me2	Dimethylation
me3	Trimethylation
MEF	Mouse embryonic fiblroblast
MEP	Megakaryocyte-

	erythrocyte progenitor
mESC	mouse ESC
MFM	Medial facial mesenchyme
MgCl2	Magnesium chloride
Micro-Ct	Micro computed tomography
min	Minute
ml	Milliliter
mM	Milimolar
mm	Millimiter
Mnp	Medial nasal process
mod	Modifier
MPP1	Short term HSC
MPP2/3/4	Multipotent progenitor
mRNA	messenger RNA
MTE	Motif ten element
mx	Maxillary arch
Myf	Myosine factor
NaAc	Sodium acetate
NaCl	Sodium chloride
NBT	Nitro blue tetrazolium chloride
NCC	Neural crest cell
Ndn	Necdin-encoding
NE	Nasal pit epithelial
ng	Nanogram
nm	Nanometer
no	Number
np	Nasal pit
NP40	Nonyl phenoxypolyethoxylethanol
NSCLP	Non-syndromic cleft lip and palate
NTMT	Alkaline phosphatase buffer
NTP	Nucleotide triphosphate
O/N	Overnight
OPT	Optical projection tomography
OPTI-MEM	Reduced-serum medium
p	p-value
p arm	Short chromosome arm
p27KIP1	Cyclin-dependent kinase inhibitor 1B
pA	Poly adenosine tail
Pabc1	Aminodeoxychorismate lyase
Pax	Paired box
PBS	Phosphate buffered saline
PBS-T	PBS-tween20
Pc-G	Polycomb-group
PCR	Polymerase chain reaction
PFA	Paraformaldehyde
pH	Potential hydrogen
PH3	Phospho histone3
PIC	Pre-initiation complex
PK	Proteinase K
PLM	Proximal limb mesoderm
PMSF	Phenylmethylsulfonyl fluoride
PRC	Polycomb repressor complex
PRE	Polycomb response element
PSQ	Penicillin Streptomycine glutamine
PVDF	Polyvinylidene difluoride
q arm	Long chromosome arm
qPCR	quantative PCR
qRT	quantitative real-time
R (primer)	Right primer
RD	Regulatory domain
rDNA	Ribosomal RNA
RIPA	Radio immunoprecipitation assay
rl	Rhombic lip
RNA	Ribonucleic acid
RNAPII	RNA polymerase II
Rnase	Ribonuclease
RNAseq	RNA sequencing
rpm	Round per minute
rpmk	reads per kilobase per million mapped reads
rtTA	reverse tetracycline-controlled transactivator
sec or s	Second
s	Somite
S.cerevisiae	Saccharomyces cerevisiae
S2 cell	Drosophila Schneider 2 cell
SB	Sleeping beauty transposon
SDS	Sodium dodecyl sulfate
SDS-PAGE	SDS - polyacrylamide gel electrophoresis

SHFM	Split hand foot malformation
Shh	Sonic hedgehog
SINE	Short interspersed nuclear element
siRNA	small interfering RNA
Slam	Signaling lymphocyte activation molecule
SNP	Single nucleotide polymorphism
SSC	Saline-sodium citrate
STARR-seq	Self-transcribing active regulatory region-sequencing
Su(Hw)	Suppressor of hairy-wing
SV40	Simian virus 40
t	Telencephalon
T cell	Lymphocyte T
t(8:14)	Translocation between chr8 and chr14
TAD	Topologically associated doamin
TAMERE	Targeted meiotic recombination
TBST	Tris-buffered saline tween20
TCDD	2,3,7,8-tetrachlorodibenzo-p-dioxin
Tel_Ins	Telomeric insertion
TF	Transcription factor
TFBS	Transcription factor binding site
TFF	Trefoil factor
TRACER	Transposase and recombinase-associated chromosomal engineering resource
TRIS	Tris(hydroxymethyl)aminomethane
tRNA	transferase RNA
TSS	Transciption starting site
UCSC	University of California, Santa Cruz
USA	United States of America
UV light	Ultraviolet
v2.0	version 2.0
wS	Widespread
wt	Wild-type
YAC	Yeast artificial chromosome
ZRS	Zone of polarizing activity regulatory sequence
µl	Microliter
µM	Micromolar

8.4 Supplementary Tables

Supplementary Table 1: Reference List for GWAS in 8q24

SNP Position	SNP no	PMID
chr8 12733265	rs11986011	23326517
chr8 12733265	rs11986011	23326517
chr8 12790200	rs4871750	23251661
chr8 12791159	rs2220321	23251661
chr8 12809329	rs1016343	21743057
chr8 12809329	rs1016343	18264097
chr8 12809515	rs13252298	21743057
chr8 12810393	rs1456315	23023329
chr8 12810393	rs1456315	20676098
chr8 12810434	rs13254738	22923026
chr8 12810687	rs6983561	22923026
chr8 12812491	rs16901979	19767754
chr8 12812491	rs16901979	17401366
chr8 12850449	rs17766217	22041458
chr8 12851757	rs4242382	18264096
chr8 12851855	rs4242384	21743057
chr8 12851855	rs4242384	19767753
chr8 12851855	rs4242384	18264097
chr8 12853213	rs10090154	22923026
chr8 12853935	rs7837688	20676098
chr8 12871806	rs9642880	20972438
chr8 12871806	rs9642880	20348956
chr8 12871806	rs9642880	18794855
chr8 12881502	rs4410871	21833088
chr8 12907216	rs2648875	17395743
chr8 12907583	rs2608053	21037568

chr8	12812519	rs10505483	22923026
chr8	12819298	rs2456449	20062064
chr8	12832034	rs16902094	19767754
chr8	12832318	rs445114	21743057
chr8	12832318	rs445114	19767754
chr8	12835561	rs13281615	17529967
chr8	12838785	rs1562430	21263130
chr8	12838785	rs1562430	20453838
chr8	12840744	rs10505477	17618283
chr8	12841330	rs6983267	23266556
chr8	12841330	rs6983267	21743057
chr8	12841330	rs6983267	21242260
chr8	12841330	rs6983267	18372905
chr8	12841330	rs6983267	18264097
chr8	12841330	rs6983267	18264096
chr8	12841330	rs6983267	17618284
chr8	12841330	rs6983267	17401363
chr8	12842479	rs7014346	18372901
chr8	12848503	rs1447295	19767754
chr8	12848503	rs1447295	17401366
chr8	12848503	rs1447295	17401363
chr8	12919227	rs2019960	21833088
chr8	12919227	rs2019960	21037568
chr8	12922007	rs11995854	23251661
chr8	12924641	rs10492294	20694011
chr8	12926458	rs9792269	20190752
chr8	12931601	rs975730	21383967
chr8	12942751	rs7815944	23042114
chr8	12954394	rs10088218	23535730
chr8	12954394	rs10088218	23535730
chr8	12954394	rs10088218	20852632
chr8	12956718	rs6651252	23128233
chr8	12956718	rs6651252	21102463
chr8	12994615	rs987525	22863734
chr8	12994615	rs987525	22863734
chr8	12994615	rs987525	22863734
chr8	12994615	rs987525	20436469
chr8	12994615	rs987525	19656524
chr8	12994615	rs987525	19270707
chr8	13049175	rs891835	19578367
chr8	13057210	rs10956483	21738478
chr8	13062410	rs1991866	23128233
chr8	13067663	rs9918807	23393555
chr8	13068545	rs4295627	21531791
chr8	13068545	rs4295627	19578367
chr8	13072566	rs6470764	20881960
chr8	13082003	rs2128382	23266556
chr8	13098047	rs10092658	18464913
chr8	13109241	rs6984045	19525955

Supplementary Table2:

GO term enrichment analysis of the highly expressed (rpmk>100) differentially regulated gene upon del(8a-17a): Yellow highlight indicates translation related terms, red highlight indicate hematopoietic system related terms, and the purple highlight indicates metabolism related terms.

GO term	**Description**	P-value
GO:0006412	translation	1.03E-06
GO:0034109	homotypic cell-cell adhesion	5.37E-05
GO:0006919	activation of cysteine-type endopeptidase activity involved in apoptotic process	1.64E-04
GO:0097202	activation of cysteine-type endopeptidase activity	1.64E-04
GO:0031331	positive regulation of cellular catabolic process	1.86E-04
GO:0051246	regulation of protein metabolic process	2.40E-04
GO:0051247	positive regulation of protein metabolic process	2.93E-04
GO:0071345	cellular response to cytokine stimulus	3.29E-04
GO:0070527	platelet aggregation	3.37E-04
GO:0002761	regulation of myeloid leukocyte differentiation	3.66E-04
GO:0019221	cytokine-mediated signaling pathway	4.14E-04
GO:0006417	regulation of translation	5.26E-04
GO:0009896	positive regulation of catabolic process	6.35E-04
GO:0040012	regulation of locomotion	6.65E-04
GO:0043407	negative regulation of MAP kinase activity	7.53E-04
GO:0044267	cellular protein metabolic process	7.70E-04
GO:0034097	response to cytokine stimulus	8.01E-04
GO:0090066	regulation of anatomical structure size	9.96E-04

8.5 Acknowledgements

The PhD has been the hardest thing I had to get through in my life. Thanks to many people, it was fun, meaningful and fulfilling.

First of all, I would like to thank to François Spitz for hosting me in the lab, introducing the fascinating puzzle about the genome organization, providing me an invaluable supervision during this period and creating a pleasant lab atmosphere. More than these, I am truly thankful to him for teaching me the fundamentals of scientific research. In addition, I would like to thank to Jochen Wittbrodt, Anne Ephrussi and Eileen Furlong for their valuable feedback and input to the project

Spitz lab members and alumni have all been a very helpful, supportive and most importantly friendly attitude during my PhD period. It was a pleasure to know and to work together especially with Tugce Aktas and also Mirna Marinic, Katja Langenfeld, Sandra Ruf, and Wibke Schwarzer, who help me to tackle with countless obstacles.

I would like to thank to Ahmet Ucar, Ihsan Gursel and Isik Yulug for their very valuable scientific guidance, to Andreas Trumpp and Lisa von Paleske for the fruitful collaboration in hematopoiesis project, to Taro Tsujimura, Massimo Petretich, Leslie Pan and especially to Orsolya Symmons for their constructive scientific criticisms and collaborative attitude.

I am sure I would not enjoy my stay in Heidelberg more, if Martin Schneider, Nathanael Harfst, Anna-Lisa Harfst did not make me a part of the legendary WG in Lessingstr.6. Innumerable thanks to Martin S, Tani H, Anna-Lisa H, Robert W, Dr. Anna, Justin S, Michaela S, Jannika J, Philip, Oli P, Anne K, and Lisa B from Lessingstr.6 for cheering me up even in the most exhausting and depressing day.

Predoc 2009 was a great group. Among them, it was a particular pleasure to meet and spend time with Christopher Amourda, Claudio Alfieri, Maria Secrier, Philip Brennecke, Volker Lauschke, Kasia Kasia, Sebastian Glatt, Anna Anna, Erik Mell, Tilman Plass, Ana-Rita, Comazzetto Stefano, Sander Timmer, Daniel Mende, and Mathieu Foux de Botte, who always made me feel that I am in the right place for the PhD. Besides, I would like to acknowledge Federico Maria Rossi, Carmen Aguilar, Martin Schorb, Tze Hang Tan, Aino Jarvelin, Alexandra Bebel, Nurlan, and Martein, Andreea Gruia, Adela Valceanu, for great time we had together.

The support and the company of Murat Iskar and Sevi Durdu deserve much more than an acknowledgement. I am very lucky to have them as a part of my Turkish family in Heidelberg. In addition, I would like to thank Ege Ilicak for his great company in Heidelberg.

There are three very special people that I would like to mention here but I will never be able to thank them enough for what they have done for me. I met Anne-Laure Duchemin during the first year of my PhD at EMBL. Dear Cancan Cherie: Since the moment that I set across you in the train to Freiburg, your presence is enough for me to be happy, to make me feel alive, and to enjoy every single moment of the life. Without your support, I would not be a healthy, sane person at the end of the PhD. Seni çok seviyorum. Finally, I would like to mention my dear parents, Leyla and Zafer Uslu: Sevgili Annem ve Babam, Size nasıl teşekkür edeceğimi, nereden baslayacağımı düşünürken, dört sene sonra gözümden yaşlar dökülüyor. Acaba diyorum bana gösterdiğiniz sevginin, Çiftlik İlkokulundan EMBL'a uzanan bu yolda verdiğiniz desteğin, güven duygusunu anlatmanin bir yolu var mı, bulamıyorum. İkinizi de çok seviyorum, çok özlüyorum.

8.6 ANNEX

1. Promoters.

1.1 Nucleosome Occupancy and Marks of Promoters:

PIC forms on the linker DNA. Therefore, the active promoter sequences are excluded from the nucleosomes. Nevertheless the histones of neighboring nucleosomes are informative about the promoter activity(Li Z *et al* 2011). For example, presence of H3K4me3 rich nucleosomes together with H3K4me1, H3 acetylation, H4K20me3 signal implies active promoters in the vicinity (Barski A *et al* 2007). In addition to histone signature, Serine 5 Phosphorylated polymerase binding correlates with promoter activity (Bernstein BE *et al* 2006)

1.2 Mutations on Promoter Sequences:

Mutations in the promoter sequence or in the activator proteins are shown to lead to human diseases. A common C to T mutation in the CACCC motif at -87bp position of ß-globin gene that causes a moderate form of ß-thalassemia is experimentally demonstrated to decrease the promote activity to half in HeLa cell line (Kulozik AE *et al* 1991). Similarly, when C/EBP binding is disrupted by A to G mutation at +13 position of factor IX gene it causes Hemophilia B (Crossley M and Brownlee GG, 1990).

2.Enhancers:

2.1 Historical Perspective and Definition of Enhancers:

Early studies on the sea urchin H2A promoter, the Simian Virus 40 (SV40), and the immunoglobin heavy chain gene promoter identified regulatory sequences up to 3300bp outside of promoter region that can enhance both endogenous and heterologous cis-linked genes in an orientation independent ad tissue-specific manner only in the presence of initiator sequences (or core promoter elements) (Grosschedl R and Birnstiel ML, 1980; Banerji J *et al* 1981, Banerji J *et al* 1983).

2.2 Enhanceosome Structure and Modularity.

Enhanceosomes modify the conformation of local chromatin around promoter to recruit PIC (reviewed in (Panne D, 2008)). Enhanceosome structure is well characterized in interferon-ß (IFN-ß) locus, which integrates extensive environmental signals, such as virus infection for activation. On contrary to the weak binding of transcription factors to enhancers or proximal promoters, IFN-ß enhanceosome is a very stable complex. Daniel Panne argues that the stability of the complex compromises the modularity of enhancers to make them more operative (Panne D, 2008). More enhanceosome structures have been revealed in mammalians, particularly in immune system related genes like TNF-α and TCR-α (reviewed in (Merika M and Thanos D, 2001).

2.2 The Mechanism of Enhancer Action:

The nature of the transcriptional information encoded in enhancers is not very well understood at single cell level. The tools like qPCR and luciferase, which are used to measure the reporter gene expression in the presence and the absence of enhancers, give only the average expression levels in a population of cells. Therefore, until the availability of single cell resolution tools, reporter gene expression measurements did not distinguish whether an enhancer increases the average reporter gene expression via increasing the rate of transcription ("rheostat model") or via increasing the number of the cells that express the reporter gene ("binary model") (reviewed in (Arnosti DN and Kulkarni MM, 2005; Fiering S *et al* 2000)). Several lines of experiments in cell culture via SV40 enhancer or endogenous enhancers suggested that enhancers increase the number of cell that express the reporter gene rather than the level of expression in each cells and it supports the binary model This model also fits to the working principles of enhanceosome dependent enhancers that switch on transcription solely in the presence of whole complex (Sutherland HG *et al,* 1997; Weintraub H, 1988, Walters MC *et al, 1995*). On the other hand, altering activator or repressor proteins bound to enhancer sequence results in gradual changes in the reporter gene expression, which fits to "rheostat model" (Rossi FM *et al* 2000). In addition to this, synthetic enhancers of developmental genes in drosophila lead to intermediate level of expression (Kulkarni MM and Arnosti DN, 2003). Functional dissection of ß-globin enhancers in erythroleukemia cells demonstrated that each single tested enhancer increase both the rate of transcription and the number of cells with active transcription of the reporter gene (Bouhassira EE *et al* 1997). The difference of in the nature of information encoded in the enhancers may be due to the evolutionary history of the target gene. For example, a non-modular and rapidly assembling enhanceosome may effectively trigger immune reaction immediately after viral infection. Due to non-modular nature of immune response, all the TF binding sites in IFN-ß enhancer have remained unchanged for the last 100 million years (Borok MJ *et al* 2010). On the other hand, organization and regulation of developmental genes undergo substantial changes throughout the evolution. For example, stripe 2 enhancer of eve gene in different Drosophila species demonstrated divergent protein binding site profiles despite keeping identical expression pattern of the gene (Ludwig MZ *et al* 1998). This modularity of developmental enhancers, which allows evolutionary tinkering correlates with the gradual changes in gene expression.

2.3 eRNA:

Michael Greenberg’s lab found out RNAPII also bind to H3K4me1 and H3K27ac marked enhancers like the enhancer of mouse *Arc* gene and start bi-directionally transcribing non-coding RNAs called eRNA. Although it is not clear whether these RNAs are transacting, the correlation between H3K4me1 signal on enhancer and eRNA expression levels only when the target gene of the enhancer is transcribed. Therefore, apart from the histone marks, eRNA expression potentially indicates enhancers of actively transcribed genes (Kim TK *et al* 2010).

3.Silencers

3.1 Mode of Action of Silences:

There are several mechanisms described for silencer activity:

In PDGF-A chain receptor locus, a silencer which is more than 1000bp far from the promoter of the gene, is proposed to mediate repression via inducing changes DNA topology (Liu B *et al* 1996). According to a different mechanism, silencer binding changes the nuclear localization as in human TSHß gene example. Moreover, studies in IL-4 promoter showed that repressor BCL-6 binds to IL-4 promoter in competition with STAT6 binding in a neighboring locus. In this situation, repressor works by inhibiting activator binding (Harris MB *et al* 2005). Another unusual example is in TFF1 promoter, where Sp1 activator and Sp3 repressor proteins bind to the same sequence. Therefore, a silencer sequence may act as a promoter proximal element depending on the protein associated with the sequence (He S *et al* 2005). Furthermore, a very fundamental silencing mechanism pointed out that the Sir2 and Dsp1 repressors inhibit PIC recruitment to the core promoter in yeast and in drosophila, respectively (Chen L and Widom J, 2005; Kirov NC *et al* 1996). There are cases where silencers work orientation independently such as in human growth hormone locus, and cases where the orientation makes a difference in the silencing activity as in c-fos promoter (Trujillo MA *et al,* 2006; Natesan S and Gilman MZ, 1993). There are cases where silencers are position independent such as human IFN-ß locus, and cases where they are position dependent as in human AFP gene locus (Nourbakhsh M *et al* 1993; Nakabayashi H *et al* 1991). Moreover, recent ChIP studies globally associated a number of histone marks with repressive activity in mammalian (Ku M *et al* 2008). The molecular characterization of Pc-G and its interaction partners improved the understanding of molecular mechanisms of transcriptional silencing, which provided a very valuable insight into understanding gene regulation in a developmental context like *Hox* gene cluster. To sum up, silencers impede active transcription in many different ways and the assays to test silencer activity are not able to reflect very different forms of repression.

4. Insulators

4.1 The mode of Action of Insulators:

The molecular mechanisms of insulator function are still elusive. Observations coming from insulator assays implied several different mechanisms underlying the disruption or establishment of enhancer-promoter interaction. Recently, genome-wide analysis of CTCF, cohesin binding implicate more global role for insulator sequences (reviewed in (Bushey AM *et al*, 2008)).

Molecular characterization of the mouse H19 locus insulator, which mediates imprinting of this locus; revealed a CTCF dependent insulator sequence (ICR) in the close proximity of both the enhancer and the promoter (Yoon YS *et al* 2007). This observation fits to the decoy model proposed by Pamela Geyer, which suggests a competition between the promoter and the insulator for enhancer (Geyer PK and Clark I, 2002). Further observations challenge but not necessarily falsify this model. For example, 5'HS4 insulator functions is observed only when the insulator is in between the enhancer and the promoter. The insulator on the opposite side of the enhancer does not interfere with the enhancer activity regardless of its distance (Figure 1).

Intriguingly, in transiently transfected cells if the plasmid is circular, insulator works in a position independent way. However, when the plasmid is linearized, insulator works only when it is in between the enhancer and the promoter sequence (Recillas-Targa F *et al* 1999). This position dependency of the insulator suggests the presence of additional mechanisms for insulator activity. For example, insulators are proposed to stall the enhancer when it scans along the DNA towards the promoter by obstructing the transcription activator signal, such as histone acetylation on enhancers (Zhao H and Dean A, 2004). This obstruction model also fits with working principle of *gypsy* insulator when it is single copy. However, abolishing the activity of *gypsy* insulator with another gypsy insulator does not fit with the model of insulators being physical barriers.

Genome-wide studies proposed that compartmentalization of the genome the main mechanism that underlies insulator activity. For instance, it has been proposed that drosophila genome is compartmentalized into approximately 100kb long segments, which show similar expression/repression pattern (Spellman PT and Rubin GM, 2002). Insulators are located all along the genome in a heterogeneous way. In silico analysis of Su(Hw) binding sites suggested that the insulator sequences are more likely to be in the boundaries of these transcriptionally distinct compartment (Ramos E *et al* 2006). Compartmentalization can be are explained with a loop model, in which the insulators work as anchor points of the loops stems. Biochemical analysis and microscopic visualization of the interphase chromatin, together with genetic tools, support that insulator mediate loop formation and shaping of global genome architecture. For example, two studies suggest that there chromatin loops form due to the interaction between the CTCF bound insulators and matrix attachment region (MAR3) as well as interaction between 2 CTCF bound sites (Kurukuti S *et al* 2006; Handoku L *et al* 2011). Despite the presence of a number of correlation-based studies on the loop formation and gene expression, it is not clear why only a very small subset of the CTCF bound sites contribute to the compartmentalization (Dixon JR *et al* 2012). Moreover, enhancer promoter loops do not always induce gene activation as it has been shown in *Shh* locus (Amano T *et al* 2009). Therefore, the mechanistic link between loops and regulation of gene expression is completely missing.

4.2 Insulator Deregulation:

Mutations in the proteins associated with insulator activity (Gause M *et al* 2008) are implicated in the etiology of the Cornelia de Lange syndrome (CdLS). Mutations in NIPBL, SMC1 and SMC3 genes are shown to be present in the patients with a changing tissue specific severity within the spectrum of Cornelia de Lange Syndrome. NIPBL gene is a loading factor for cohesin complex, and SMC1 and SMC3 are structural units of cohesin complex (Liu J and Krantz ID, 2009). Drosophila homologue of NIPBL facilitates enhancer-promoter interaction as well as regulating insulator functions (Rollins RA *et al* 1999). Recently, more evidence in mammalian cell culture and developmental systems indicate a possible role for cohesin in long-range interactions (Seitan VD *et al,* 2011). Therefore, CdLS can appear as a reflection of global deregulation of long-range interactions.

5. Locus Control Region Deregulation:

Historically one of the earliest cases that implicate long-range regulation in gene regulation revealed deregulation of LCRs in the thalassemia patients. Lack or imbalance of hemoglobin genes (α-globin and ß-globin) causes thalassemia. Mostly α-globin but also ß-globin gene deletions are identified in thalassemia patients (reviewed in (Thein SL, 2005)). Studies on thalassemia patients showed that a deletion in the upstream of ß-globin gene keeps the gene intact but recapitulates the ß-globin gene deletions. In the discussion of these studies, along with several other mechanisms, it was postulated that ß-globin gene was regulated by distant cis-regulatory elements (Kioussis D *et al* 1983). Further studies demonstrated that there is an LCR in the deleted regions of these thalassemia patients.

6. Methodology of ChIP

Starting from the very first operon model, protein-binding to core promoter or proximal promoter sequences is known to regulate transcription (Gierer A, 1966) In order to monitor protein-DNA interaction several methods including dimethyl sulfate based chemical modifications, DNase, UV light, and formaldehyde-mediated footprinting have been extensively used (Siebenlist U *et al* 1980; Solomon MJ and Varshavsky A, 1985). Recently, new generation sequencing contributed to the understanding of transcriptional regulation in global scale via revealing chromatin accessibility. Global-scale footprinting studies using FAIRE-seq, DNase I HS-seq give an idea of chromatin availability for protein binding as shown in the loci regulated by Androgen Receptor (AR) (Tewari AK *et al* 2012). Although these footprinting methods informed the protein occupancy in a locus, they were not informative about the identity of the proteins. Initially in 1984, Polymerase II antibody was used to detect polymerase-DNA interaction on IPTG induced lac operon in E.coli via UV-light crosslinking (Gilmour DS and Lis JT, 1984). Then, in 1988 Mark Solomon and colleagues replaced UV-crosslinking with formaldehyde-based crosslinking to estimate histone occupancy in drosophila hsp70 gene (Solomon M *et al* 1988). In 1996, ChIP protocol was coupled to PCR for the first time and gained higher throughput (Hecht A *et al* 1996). The initial low throughput ChIP analysis supported the influence of transcription factor binding and certain histone modifications to gene expression by showing occupancy of these proteins to the regulatory sites, particularly to proximal promoter element (For example: Orlando V and Paro R, 1993). With the relatively recent microarrays and state-of-art deep sequencing methods (ChIP-chip, ChIPseq), revealed protein binding-profile of the genome globally a tissue-specific way (Ho JW *et al* 2011). Around 80.4% of the human genome exhibited a protein-binding related biochemical activity in at least one cell type (Kavanagh DH *et al* 2013). Therefore, analysis of the protein-DNA interaction moved from a gene-centric view to more global understanding of biochemical marks.

7. DamID Methodology:

The correlations for regulatory activity and transcription factor binding obtained by ChIP experiments are supported by an independent method called DamID. This method is developed by Bas van Steensel and it is based on linking bacterial DNA adenine methylase (Dam) enzyme to a transcription factor. Since adenine methylase does not exist in eukaryotic systems, the adenine methylation takes place only in the presence of transcription factor fused to Dam. Adenine methylation is enriched in the proximity of the transcription factor binding site (TFBS). Methylated adenines are distinguished from non-methylated adenines by a restriction enzyme. Successive adapter ligation and deep sequencing identifies the transcription factor binding sites (Vogel Maartje MJ *et al* 2007). Unlike ChIP, DamID does not require very specific antibodies and extensive amount of cells. However, it is not possible to track post-translationally modified proteins with DamID. In addition, while ChIP gives the average occupancy of the analyzed cell population at the time of fixation, DamID can give the history of transcription binding from the time of induction till the experiment. LADs are discovered by DamID: the lamin protein was linked to Dam and it marked the nuclear regions that are located in the proximity of nuclear membrane.

8.Methods to Reveal the Physical Structure of the Genome:

Microscopy based approach allowed single cell analysis. However, biochemical methods were more suitable to perform experiments in large cell populations.

8.1 Methodology for Microscopy Based Studies:

Microscopy methods date back to the beginning of 20th century. The first observations, which suggest that the chromatin is not homogeneously distributed in the nucleus, are made by Santiago Ramon y Cajal and by Emil Heitz (reviewed in de Wit E and de Laat W, 2012). Microscopy observations using specific probes mapped chromosomes to partially intermingling territories in diploid human fibroblast nuclei (Bolzer A *et al* 2005; Branco MR and Pombo A, 2006). According to these chromosome territories, the gene rich chromosomes align in the interior parts of the nuclei, whereas, the gene poor chromosomes were at the periphery. Moreover, introduction of ß-globin HS1 site to heterochomatic satellite sequence or transcription factor binding to the heterologous promoters changed the position of the locus in the nuclei (Lundgren M *et al* 2000). Recently, microscopy studies showed that tethering active loci to nuclear lamina decreased the level of gene expression in majority of the studies in the presence of H3K9 specific methyltransferase enzymes (Towbin BD *et al* 2012). This correlation has been stepped forward by the observation that *Hoxb* gene expression increases upon displacement of the genes from their native chromosomal territory (Chambeyron S and Bickmore WA, 2004). Consequently, microscopy methods give a lot of information on the nuclear organization of the genome and also respective location of one region in the genome with respect to another one or a particular nuclear landmark like lamina. However, the resolution obtained by light microscopy is still limited to 100-200nm. This limitation will get better with the advancements in high-resolution microscopes (Weiland Y *et al* 2011). In addition, future applications of super resolution microscopy and correlative electron-fluorescence microscopy to nuclear organization can possibly bring unprecedented resolution and insight to the field (Kukulski W *et al* 2011) Nevertheless, the number of items co-visualized is currently limited to the light emission features of probes in the visual spectrum.

8.2 Methodology for Chromosome Conformation Capture Based Studies:

Biochemical approaches to understand the genome organization kicked off a decade ago by introduction of "Chromosome Conformation Capture (3C)" method. 3C aims to get a linear footprint of 3D structure of the chromatin via proximity ligation. In 3C and 3C-derived method, the cells or nuclei are cross-linked by PFA. Cross-linking step is followed by restriction enzyme digestion by 4bp or 6bp cutters. After this step, these cross-linked and cut sites are ligated. Therefore, the sequences, which are close to each other in 3D, can ligate in one fragment. The latter work is to quantify the ligation frequency to reveal the average 3D structure of a locus or the whole genome in a population of cells (reviewed in de Wit E and de Laat W, 2012).

The level of interaction in between two sites is investigated in a 3C experiment. Therefore, primers are designed specific to the sites of interest. Quantitative amplification of the ligation products with selected primers give a relative 3D proximity of these sites. In addition to the examples given in the introduction, loops between the 5' and 3' end of ribosomal gene promoters were identified. These loops correlated with facilitated loading of the RNAPII to the promoter of the genes, thereby, elevated gene expression (Nemeth A *et al* 2008). On the contrary to rDNA expression, promoter-terminator loops in BRCA1 tumor suppressor gene is shown to have a negative effect in the expression of this gene (Tan-Wong SM *et al* 2008). To sum up, 3C method is used to show physical proximity or loop structures particularly in well-characterized loci. However, the presence of loops does not necessarily indicate whether a locus is active or not.

5C is a higher throughput version of 3C, where multiple primers are designed in a locus and the ligation product is amplified in a multiplex PCR. 5C gives an overall interaction map of a locus rather than specific enhancer-promoter interaction (Dostie J *et al* 2006). The interaction map obtained in α-globin locus by 5C has been converted into a 3D distance map based on frequency of collisions (ligations) and this distance map is shown to be consistent with the distance measured by DNA-FISH in two distinct cell types. Chromatin modeling based on 5C data postulated a globular architecture in sub-megabase level, where the active genes locates at the core of the globule (Bau D *et al* 2011).

3C method is also used for relatively less uncharacterized regions like CFTR gene by investigating interaction of the whole DNaseI hypersensitive sites with the TSS of CFTR (Gheldof N *et al* 2010). However, the employment of next-generation sequencing technology to 3C made it possible to reveal all the DNA fragments ligated to the fragment of interest (or "viewpoint") by using viewpoint specific primers. This modified method is called 4C. Due to extra PCR steps and possible biases introduced at next generation sequencing, 4C is less quantitative than 3C. Noteworthy, single fragments interacting with the viewpoint are not reproducible in 4C experiments but the interaction of the viewpoint with a window of several fragments is reproducible. Therefore, smoothened data is favored over raw sequencing counts in 4C data. Further statistical analysis narrows down the interaction sites and gives an unbiased architectural insight of the viewpoint (reviewed in de Wit E and de Laat W, 2012). Some of the physical interactions discovered by 4C appeared to be important when complemented with ChIPseq data and human genetic disorders. However, for majority of the cases, it is not clear what physical interactions correspond in terms of gene expression. For example, repressed X chromosome regains active X chromosome conformation upon Xist deletion. However, the gene activity of conformationally restored X chromosome is not regained upon Xist depletion (Splinter E *et al* 2011). This suggest that gene expression and chromosome conformation do not solely dependent on each other. Nevertheless, currently 4C data irreplaceably sheds light on promoter-enhancer interactions from both "viewpoints". Moreover, the anisotropic distribution of physical interactions of a site between its centromeric side and telomeric side is suggested to be an indicator of differential chromatin compartmentalization and correlate with transcriptional activity (Montavon T *et al* 2011).

In order to highlight the involvement of proteins in establishing interactions, proximity ligation is combined with ChIP experiments, called ChIA-PET. The loop structures are postulated by 3C and 4C experiments such as in ß-globin locus, depend on the interaction between CTCF sites. However, not all the CTCF bound sites are looping with each other. In order to find out which sites are brought together via the same CTCF molecule, ChiA-PET is used in mESC. Two interesting results came out of this experiment. First one is only very few CTCF sites were found to be associated with loops. Considering that some of these interacting sites are enriched by p300, this supports the idea that only certain subset of CTCF sites influences enhancer-promoter specificity. Second, trans interactions, which were rarely observed in 4C assay, taking place via CTCF molecule appeared to be higher than 20% of the cis interactions (Handoko L *et al* 2011). This suggests that intermingling surfaces of chromosomal territories may have a substantial impact on gene expression in trans.

Sequencing the whole proximity ligation library revealed the conformation of the genome by a new method called HiC. In order to pull down and sequence ligation products, biotin-labeled nucleotides or adapters are used to tag the ligated fragments. Deep sequencing of these all possible ligation products revealed the topology of the genome both in mammals and in yeast (Lieberman-Aiden E *et al* 2009, Duan Z *et al* 2010). Due to limited sequencing depth in mammals, the resolution of the interactions were restricted to one megabase scale in Lieberman-Aiden 's study and due to small size of yeast genome this resolution is at kilobase range in Duan Z *et al* 2010. HiC data in mammals showed that the active regions cluster together, and inactive regions cluster together in megabase scale. Extensive modeling based on proximity between sites and knot-free structure of the genome, Lieberman-Aiden *et al* postulated that genome is organized in fractal globule structures.

Recently, Bing Ren's lab obtained extreme read depth in HiC experiment (170 times more than what Lieberman-Aiden *et al* obtained) and increased the resolution of the interactions to 40-60kb levels. This high-resolution interaction map in the mouse and the human cells revealed a sub-megabase chromatin organization called topologically associated domains (TADs). 5C experiments done on X-chromosome by Edith Heard's lab also revealed TADs with much higher coverage in the region (Dixon J *et al* 2012). These TAD structures obtained by biochemical assays are independently confirmed by DNA fish experiments (Nora EP *et al* 2012). The TAD structures indicate that the genome is compartmentalized in 200kb to 2 megabase long stretches. Therefore, most of the interactions take place within these topologically associated domains. An interesting feature of the TADs appear at the boundaries of two TADs, in which two adjacent or proximal sites show a completely different interaction profile: The centromeric site of the boundary only interacts with the centromeric regions, and the telomeric site of the boundary only interacts with the telomeric regions. Hidden Markov Models are applied to the raw reads to quantify the anisotropy of interactions. The peaks of anisotropy appeared at the TAD boundaries. It is very intriguing to understand what is the molecular basis of anisotropy at the boundary region. This is a key question to understand how the TADs emerge. Histone mark and transcription factor binding sites in different tissues, in different organisms are used to associate the boundary region with some biochemical activity. First of all, the TADs boundaries appeared to be extremely conserved in the synthenic regions of mouse and human. Moreover, despite the changes in the interaction frequency within the TADs, the boundaries are almost identical between stem cells, fibroblasts and even cortex cells. TAD boundaries lack shared motifs but they are slightly enriched in CTCF sites as well as house keeping genes and depleted for repressive marks like H3K9me3 (Dixon J *et al* 2012). Upon deletion of a TAD in X chromosome, a brand new TAD boundary formed in between two TADs and the other TADs were not affected from this deletion (Nora EP *et al* 2012). There are two different hypotheses for the formation of boundaries. One of them propose that the boundaries are consequence of small interactions in the TADs and the other one propose that the boundary elements are yet undiscovered biochemical entities that can separate two distinct domains from each other.

Last but not least, Jörg Langowski puts forward the fact that all of the biochemical methods that have been mentioned here average the interactions in a given population. Considering the dynamics of DNA as a huge polymer, average interaction frequency may not represent the actual nature of interactions (Langowski J, 2010).

8.7 Publications

1.Chen, C.-K., Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Smedley, D., & Spitz, F. (2013). TRACER: a resource to study the regulatory architecture of the mouse genome. *BMC genomics*, *14*, 215.

2. Ruf, S., Symmons, O., Uslu, V. V., Dolle, D., Hot, C., Ettwiller, L., & Spitz, F. (2011). Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor. *Nature genetics*, *43*(4), 379–86.

3. Symmons, O, Uslu, V.V….. Spitz, F (in review) Functional characteristics of mammalian regulatory landscapes, *Genome Research*

BMC Genomics

DATABASE Open Access

TRACER: a resource to study the regulatory architecture of the mouse genome

Chao-Kung Chen[1], Orsolya Symmons[2], Veli Vural Uslu[2], Taro Tsujimura[2], Sandra Ruf[2], Damian Smedley[1,3*] and François Spitz[2*]

Abstract

Background: Mammalian genes are regulated through the action of multiple regulatory elements, often distributed across large regions. The mechanisms that control the integration of these diverse inputs into specific gene expression patterns are still poorly understood. New approaches enabling the dissection of these mechanisms *in vivo* are needed.

Results: Here, we describe TRACER (http://tracerdatabase.embl.de), a resource that centralizes information from a large on-going functional exploration of the mouse genome with different transposon-associated regulatory sensors. Hundreds of insertions have been mapped to specific genomic positions, and their corresponding regulatory potential has been documented by analysis of the expression of the reporter sensor gene in mouse embryos. The data can be easily accessed and provides information on the regulatory activities present in a large number of genomic regions, notably in gene-poor intervals that have been associated with human diseases.

Conclusions: TRACER data enables comparisons with the expression pattern of neighbouring genes, activity of surrounding regulatory elements or with other genomic features, revealing the underlying regulatory architecture of these loci. TRACER mouse lines can also be requested for *in vivo* transposition and chromosomal engineering, to analyse further regions of interest.

Keywords: Gene regulation and expression, Genome organisation, Regulatory landscapes, Chromosomal engineering, Mouse models of human structural variation

Background

Genes occupy only a small fraction of mammalian genomes. Accordingly, intergenic regions can extend up to several megabases, and the functional importance of these regions is being growingly recognized [1] (Figure 1). Notably, these regions comprise important elements that control gene expression [3]. Enhancer elements are frequently found hundreds of kilobases away from the promoter of the gene they control, sometimes even separated from it by unrelated genes [4-11]. These remote enhancers can be essential for gene expression, as shown by human disorders resulting from their mutation or disruption by chromosomal rearrangements [12-16]. The importance of these intergenic regions in human phenotypic diversity and disease susceptibility is further emphasized by the significant proportion of risk alleles that have been identified in gene-desert intervals [3,17-20]. Thus, there is a pressing need to better characterize the nature of the regulatory activities embedded in such regions and to obtain animal models to help dissect *in vivo* how variations in these regions contribute to human phenotypes.

Recent progress in whole genome chromatin profiling has led to the identification of chromatin features that are strongly correlated with gene regulatory elements [21-26], opening ways to obtain a comprehensive catalogue of these elements, and a better annotation of the regulatory genome [27]. Databases that document the *in vivo* activities of experimentally validated regulatory elements – mostly enhancers – further complement these approaches [28]. Such datasets on regulatory activity can be compared to gene expression data in

* Correspondence: damian@ebi.ac.uk; spitz@embl.de
[1]European Bioinformatics Institute - European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
[2]Developmental Biology Unit - European Molecular Biology Laboratory, Heidelberg, Germany
Full list of author information is available at the end of the article

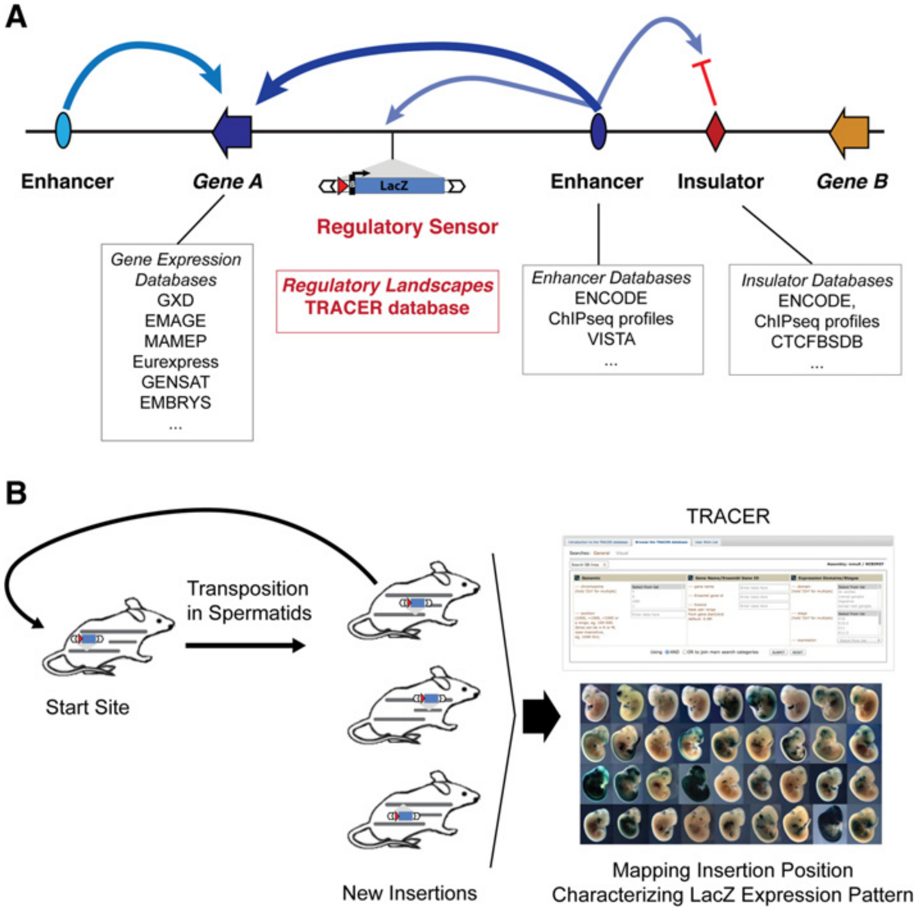


Figure 1 Genome organisation and TRACER. (**A**) Schematic representation of a genomic locus, with the different elements that control gene expression, and the specialized databases dedicated to their description (see Websites). The TRACER database displays information from a regulatory sensor that detects the regulatory influences active at its insertion site, outlining the characteristics and extent of genomic regulatory landscapes. (**B**) The GROMIT strategy. A transposon that contains the regulatory sensor can be remobilised *in vivo* from its start site by a transposase transgene (not shown) active only in the male germline [2]. The genomic position of the new insertions and the associated expression pattern defined by LacZ staining in mouse embryos is displayed in the TRACER database.

developing mouse embryos [29-35]. However one cannot reduce gene expression to a catalogue of the many potential regulatory elements present in the genome (from few hundred thousands to millions [22]). It is equally important to understand the interplay between the different elements present at a locus and how their different inputs are integrated and conveyed to target gene(s). Yet, compared to enhancers, other *cis*-regulatory elements such as silencers are much more elusive, despite their essential role in gene expression. Similarly important are the mechanisms that define the range and specificity of enhancer-promoter interactions. Indeed, changes in the relative position of genes and regulatory elements by chromosomal rearrangements and structural variations can alter gene expression with dramatic consequences [36-40]. Understanding these situations and the associated mechanisms requires approaches that complement the available catalogues of elements and provide a functional integrated view of the genome regulatory architecture.

For this purpose, we have developed an approach based on the distribution of a regulatory sensor gene throughout the mouse genome [2] (Figure 1B). The regulatory sensor consists of a LacZ reporter gene, which is driven by a minimal promoter that has no specific activity on its own but responds faithfully to endogenous enhancers. This regulatory sensor therefore uncovers the *regulatory potential* associated with a given genomic position, which results from the collective action of the different regulatory elements that act on this position. It thus reveals, in an operational manner, the gene regulatory activities within poorly characterized regions, or where annotation for activity is indirect (eg. chromatin profiling) or out of the proper genomic context (eg. transgenic assays). Importantly, the minimal promoter used does not display any obvious tissue- or enhancer-type bias, and the observed expression patterns often overlap with the ones of neighbouring genes [2]. The basic principle of the strategy is analogous to an enhancer-trap [41]; however, the sensor used in our approach has minimal impact on endogenous gene expression [2] and therefore reveals regulatory activities without titrating them away from their natural target genes.

This regulatory sensor is carried in a *Sleeping Beauty* transposon, which can be distributed randomly in the mouse genome, by remobilisation in the male germline [2]. Owing to the efficiency of this *in vivo* transposition system, we have recovered, identified and characterized a large number of insertions that provide a direct view of the regulatory activities associated with specific genomic regions. Furthermore, as the transposons used also carry a *loxP* site, the different lines can be used for *in vivo* chromosomal engineering, to generate mice with targeted deletions or duplications, or segmental aneuploidies [2,42-44]. The local hopping behaviour of *Sleeping Beauty* makes each line a potential starting point to scan a region of interest [45]: with our germline-specific transposase transgene, the remobilization rate ranges from 10 to 45%, depending on the starting site, and more than 15% of new insertions are within 1 Mb of the starting point. Thus, a research group with access to a limited number of cages can nonetheless set up a regional screen for its region of interest.

To provide a simple and useful access to the expression patterns and the mouse insertion strains generated with this on-going project, we have designed the **T**ransposon- and **R**ecombinase-**A**ssociated **C**hromosomal **E**ngineering **R**esource (TRACER) database. This new database is freely accessible at http://tracerdatabase.embl.de/. It constitutes a substantial improvement over the previous one that was established to display the data from a limited pilot screen [2]. The new database comprises novel features that allow users to browse and perform refined searches of insertion sites by position and/or expression patterns. The dataset is also now much larger (4-fold increase, with about 1500 insertions in July 2012), and is growing steadily. This web-based database not only provides information on regulatory activities present along the mouse genome but also gives access to a large collection of mice for engineering chromosomal rearrangements in non-genic intervals.

Construction and content

Dataset

In July 2012, the TRACER database contained information on 1467 insertions, 643 of which had been characterized for expression in mouse embryos (mostly at stage E11.5). Specific expression patterns were reported for 344 insertions, documented and annotated by 852 pictures. The dataset is updated regularly, with new insertions and new expression data. Most insertions were obtained with SB9 or SB8 transposons, which contain the regulatory sensor and a *loxP* site cloned in one or the other orientation in the *Sleeping Beauty* transposon. Newer versions of the transposon with additional features have been developed (Figure 2) and will be introduced in the database when mice with such insertions will be available.

Methodology and population of database

All our data is stored on a MySQL 5.5.15 RDBMS community server (GPL). Server side programming languages are PHP and PERL CGI. CSS and the javascript framework jQuery render the client data display and graphical user interfaces.

As well as the external user interfaces described below, the TRACER database has internal interfaces restricted to contributing members and requiring login for authentication. These internal interfaces have all the LIMS (laboratory information management system) components

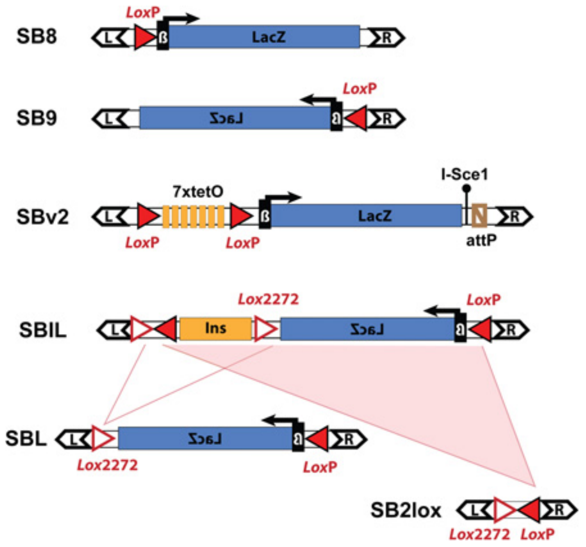


Figure 2 The different transposons used in TRACER. Within the left and right inverted/direct repeats of the SB transposon (black double arrows) different cargoes have been cloned that can be utilised for various purposes. Most lines contain an insertion of either the SB9 or SB8 transposon, comprising a LacZ reporter gene with a SV40 polyA sequence (blue rectangle), driven by a synthetic promoter composed of 50 bp of the human beta-globin promoter (black box with β), and a *loxP* site (red triangle). Newer transposons with additional features have been constructed, and mice with these transposons are being produced. Additional modules comprise sites for the PhiC31 integrase (attB), and the I-Sce1 meganuclease, which open possibilities to use the transposon as a docking site for incoming cassettes [46-48]. TetO binding sites can be used to recruit fusion proteins [49]. SBIL contains an insulator/enhancer-blocker element (Ins, orange block, from the chicken HSS4 element [50]) flanked by *loxP* and *lox2272* (white triangle with red contour). Cre-mediated recombination can transform this transposon into different derivatives (SBL or SB2lox), depending on which *lox* sites are used.

required for uploading data, curation of lines and various administration purposes.

The main internal interface allows lab staff to add all the text annotation, and insert sequence and image files associated with a particular TRACER line. There is also a batch upload interface for multiple insert sequences. The backend code automatically cuts the sequence down to just the insert, verifies the mutagen tag is present and the genomic sequence starts with 'TA'. The batch sequence submission tool is automatically coupled to the UCSC BLAT service with standard parameters (http://genome.ucsc.edu/) to determine the best alignment and genomic location for each insert. When there are multiple good alignments, user intervention is possible to select the best genomic location. An input form is then populated for the aligned sequence along with any existing data for the line. A similar batch upload interface exists for the parsing of the expression image and annotation files. Internal users can also edit annotations for existing lines using a separate curation interface.

Many of the interfaces utilise a controlled vocabulary of terms to populate the drop-down menus, reducing the number of typos in the database and preserving the integrity of the data stored in TRACER. An administrative database exists to edit these controlled vocabularies.

The external interface allows users to register interest in particular lines, or - if the user's genomic region of interest is not yet covered - to wish for such a line when it becomes available. These requests are captured in the database and matching lines are displayed for the curators so they can contact the requesting researcher. For user-defined regions of interest, new matching lines are

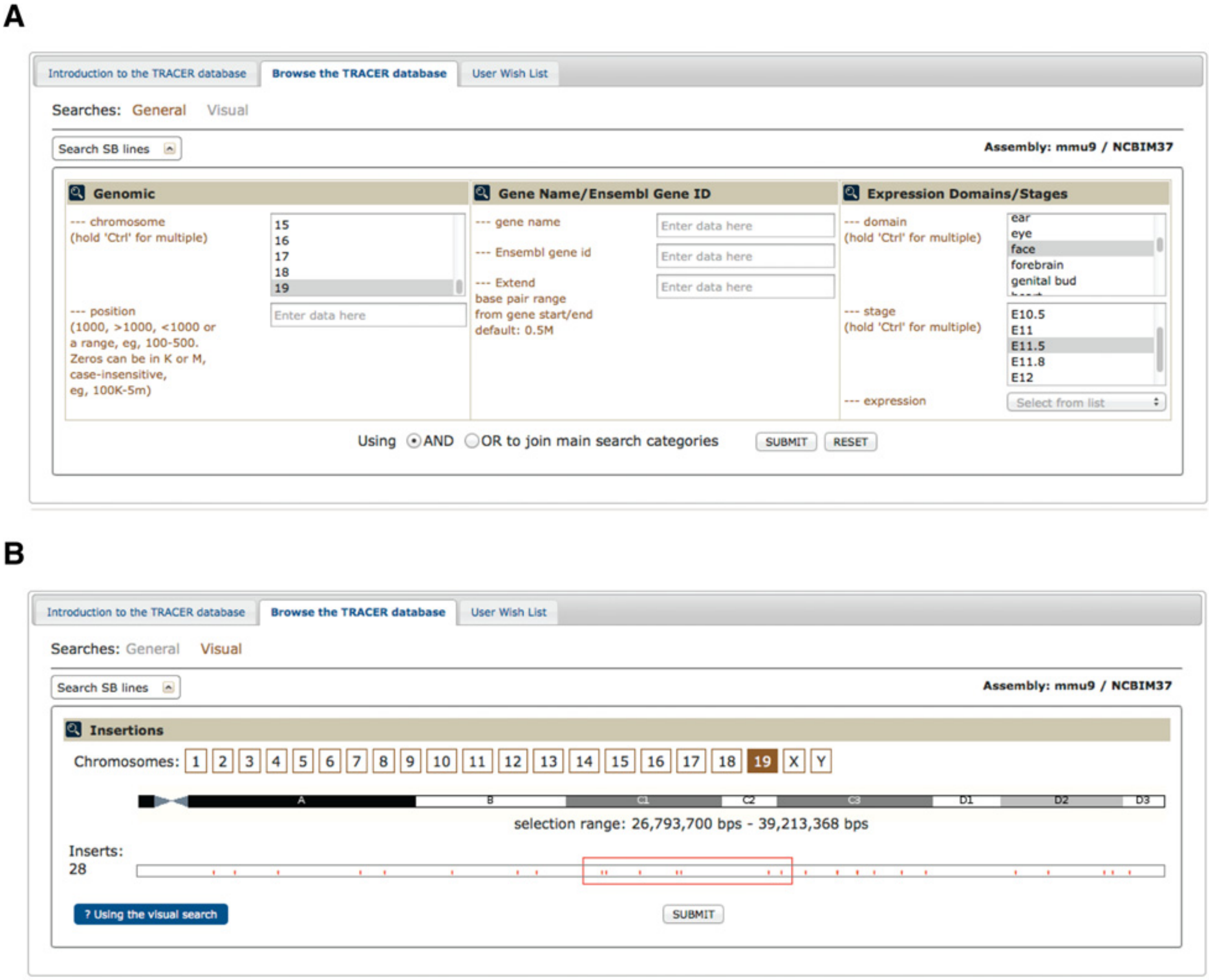
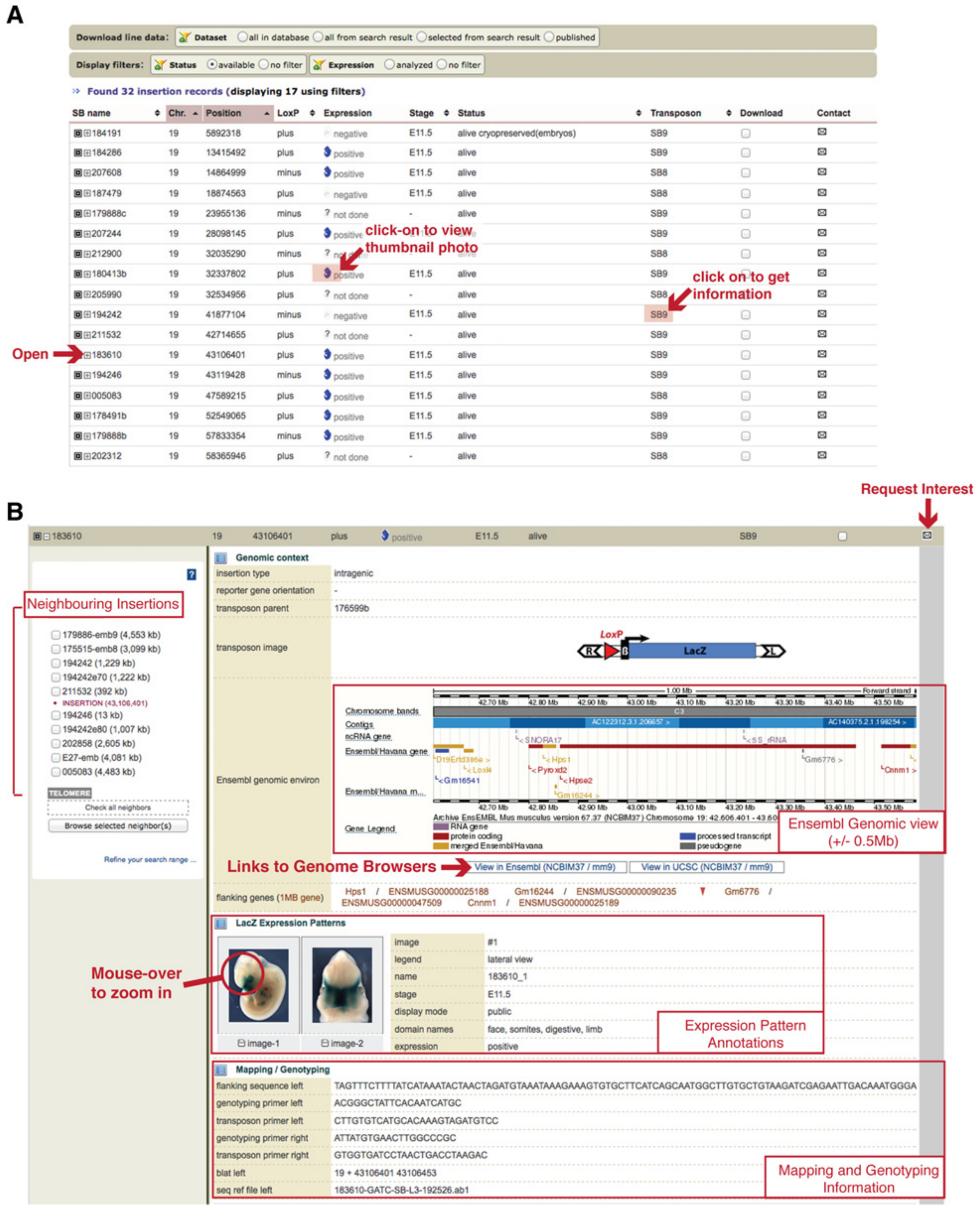

A
Introduction to the TRACER database
Browse the TRACER database
User Wish List
Searches: General Visual
Search SB lines
Assembly: mmu9 / NCBIM37
Genomic
--- chromosome
(hold 'Ctrl' for multiple)
15
16
17
18
19
--- position
(1000, >1000, <1000 or a range, eg, 100-500. Zeros can be in K or M, case-insensitive, eg, 100K-5m)
Enter data here
Gene Name/Ensembl Gene ID
--- gene name
--- Ensembl gene id
--- Extend base pair range from gene start/end default: 0.5M
Enter data here
Enter data here
Enter data here
Expression Domains/Stages
--- domain
(hold 'Ctrl' for multiple)
ear
eye
face
forebrain
genital bud
--- stage
(hold 'Ctrl' for multiple)
E10.5
E11
E11.5
E11.8
E12
--- expression
Select from list
Using ● AND ○ OR to join main search categories
SUBMIT
RESET
B
Introduction to the TRACER database
Browse the TRACER database
User Wish List
Searches: General Visual
Search SB lines
Assembly: mmu9 / NCBIM37
Insertions
Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y
A
B
C1
C2
C3
D1
D2
D3
selection range: 26,793,700 bps - 39,213,368 bps
Inserts:
28
? Using the visual search
SUBMIT

Figure 3 Searching the TRACER database. (**A**) TRACER lines can be retrieved ("General search") either based the genomic position of the insertion as defined by a genomic range or a region around a particular gene, or using the result of the expression assays, with the possibility to specify stage and expression domains of interest. (**B**) Alternatively ("Visual search"), the distribution of insertion sites along a chromosome is visualised, and those within the range defined by a dragable and expandable red rectangle will be returned.



A

Download line data: Dataset ○ all in database ○ all from search result ○ selected from search result ○ published

Display filters: Status ● available ○ no filter Expression ○ analyzed ○ no filter

» Found 32 insertion records (displaying 17 using filters)

SB name	Chr.	Position	LoxP	Expression	Stage	Status	Transposon	Download	Contact
⊞184191	19	5892318	plus	negative	E11.5	alive cryopreserved(embryos)	SB9	□	✉
⊞184286	19	13415492	plus	positive	E11.5	alive	SB9	□	✉
⊞207608	19	14864999	minus	positive	E11.5	alive	SB8	□	✉
⊞187479	19	18874563	plus	negative	E11.5	alive	SB8	□	✉
⊞179888c	19	23955136	minus	? not done	-	alive	SB9	□	✉
⊞207244	19	28098145	plus	positive	E11.5	alive	SB9	□	✉
⊞212900	19	32035290	minus	? not done	-	alive	SB8	□	✉
⊞180413b	19	32337802	plus	positive	E11.5	alive	SB9	□	✉
⊞205990	19	32534956	plus	? not done	-	alive	SB8	□	✉
⊞194242	19	41877104	minus	negative	E11.5	alive	SB9	□	✉
⊞211532	19	42714655	plus	? not done	-	alive	SB9	□	✉
⊞183610	19	43106401	plus	positive	E11.5	alive	SB9	□	✉
⊞194246	19	43119428	minus	positive	E11.5	alive	SB9	□	✉
⊞005083	19	47589215	plus	positive	E11.5	alive	SB8	□	✉
⊞178491b	19	52549065	plus	positive	E11.5	alive	SB9	□	✉
⊞179888b	19	57833354	minus	positive	E11.5	alive	SB9	□	✉
⊞202312	19	58365946	plus	? not done	-	alive	SB8	□	✉

B

⊟183610	19	43106401	plus	positive	E11.5	alive	SB9	□	✉

?

- □ 179886-emb9 (4,553 kb)
- □ 175515-emb8 (3,099 kb)
- □ 194242 (1,229 kb)
- □ 194242e70 (1,222 kb)
- □ 211532 (392 kb)
- • INSERTION (43,106,401)
- □ 194246 (13 kb)
- □ 194242e80 (1,007 kb)
- □ 202858 (2,605 kb)
- □ E27-emb (4,081 kb)
- □ 005083 (4,483 kb)

TELOMERE

Check all neighbors

Browse selected neighbor(s)

Refine your search range ...

Genomic context

insertion type	intragenic
reporter gene orientation	-
transposon parent	176599b
transposon image	
Ensembl genomic environ	Chromosome bands Contigs ncRNA gene Ensembl/Havana gene Ensembl/Havana m... Archive EnsEMBL Mus musculus version 67.37 (NCBIM37) Chromosome 19: 42.606.401 - 43.60 Gene Legend: ▨ RNA gene ▨ protein coding ▨ merged Ensembl/Havana ▨ processed transcript ▨ pseudogene
	View in Ensembl (NCBIM37 / mm9) View in UCSC (NCBIM37 / mm9)
flanking genes (1MB gene)	Hps1 / ENSMUSG00000025188 Gm16244 / ENSMUSG00000090235 ▼ Gm6776 / ENSMUSG00000047509 Cnnm1 / ENSMUSG00000025189

LacZ Expression Patterns

⊟ image-1 ⊟ image-2

image	#1
legend	lateral view
name	183610_1
stage	E11.5
display mode	public
domain names	face, somites, digestive, limb
expression	positive

Mapping / Genotyping

flanking sequence left	TAGTTTCTTTTATCATAAATACTAACTAGATGTAAATAAAGAAAGTGTGCTTCATCAGCAATGGCTTGTGCTGTAAGATCGAGAATTGACAAATGGGA
genotyping primer left	ACGGGCTATTCACAATCATGC
transposon primer left	CTTGTGTCATGCACAAAGTAGATGTCC
genotyping primer right	ATTATGTGAACTTGGCCCGC
transposon primer right	GTGGTGATCCTAACTGACCTAAGAC
blat left	19 + 43106401 43106453
seq ref file left	183610-GATC-SB-L3-192526.ab1

Figure 4 (See legend on next page.)

(See figure on previous page.)
Figure 4 Presentation of TRACER data. (**A**) Summary display of results for a TRACER search. For each line the TRACER name, genomic position, orientation of the *loxP* site, summary of expression (expressed/positive, not expressed/negative, not done; developmental stage(s) assayed), status (alive, cryopreserved, not maintained, newly created) and the transposon type are displayed. The final two columns allow users to select the data for download, or to register interest in the line. Clicking on the "open" icon gives access to the detailed information of an insertion of interest. Quick access to a thumbnail photo of a representative embryo and to the corresponding transposon is possible by a simple click on the corresponding zone (**B**) Detailed view of a TRACER mouse line. The top section reveals basic information about the insertion, including the type of insertion, the name of the parent insertion, the orientation of the *loxP* site and a visual representation of the transposon in the correct orientation. The first panel shows the genomic context of the insertion in a snapshot from the Ensembl genome browser (Ensembl Genomic view +/−0.5 Mb) along with links to view the insertion point in the Ensembl or UCSC genome browsers. The second panel shows photos and annotations of the expression patterns of the regulatory sensor. The images can be mouse-overed to reveal a high-resolution zoomed-in view, and the annotation of expression domains is displayed. The third panel shows information related to mapping and genotyping, including the sequence obtained from the mapping procedure, and the sequences of the primers used to genotype animals carrying this specific insertion. The interface in the panel on the left allows neighbouring insertions to be selected for detailed analysis of a region's regulatory potential.

automatically searched every week, or when triggered through the curator interface.

Utility

Searching the TRACER database

The "Browse the TRACER database" tab takes users to the main search interface of the website (Figure 3A). Insertions of interest can be identified by a variety of options. For genome-centric views, a genomic region of interest can be specified, defined either by chromosomal coordinates (reference genome is MGSCv37/mm9), or by a gene name ("associated gene name" from Ensembl database) and an optional user-defined flanking region (default is 0.5 Mb). In addition to this "General" option, one can perform a "Visual" search by clicking on the link at the top of the search window (Figure 3B). Users can view the distribution of insertions across each chromosome and drag a rectangle to define the region they want to retrieve lines. Searches can also be carried out based on expression patterns, using criteria such as positive/negative, expression domains and expression stages. The label "negative" for expression means that no specific expression patterns was scored for this insertion at the embryonic stages assayed, whereas an insertion is labelled as "positive" if specific expression is detected at least at one stage of development. The majority of the insertions have been assayed at E11.5, but some data is available at other stages (E10 to E13). Expression domains are annotated using a simplified controlled vocabulary (e.g. branchial arches, cranial ganglia, digestive, dorsal root ganglia, ear, eye, face, forebrain, genital bud, heart, hindbrain, limb, midbrain, neural tube, somites, urogenital, others or widespread), which is compatible with the one used by the Vista Enhancer Database [28], in order to facilitate comparison of the two datasets.

Display and download of data

Results are returned as a table (Figure 4A) with one row per insertion and sortable columns displaying:

- The internal identifier of the mouse line in the TRACER database.
- The genomic position of the insertion (chr/position ; based on MGSCv37/mm9 genome assembly).
- The orientation of the *loxP* site in the transposon. "Plus" corresponds to the following orientation: centromere – 5′-ATAACTTCGTATA GCATACATTATACGAAGTTAT- 3′ telomere. For comparison, *loxP* sites targeted by the International Knockout Mouse Consortium in genes transcribed from the plus strand (http://www.knockoutmouse.org/about/targeting-strategies) have the same orientation than TRACER "plus" *loxP*. Depending on the specific transposon, the orientation of the other features (transposon ends, reporter gene) varies: they are indicated and represented in the expanded view available by clicking on the "expand" icon.
- An icon and text, indicating whether expression analysis has been performed and whether LacZ reporter expression has been detected. The developmental stage(s) for which information is available are indicated in the next column. Expression assay is "positive" if the insertion showed LacZ staining at least at one of the stages assessed.
- The status of the insertion, indicating whether animals carrying the insertion are available. Insertions that were identified in F0 embryos, that couldn't be established from the founder or were discontinued, are labelled as "not maintained". Insertions "available" for further use or analysis fall under three categories: "alive" (line established with mice available in small numbers), "cryopreserved" (either as embryos or sperm) and "new" (usually corresponding to a new insertion, with only the founder animal). The status of an insertion is dynamic: not all "new" insertions are established, and depending on circumstances, "alive" ones may become "cryopreserved" or "not maintained".
- Transposon type: most of the available lines harbour a simple regulatory sensor with a lacZ reporter and

a single *loxP* site, in one or the other orientation relative to the transposon ends (SB8 and SB9). New transposons with additional features have been constructed (see Figure 2), and lines containing them are being established and will be added to the resource. Detailed maps and sequences of available transposons are available on the Tracer website.

The final two columns display a checkbox to download the complete set of information available for an insertion, and an email link to indicate interest in a specific insertion. The toolbar buttons above the results table can be used to filter the search results, and to show only available lines and/or lines with expression data.

Further details on a given insertion can be seen by clicking the expand icon next to each record (Figure 4B). The first section describes the genomic context of the insertion. It lists whether the insertion is located in a gene desert (a gene-free region larger than 500 kb), intergenic (less than 500 kb-long), intronic or exonic region, specifies the orientation of the reporter gene, and the parental insertion line from which the insertion was obtained. This section also contains a schematic of the transposon construct, the genomic environment and flanking genes in a snapshot from the Ensembl genome browser [51] along with links to view the insertion point in Ensembl or the UCSC genome browser [52].

The second section shows the LacZ expression patterns obtained for the insertion, when available. Mousing over each thumbnail image show a zoomed-in, trackable high-resolution view of the image. In addition, the stage and viewpoint of the image is recorded along with annotations using the expression domain categories detailed above. One can switch from one image to another one by clicking on the corresponding thumbnail.

The final section shows details regarding how the genomic position of the insertion was determined, such as the flanking sequence(s) obtained (trimmed to the TA dinucleotide duplicated upon Sleeping Beauty insertion [53]), and where this sequence mapped where this sequence mapped to genome using BLAT [54]. When available, primers that have been used to genotype embryos and mice for this specific insertion are indicated.

The left hand panel of the expanded section contains an interface that displays lines with insertion points within 5 Mb (or a user-selected range) (Figure 4B). Users can select one or more of these lines, and open a new tab displaying these flanking lines. This feature is particularly useful to compare regulatory activities across large regions, and to delineate the extent of regulatory domains.

Finally, the toolbar below the search interface allows data to be downloaded for the whole TRACER database, the search results, user selected lines or just the lines described in publications referring to the dataset. Additionally, all available images can be downloaded. Requests for higher resolution photos and other questions can be sent to gromit@embl.de. Most LacZ stained embryos has been archived, albeit in limited numbers for each insertion, and may be made available upon request.

User wish list

Although the TRACER database already covers a substantial proportion of the genome, it is likely that individual researchers will be interested to get information and mice with transposons in regions where we haven't yet identified an insertion. Given the high efficiency of transposition, the number of new insertions identified in on-going remobilisation efforts (~ 10 per week) exceeds our current capacity to keep, expand and cryopreserve all of them (Figure 5). The "User wish list" tab allows scientists to indicate particular genomic regions they are interested in, along with their contact details (Figure 5C). Once an insertion in this region is identified, it is "flagged" for the producing group, so that the corresponding animal is kept, and the interested group will be contacted.

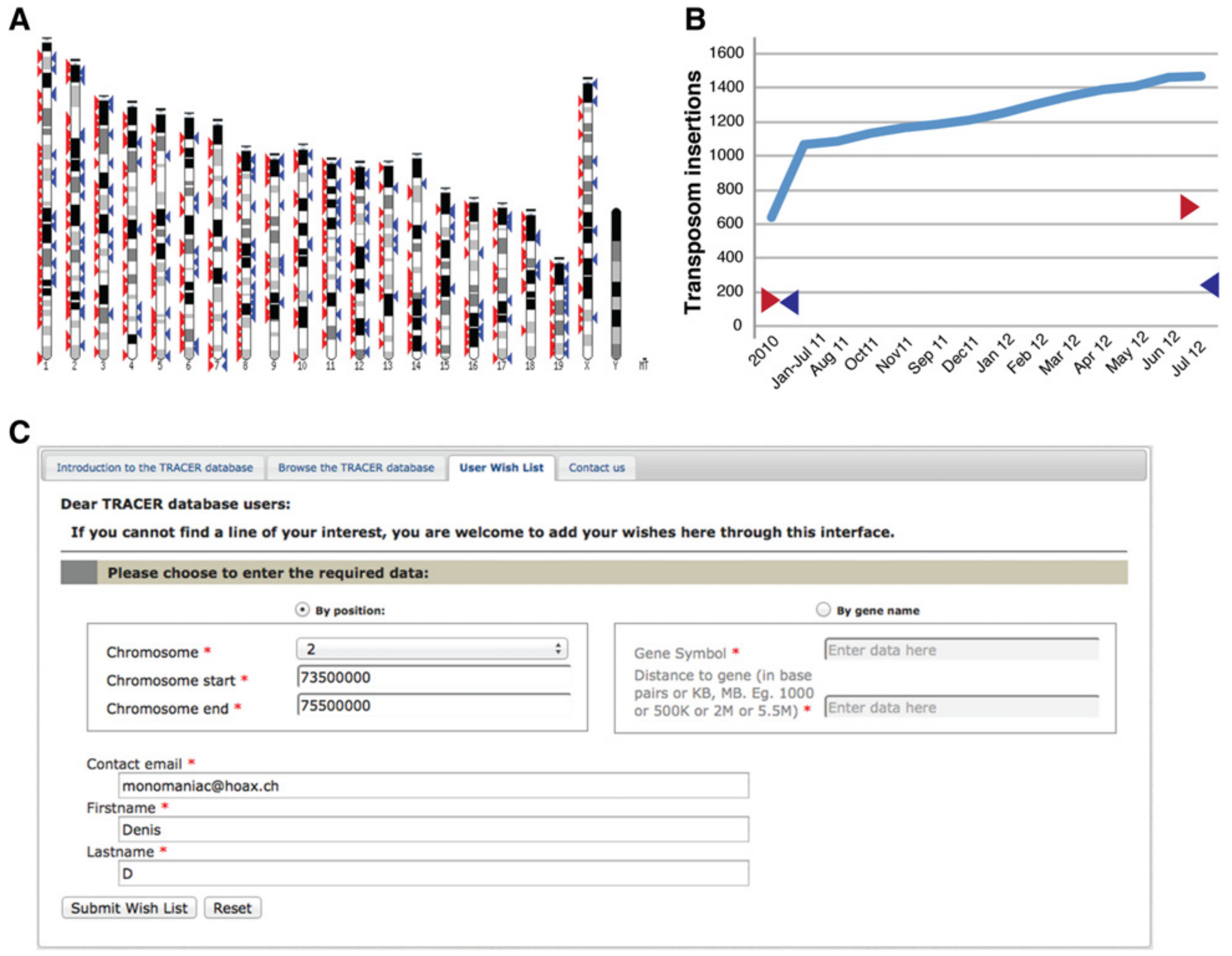
Discussion

A functional view of the genome with TRACER

The introduction of a "regulatory sensor" in the genome provides a direct operational readout of the activities that can contribute to gene expression, which surround the insertion point. Similar enhancer-trap screens have widely been used in *Drosophila* [41] and to some extent in zebrafish [55-58], providing information about genes and genomes, as well as a series of useful markers and tools. Their use in mice has been limited [59,60], in part due to the low throughput of transgenesis, and technical difficulties of generating single-copy insertions. The development of robust and efficient *in vivo* transposition systems [2,61-63], as shown here, or the use of lentiviral transgenesis, as recently described elsewhere [64], open new exciting possibilities to conduct such screens in an efficient and affordable manner.

Collections of insertions generated by these approaches can provide useful information and tools, and the TRACER database represents a substantial step to capitalise on such a collection, by centralizing and giving access to data and to mouse lines. We present and discuss here briefly some of the possible uses of this database and of the information therein (Figure 6).

By querying the database for a gene or a region of interest, one can identify expression patterns and regulatory activities associated with that location and its surroundings. The observed activity may indicate possible developmental or tissue-specific regulation of genes, and shed light on their physiological roles *in vivo* (Figure 6A).



Dear TRACER database users:

If you cannot find a line of your interest, you are welcome to add your wishes here through this interface.

Please choose to enter the required data:

● By position:

Chromosome * 2
Chromosome start * 73500000
Chromosome end * 75500000

○ By gene name

Gene Symbol * Enter data here
Distance to gene (in base pairs or KB, MB. Eg. 1000 or 500K or 2M or 5.5M) * Enter data here

Contact email * monomaniac@hoax.ch
Firstname * Denis
Lastname * D

Figure 5 Evolution of TRACER dataset. (**A**) Genomic distribution of the insertions with expression annotations (red) and available as lines (blue) (generated with the Karyoview tool from Ensembl – status in July 2012). Insertions on the Y chromosome are not shown as their positions are ambiguous (**B**) The number of mapped insertions increases at a regular pace (light blue line), almost paralleled by expression reports (red cursor). A smaller proportion of these are kept alive or archived by cryopreservation (blue cursor). (**C**) The user wish list allows researchers to register interest for a genomic interval. It ensures that mice with new insertions in this region will be preserved for the scientist who indicated this request. The interval of interest can be defined by position or by gene name.

However, we wish to emphasize that the regulatory sensor sometimes reflects only a subset of the expression domains of a given gene [2]. Although the sensor responds accurately to influences from long-range remote enhancers, it is less likely to capture the input of promoter elements that have a limited range of action: tissue-restricted expression of the sensor may therefore represent a tissue-specific modulation of an otherwise broadly expressed gene; yet, this modulation may correspond to important biological functions.

Also, the expression pattern associated with an insertion does not necessarily imply that a corresponding enhancer lies nearby, as illustrated by the shared expression of distant insertions (Figure 6A,C; other examples in [2]). Instead, the sensor reports the collective input at a given position of both positive and negative regulatory elements. Accordingly, comparing the expression pattern of neighbouring insertions to each other and to known enhancer activities [21,65] can reveal important regulatory features. These include the range of action of enhancers, the boundaries of expression domains, the presence of silencers or other repressive or insulating elements that modulate enhancer activity and cannot be obtained from other types of datasets and approaches. In essence, TRACER provides an operational view of the regulatory structure of the mammalian genome, and delineates the extent of the large *regulatory landscapes* [6] that subdivide the genome into functional units. It constitutes a functional counterpart to views obtained by different methods; including, for example, *Genome Regulatory Blocks* that are delineated by the density of conserved non-coding elements and synteny conservation

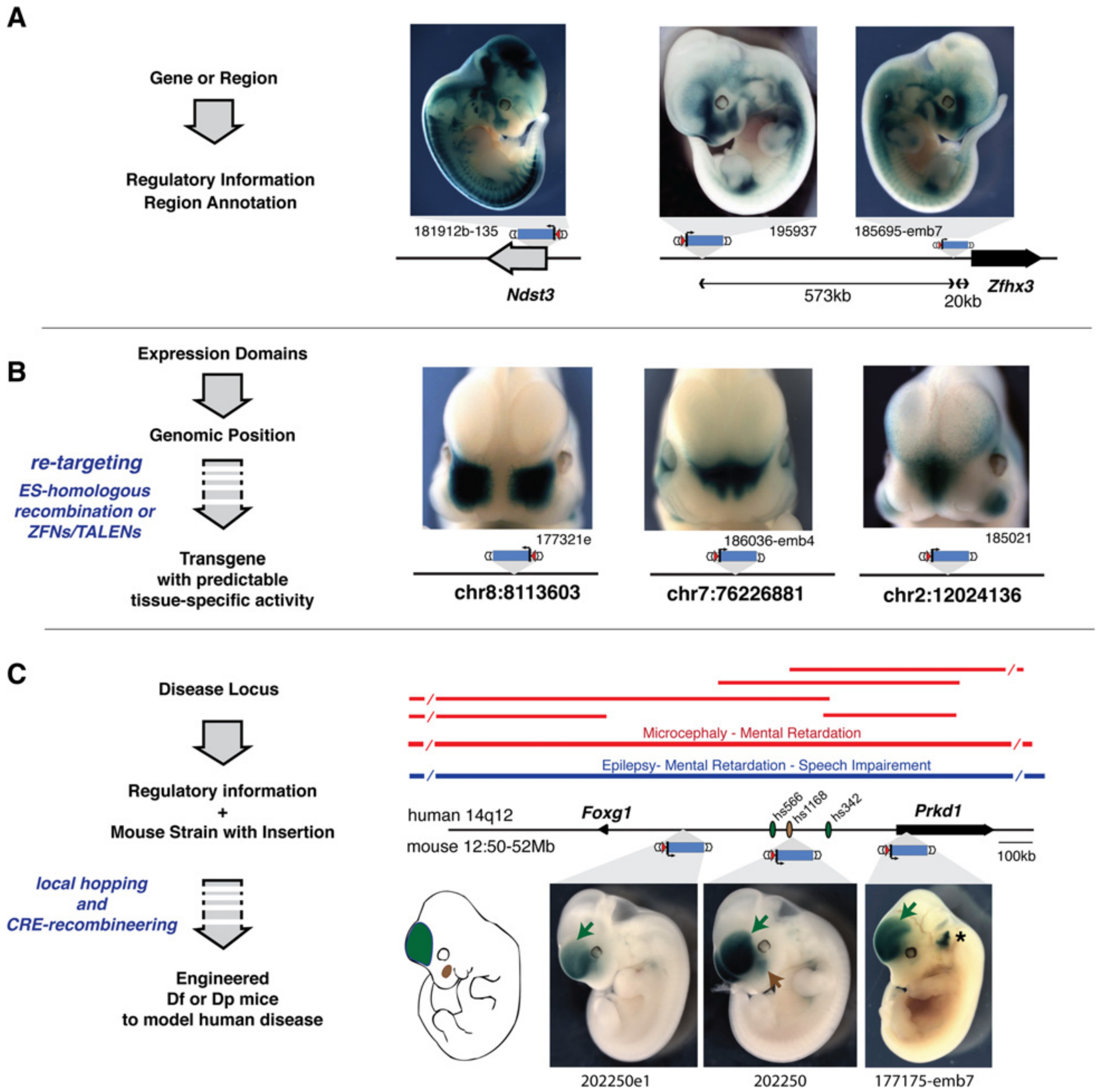

A
Gene or Region
Regulatory Information
Region Annotation
181912b-135
Ndst3
195937
185695-emb7
573kb
20kb
Zfhx3
B
Expression Domains
Genomic Position
re-targeting
ES-homologous recombination or ZFNs/TALENs
Transgene with predictable tissue-specific activity
177321e
chr8:8113603
186036-emb4
chr7:76226881
185021
chr2:12024136
C
Disease Locus
Regulatory information + Mouse Strain with Insertion
local hopping and CRE-recombineering
Engineered Df or Dp mice to model human disease
Microcephaly - Mental Retardation
Epilepsy- Mental Retardation - Speech Impairement
human 14q12
mouse 12:50-52Mb
Foxg1
hs566
hs1168
hs342
Prkd1
100kb
*
202250e1
202250
177175-emb7

Figure 6 Examples of potential applications of TRACER. (**A**) Using TRACER to obtain information about the regulatory potential of a specific interval. Two loci (*Ndst3* and *Zfhx3*) are given as an example, with photos of lacZ-stained embryos with insertions close or far from these genes. (**B**) Searching TRACER for insertions with specific expression patterns can identify genomic positions where an inserted transgene will display a predicted activity. Three insertions with different patterns of expression in the fronto-nasal process are shown as examples. Such specific patterns can subsequently be exploited to re-target other transgenes with expression in the predicted domains through different techniques. (**C**) Overlapping deletions (red bars) and duplications (blue bars) over 14q12 entered in the DECIPHER database (http://decipher.sanger.ac.uk/) are associated with similar clinical phenotypes. TRACER insertions in the orthologous interval show that it corresponds to a large regulatory landscape associated with forebrain (green) and facial (brown) expression domains in E11.5 (SB-202250e1 and SB-202250) and E12.5 (SB-177175emb7) mouse embryos. LacZ expression in the ear of embryo SB-177175emb7 (star) is due to the presence of an additional insertion, SBlacC [2]. Enhancer regions (ovals, color-coded depending on their pattern of activity as shown in the schematized embryo; data from VISTA enhancer browser [28]) present in the proximity may contribute to these activities. By local remobilisation and Cre-mediated recombineering, the mouse line SB-202250 may be used to produce mouse models of the human aneuploidies associated with this largely non-coding interval.

[66,67], *Topological Associated Domains* defined by chromosomal interaction biases [68,69], and *Enhancer-Promoter Units* that are revealed by clusters of coincident promoter-enhancer chromatin signatures [70].

The data present in TRACER identifies genomic positions where an inserted transgene will adopt a highly specific expression profile (Figure 6B). Transgenes that drive the expression of markers to label specific cells (such as fluorescent markers) or of effector genes (for example Cre recombinase) in defined cell-types or embryonic tissues have proven very useful to dissect biological and genetic processes. "Position-effects" (the action of endogenous regulatory elements on transgenes) are usually considered as a problem for transgenic experiments because they lead to partially unpredictable outcomes. With the information displayed in TRACER, one can instead exploit position effects, and select genomic sites that will convey an expression pattern of interest. Importantly, many of these sites are located far from genes, implying that their use would have less functional impact than a gene knock-in. The sensor integrates the inputs of both enhancers and silencers that are acting at its position: consequently, the observed pattern is often more restricted than the one driven by enhancer-only constructs or displayed by the neighbouring genes [2]. Hence, retargeting positions identified in TRACER with a transgene of interest should provide a reliable method to create new tissue- and cell-type specific transgenes. This can be done by homologous recombination in mouse ES cells, but the rapid development of Zinc-Finger or TALE Nuclease-associated targeted transgenesis may offer more efficient alternatives [46,71,72].

In addition to maps of genomic "regulatory landscapes", TRACER provides access to a large and growing collection of mice with different transposon insertions (around 200 in July 2012). Only few insertions are likely to disrupt genes or key/highly conserved regulatory elements directly. Instead, these mice can be used for other purposes, and in particular for engineering aneuploidies and structural variants. Chromosomal aneuploidies are often found in patients suffering developmental malformations and/or neuropsychiatric disorders. In some cases, single gene-knockout can reproduce the phenotypes observed in human patients; however, for numerous other conditions, such as contiguous gene diseases, chromosomal duplications or rearrangements in non-coding intervals, gene-based alleles do not provide accurate models. Because *Sleeping Beauty* transposons frequently re-insert in the vicinity of their initial position, it is possible to use one insertion in a region of interest to generate additional local re-insertions. These insertions can be (re)combined owing to the associated *loxP* sites, to produce a series of rearrangements of this locus that model genomic alterations found in human patients, and help determine the causal elements or genes (Figure 6C). Such a use of the TRACER resource and GROMIT strategy can be particularly well suited for large gene clusters (eg. proto-cadherins, KRAB-zinc finger genes, olfactory receptors) or gene-deserts associated with human pathologies, complementing the gene-centric resource provided by the International Knockout Mouse Consortium. Given the growing recognition of the biological importance of genomic structural variants for human diseases, we anticipate that TRACER will be a useful resource to rapidly engineer allelic series of structural variants in mouse orthologous intervals, helping to create novel models of human genomic disorders.

Conclusion

TRACER database and community

Owing to the dynamic nature of transposon elements, the resource present in TRACER will expand steadily with the number of users. Each lab using this transposon technology to investigate a region of interest by "local" hopping will produce a substantial number of by-products (~ 80% of the new insertions). Even if these insertions may not be useful for the producing lab, they can be of interest for others. TRACER is designed to serve as a central "virtual" repository to share those mice. Further information, including references, detailed maps and sequences of the different transposons and transgenes in use, and protocols for mapping of new insertions are available through the pages of the TRACER website.

To facilitate exchanges, the TRACER database incorporates several features and internal interfaces for contributing groups (automated insertion mapping, annotation and administration). In particular, the "User wish list" feature offers a simple manner to readily "tag" newly generated mice of interest without a major investment or commitment of the producing labs.

Availability and requirements

The database is accessible at the web addresses:

http://tracerdatabase.embl.de
http://www.ebi.ac.uk/panda-srv/tracer/index.php

Websites – links

ENSEMBL: http://www.ensembl.org/Mus_musculus/Info/Index

UCSC Genome Browser: http://genome.ucsc.edu/index.html

CTCFBSBD: http://insulatordb.uthsc.edu/

VISTA: http://enhancer.lbl.gov/frnt_page_n.shtml

GXD: http://www.informatics.jax.org/mgihome/GXD/aboutGXD.shtml

EMAGE: http://www.emouseatlas.org/emage/

GENSAT: http://www.gensat.org/index.html

MAMEP: http://mamep.molgen.mpg.de/index.php

EUREXPRESS: http://www.eurexpress.org/
EMBRYS: http://embrys.jp/embrys/html/MainMenu.html
DECIPHER: http://decipher.sanger.ac.uk/

Competing interests

The authors declare that they have no competing interest.

Authors' contributions

CKC, DS and FS designed the database with critical input from OS, VVU, TT, SR. CKC wrote the code, the different interfaces and tools associated with TRACER. OS, VVU, TT, SR and FS provided all data present in the database. DS and FS wrote the manuscript, with input and suggestions from all the other authors. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the members of the EMBL Laboratory Animal Resource, in particular Andrea Schulz, Silke Feller, Michaela Wesch and Klaus Schmitt; the development and maintenance of this mouse resource would not be possible without their dedication and constant support. We thank as well Manuela Borchert and Anne Hermelin for advice and design of the TRACER webpages. The TRACER database and resource is funded by EMBL. The production of some of the strains described in the database was made in the course of projects supported by the European Commission-FP7 (grant Health 223210/CISSTEM) and Human Frontier Science Program (grant RGY0081/2008-C) to FS.

Author details

[1]European Bioinformatics Institute - European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [2]Developmental Biology Unit - European Molecular Biology Laboratory, Heidelberg, Germany. [3]Present address: Welcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Received: 31 October 2012 Accepted: 22 March 2013
Published: 2 April 2013

References

1. Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB: **Annotating non-coding regions of the genome.** *Nat Rev Genet* 2010, **11:**559–571.
2. Ruf S, Symmons O, Uslu VV, Dolle D, Hot C, Ettwiller L, Spitz F: **Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor.** *Nat Genet* 2011, **43:**379–386.
3. Visel A, Rubin EM, Rubin EM, Pennacchio LA, Pennacchio LA: **Genomic views of distant-acting enhancers.** *Nature* 2009, **461:**199–205.
4. Jeong Y, El-Jaick K, Roessler E, Muenke M, Epstein DJ: **A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers.** *Development* 2006, 133:761–772.
5. Zuniga A, Michos O, Spitz F, Haramis A-PG, Panman L, Galli A, Vintersten K, Klasen C, Mansfield W, Kuc S, Duboule D, Dono R, Zeller R: **Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression.** *Genes Dev* 2004, **18:**1553–1564.
6. Spitz F, Gonzalez F, Duboule D: **A global control region defines a chromosomal regulatory landscape containing the HoxD cluster.** *Cell* 2003, **113:**405–417.
7. Montavon T, Soshnikova N, Mascrez B, Joye E, Thevenet L, Splinter E, de Laat W, Spitz F, Duboule D: **A regulatory archipelago controls Hox genes transcription in digits.** *Cell* 2011, **147:**1132–1145.
8. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM: **Scanning human gene deserts for long-range enhancers.** *Science* 2003, **302:**413.
9. Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.** *J Neurosci Res* 2003, **12:**1725–1735.
10. Uchikawa M, Ishida Y, Takemoto T, Kamachi Y, Kondoh H: **Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals.** *Dev Cell* 2003, **4:**509–519.
11. Kleinjan DA, Seawright A, Mella S, Carr CB, Tyas DA, Simpson TI, Mason JO, Price DJ, van Heyningen V: **Long-range downstream enhancers are essential for Pax6 expression.** *Dev Biol* 2006, **299:**563–581.
12. Lettice LA, Horikoshi T, Heaney SJH, van Baren MJ, van der Linde HC, Breedveld GJ, Joosse M, Akarsu N, Oostra BA, Endo N, Shibata M, Suzuki M, Takahashi E, Shinka T, Nakahori Y, Ayusawa D, Nakabayashi K, Scherer SW, Heutink P, Hill RE, Noji S: **Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly.** *Proc Natl Acad Sci USA* 2002, **99:**7548–7553.
13. Kleinjan DA, van Heyningen V: **Long-range control of gene expression: emerging mechanisms and disruption in disease.** *Am J Hum Genet* 2005, **76:**8–32.
14. Jeong Y, Leskow FC, El-Jaick K, Roessler E, Muenke M, Yocum A, Dubourg C, Li X, Geng X, Oliver G, Epstein DJ: **Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein.** *Nat Genet* 2008, **40:**1348–1353.
15. Benko S, Fantes JA, Amiel J, Kleinjan D-J, Thomas S, Ramsay J, Jamshidi N, Essafi A, Heaney S, Gordon CT, McBride D, Golzio C, Fisher M, Perry P, Abadie V, Ayuso C, Holder-Espinasse M, Kilpatrick N, Lees MM, Picard A, Temple IK, Thomas P, Vazquez M-P, Vekemans M, Crollius HR, Hastie ND, Munnich A, Etchevers HC, Pelet A, Farlie PG, *et al:* **Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence.** *Nat Genet* 2009, **41:**359–364.
16. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, Lie RT, Wilcox AJ, Fitzpatrick DR, Green ED, Mossey PA, Little J, Steegers-Theunissen RP, Pennacchio LA, Schutte BC, Murray JC: **Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip.** *Nat Genet* 2008, **40:**1341–1347.
17. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Björklund M, Wei G, Yan J, Niittymäki I, Mecklin J-P, Järvinen H, Ristimäki A, Di-Bernardo M, East P, Carvajal-Carmona L, Houlston RS, Tomlinson I, Palin K, Ukkonen E, Karhu A, Taipale J, Aaltonen LA: **The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling.** *Nat Genet* 2009, **41:**885–890.
18. Wasserman NF, Aneas I, Nobrega MA: **An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer.** *Genome Res* 2010, **20:**1191–1197.
19. Visser M, Kayser M, Palstra RJ: **HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.** *Genome Res* 2012, **22:**446–455.
20. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu X-D, Topol EJ, Rosenfeld MG, Frazer KA: **9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response.** *Nature* 2011, **470:**264–268.
21. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA: **ChIP-seq accurately predicts tissue-specific activity of enhancers.** *Nature* 2009, **457:**854–858.
22. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression.** *Nature* 2009, **459:**108–112.
23. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R: **Histone H3K27ac separates active from poised enhancers and predicts developmental state.** *Proc Natl Acad Sci USA* 2010, **107:**21931–21936.
24. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J: **A unique chromatin signature uncovers early developmental enhancers in humans.** *Nature* 2011, **470:**279–283.
25. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei C-L, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, *et al*: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.** *Cell* 2012, **148:**84–98.
26. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473:**43–49.
27. Consortium TEP: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57–74.
28. Visel A, Minovitsky S, Dubchak I, Pennacchio LA: **VISTA Enhancer Browser–a database of tissue-specific human enhancers.** *Nucleic Acids Res* 2007, **35:**D88–92.

29. Finger JH, Smith CM, Hayamizu TF, McCright IJ, Eppig JT, Kadin JA, Richardson JE, Ringwald M: **The mouse Gene Expression Database (GXD): 2011 update.** *Nucleic Acids Res* 2011, **39:**D835–41.
30. Heintz N: **Gene expression nervous system atlas (GENSAT).** *Nat Neurosci* 2004, **7:**483.
31. Yokoyama S, Ito Y, Ueno-Kudoh H, Shimizu H, Uchibe K, Albini S, Mitsuoka K, Miyaki S, Kiso M, Nagai A, Hikata T, Osada T, Fukuda N, Yamashita S, Harada D, Mezzano V, Kasai M, Puri PL, Hayashizaki Y, Okado H, Hashimoto M, Asahara H: **A systems approach reveals that the myogenesis genome network is regulated by the transcriptional repressor RP58.** *Dev Cell* 2009, **17:**836–848.
32. Richardson L, Venkataraman S, Stevenson P, Yang Y, Burton N, Rao J, Fisher M, Baldock RA, Davidson DR, Christiansen JH: **EMAGE mouse embryo spatial gene expression database: 2010 update.** *Nucleic Acids Res* 2010, **38:** D703–9.
33. Visel A, Thaller C, Eichele G: **GenePaint.org: an atlas of gene expression patterns in the mouse embryo.** *Nucleic Acids Res* 2004, **32:**D552–6.
34. Diez-Roux G, Banfi S, Sultan M, Geffers L, Anand S, Rozado D, Magen A, Canidio E, Pagani M, Peluso I, Lin-Marq N, Koch M, Bilio M, Cantiello I, Verde R, De Masi C, Bianchi SA, Cicchini J, Perroud E, Mehmeti S, Dagand E, Schrinner S, Nürnberger A, Schmidt K, Metz K, Zwingmann C, Brieske N, Springer C, Hernandez AM, Herzog S, *et al:* **A high-resolution anatomical atlas of the transcriptome in the mouse embryo.** *PLoS Biol* 2011, **9:**e1000582.
35. Neidhardt L, Gasca S, Wertz K, Obermayr F, Worpenberg S, Lehrach H, Herrmann BG: **Large-scale screen for genes controlling mammalian embryogenesis, using high-throughput gene expression analysis in mouse embryos.** *Mech Dev* 2000, **98:**77–94.
36. Klopocki E, Ott C, Benatar N, Ullmann R, Mundlos S, Lehmann K: **A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome.** *J Med Genet* 2008, **45:**370–375.
37. Sun M, Ma F, Zeng X, Liu Q, Zhao X, Wu F, Wu G, Zhang Z, Gu B, Zhao Y, Tian S, Lin B, Kong X, Zhang X, Yang W, Lo W: **Triphalangeal thumb-polysyndactyly syndrome and syndactyly type IV are caused by genomic duplications involving the long-range, limb-specific SHH enhancer.** *J Med Genet* 2008, **45:**589–595.
38. Dathe K, Kjaer KW, Brehm A, Meinecke P, Nürnberg P, Neto JC, Brunoni D, Tommerup N, Ott CE, Klopocki E, Seemann P, Mundlos S: **Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type A2.** *Am J Hum Genet* 2009, **84:**483–492.
39. Dimitrov BI, de Ravel T, Van Driessche J, de Die-Smulders C, Toutain A, Vermeesch JR, Fryns JP, Devriendt K, Debeer P: **Distal limb deficiencies, micrognathia syndrome, and syndromic forms of split hand foot malformation (SHFM) are caused by chromosome 10q genomic rearrangements.** *J Med Genet* 2010, **47:**103–111.
40. Kurth I, Klopocki E, Stricker S, van Oosterwijk J, Vanek S, Altmann J, Santos HG, van Harssel JJT, de Ravel T, Wilkie AOM, Gal A, Mundlos S: **Duplications of noncoding elements 5' of SOX9 are associated with brachydactyly-anonychia.** *Nat Genet* 2009, **41:**862–863.
41. Bellen HJ: **Ten years of enhancer detection: lessons from the fly.** *Plant Cell* 1999, **11:**2271–2281.
42. Hérault Y, Rassoulzadegan M, Cuzin F, Duboule D: **Engineering chromosomes in mice through targeted meiotic recombination (TAMERE).** *Nat Genet* 1998, **20:**381–384.
43. Wu S, Ying G, Wu Q, Capecchi MR: **Toward simpler and faster genome-wide mutagenesis in mice.** *Nat Genet* 2007, **39:**922–930.
44. Spitz F, Herkenne C, Morris MA, Duboule D: **Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes.** *Nat Genet* 2005, **37:**889–893.
45. Keng VW, Yae K, Hayakawa T, Mizuno S, Uno Y, Yusa K, Kokubu C, Kinoshita T, Akagi K, Jenkins NA, Copeland NG, Horie K, Takeda J: **Region-specific saturation germline mutagenesis in mice using the Sleeping Beauty transposon system.** *Nat Methods* 2005, **2:**763–769.
46. Meyer M, de Angelis MH, Wurst W, Kühn R: **Gene targeting by homologous recombination in mouse zygotes mediated by zinc-finger nucleases.** *Proc Natl Acad Sci USA* 2010, **107:**15022–15026.
47. Venken KJT, Schulze KL, Haelterman NA, Pan H, He Y, Evans-Holm M, Carlson JW, Levis RW, Spradling AC, Hoskins RA, Bellen HJ: **MiMIC: a highly versatile transposon insertion resource for engineering Drosophila melanogaster genes.** *Nat Methods* 2011, **8:**737–743.
48. Smih F, Rouet P, Romanienko PJ, Jasin M: **Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells.** *Nucleic Acids Res* 1995, **23:**5012–5019.
49. Groner AC, Meylan S, Ciuffi A, Zangger N, Ambrosini G, Dénervaud N, Bucher P, Trono D: **KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading.** *PLoS Genet* 2010, **6:**e1000869.
50. Chung JH, Whiteley M, Felsenfeld G: **A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila.** *Cell* 1993, **74:**505–514.
51. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, *et al:* **Ensembl 2012.** *Nucleic Acids Res* 2012, **40:**D84–90.
52. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2011.** *Nucleic Acids Res* 2010, **39:**D876–D882.
53. Ivics Z, Hackett PB, Plasterk RH, Izsvák Z: **Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells.** *Cell* 1997, **91:**501–510.
54. Kent WJ: **BLAT–-The BLAST-Like Alignment Tool.** *Genome Res* 2002, **12:**656–664.
55. Kawakami K, Abe G, Asada T, Asakawa K, Fukuda R, Ito A, Lal P, Mouri N, Muto A, Suster ML, Takakubo H, Urasaki A, Wada H, Yoshida M: **zTrap: zebrafish gene trap and enhancer trap database.** *BMC Dev Biol* 2010, **10:**105.
56. Balciunas D, Davidson AE, Sivasubbu S, Hermanson SB, Welle Z, Ekker SC: **Enhancer trapping in zebrafish using the Sleeping Beauty transposon.** *BMC Genomics* 2004, **5:**62.
57. Choo BGH, Kondrichin I, Parinov S, Emelyanov A, Go W, Toh W-C, Korzh V: **Zebrafish transgenic Enhancer TRAP line database (ZETRAP).** *BMC Dev Biol* 2006, **6:**5.
58. Ellingsen S, Laplante MA, König M, Kikuta H, Furmanek T, Hoivik EA, Becker TS: **Large-scale enhancer detection in the zebrafish genome.** *Development* 2005, **132:**3799–3811.
59. Allen ND, Cran DG, Barton SC, Hettle S, Reik W, Surani MA: **Transgenes as probes for active chromosomal domains in mouse development.** *Nature* 1988, **333:**852–855.
60. Gossler A, Joyner AL, Rossant J, Skarnes WC: **Mouse embryonic stem cells and reporter constructs to detect developmentally regulated genes.** *Science* 1989, **244:**463–465.
61. Ding S, Wu X, Li G, Han M, Zhuang Y, Xu T: **Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice.** *Cell* 2005, **122:**473–483.
62. Horie K, Yusa K, Yae K, Odajima J, Fischer SEJ, Keng VW, Hayakawa T, Mizuno S, Kondoh G, Ijiri T, Matsuda Y, Plasterk RHA, Takeda J: **Characterization of Sleeping Beauty transposition and its application to genetic screening in mice.** *Mol Cell Biol* 2003, **23:**9189–9207.
63. Mátés L, Chuah MKL, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, Grzela DP, Schmitt A, Becker K, Matrai J, Ma L, Samara-Kuko E, Gysemans C, Pryputniewicz D, Miskey C, Fletcher B, Vandendriessche T, Ivics Z, Izsvák Z: **Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates.** *Nat Genet* 2009, **41:**753–761.
64. Kelsch W, Stolfi A, Lois C: **Genetic labeling of neuronal subsets through enhancer trapping in mice.** *PLoS One* 2012, **7:**e38593.
65. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:**499–502.
66. Kikuta H, Laplante M, Navratilova P, Komisarczuk AZ, Engström PG, Fredman D, Akalin A, Caccamo M, Sealy I, Howe K, Ghislain J, Pezeron G, Mourrain P, Ellingsen S, Oates AC, Thisse C, Thisse B, Foucher I, Adolf B, Geling A, Lenhard B, Becker TS: **Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates.** *Genome Res* 2007, **17:**545–555.

67. Engström PG, Fredman D, Lenhard B: **Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes.** *Genome Biol* 2008, **9:**R34.
68. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485:**376–380.
69. Nora EP, Lajoie BR, Schulz EG, Giorgetti L, Okamoto I, Servant N, Piolot T, van Berkum NL, Meisig J, Sedat J, Gribnau J, Barillot E, Blüthgen N, Dekker J, Heard E: **Spatial partitioning of the regulatory landscape of the X-inactivation centre.** *Nature* 2012, **485:**381–385.
70. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV, Ren B: **A map of the cis-regulatory sequences in the mouse genome.** *Nature* 2012, **488:**116–120.
71. Kim E, Kim S, Kim DH, Choi B-S, Choi I-Y, Kim J-S: **Precision genome engineering with programmable DNA-nicking enzymes.** *Genome Res* 2012, **22:**1327–1333.
72. Wang J, Friedman G, Doyon Y, Wang NS, Li CJ, Miller JC, Hua KL, Yan JJ, Babiarz JE, Gregory PD, Holmes MC: **Targeted gene addition to a predetermined site in the human genome using a ZFN-based nicking enzyme.** *Genome Res* 2012, **22:**1316–1326.

doi:10.1186/1471-2164-14-215
Cite this article as: Chen *et al.*: TRACER: a resource to study the regulatory architecture of the mouse genome. *BMC Genomics* 2013 **14**:215.

nature genetics

Large-scale analysis of the regulatory architecture of the mouse genome with a transposon-associated sensor

Sandra Ruf[1], Orsolya Symmons[1], Veli Vural Uslu[1], Dirk Dolle[2], Chloé Hot[1], Laurence Ettwiller[2] & François Spitz[1]

We present here a Sleeping Beauty–based transposition system that offers a simple and efficient way to investigate the regulatory architecture of mammalian chromosomes *in vivo*. With this system, we generated several hundred mice and embryos, each with a regulatory sensor inserted at a random genomic position. This large sampling of the genome revealed the widespread presence of long-range regulatory activities along chromosomes, forming overlapping blocks with distinct tissue-specific expression potentials. The presence of tissue-restricted regulatory activities around genes with widespread expression patterns challenges the gene-centric view of genome regulation and suggests that most genes are modulated in a tissue-specific manner. The local hopping property of Sleeping Beauty provides a dynamic approach to map these regulatory domains at high resolution and, combined with *Cre*-mediated recombination, allows for the determination of their functions by engineering mice with specific chromosomal rearrangements.

Recent findings highlight that critical genetic information, including features commonly associated with the control of gene transcription, can be localized far from genes[1,2]. In particular, developmental genes are known to rely on enhancers localized hundreds of kb away from their promoter, sometimes inside or beyond unrelated adjacent genes[3–6]. Furthermore, genome-wide association studies have suggested the existence of many regulatory variants that could influence the expression of flanking but distant genes[7–9]. Computational and experimental strategies[10–12] have recently progressed in identifying key regulatory elements, primarily enhancers. In some cases, mutations or deletions of such enhancers have been found in individuals with different diseases[13–17]. Interestingly however, several traits or diseases are not caused by changes affecting an enhancer directly but by modifications of the surrounding genomic context[18–21]. These examples suggest that, within a locus, besides their mere presence, the relative position of the different regulatory elements could contribute to their activity and specificity. To understand how genome organization influences gene expression, we need to not only identify individual regulatory elements but also determine their range of action and specificity toward surrounding genes. Understanding the basis of this regulatory architecture will be essential in comprehending the phenotypic consequences of the widespread structural variation present in the human genome[22,23].

The mouse has proven an extremely useful model system to investigate long-range gene regulation or human aneuploidies thanks to chromosomal engineering in embryonic stem cells[24–27]. However, the effort and time needed to transform an embryonic stem cell into a live animal are restrictive. Here we present a simple and efficient strategy to explore the regulatory genome at a large scale without the need for sophisticated manipulations of embryonic stem cells. GROMIT (Genome Regulatory Organization Mapping with Integrated Transposons) relies on *in vivo* controlled mobilization of a Sleeping Beauty transposon. It distributes a regulatory sensor throughout the mouse genome and reveals the regulatory activities associated *in vivo* with each integration site. Using GROMIT, we generated several hundred mouse lines with the regulatory sensor integrated randomly in the genome. We used a subset of 165 representative insertions to survey the regulatory potential and architecture of the mouse genome in an unbiased, non–gene-based manner. At most insertion sites, the reporter sensor showed highly tissue-specific activity, revealing the intrinsic pervasive presence of regulatory influences throughout the genome. Importantly, each transposon can be remobilized to produce mice with new insertions around a selected starting point, as Sleeping Beauty often transposes locally. This property makes GROMIT a highly efficient system to define the regulatory architecture of the genome at high resolution. Combined with *in vivo* recombineering strategies[28–30], this approach offers multiple methods to create animals with segmental aneuploidies or other structural variations.

RESULTS

GROMIT, a transposon-mediated regulatory sensor system

We constructed a 'regulatory sensor' (SBlac) consisting of a *LacZ* reporter gene driven by the promoter region of the human *β-globin* gene (**Fig. 1**). This 50-bp promoter fragment has no specific activity on its own, but it responds faithfully to the activity of enhancers cloned next to it[4,31]. Thus, SBlac activity should reflect the influence of endogenous regulatory elements, thus revealing the regulatory potential existing at its insertion point. To easily

[1]Developmental Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. [2]Centre for Organismal Studies, Heidelberg University, Heidelberg, Germany. Correspondence should be addressed to F.S. (spitz@embl.de).

Received 17 November 2010; accepted 16 February 2011; published online 20 March 2011; doi:10.1038/ng.790

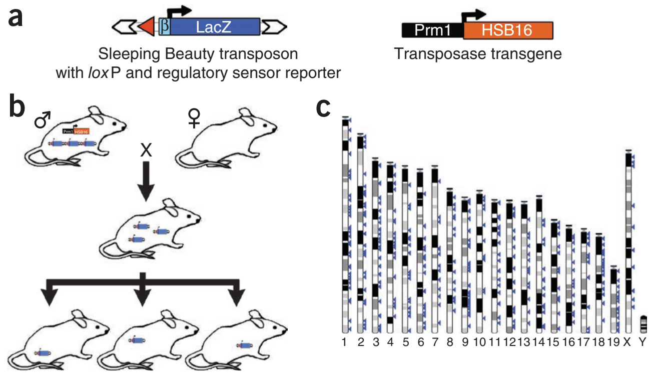

a
β
LacZ
Sleeping Beauty transposon
with loxP and regulatory sensor reporter
Prm1
HSB16
Transposase transgene
b
♂
♀
X
c
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y

Figure 1 The GROMIT strategy. (**a**) Schematic representation of the transposase-expressing transgene and the regulatory sensor. The coding sequence of the improved HSB16 Sleeping Beauty transposase[35] is under the control of the mouse *Prm1* promoter region, which is specifically active in the male germline[36]. The SBlac transposon contains a *LacZ* reporter gene under the control of the minimal promoter of the human β-globin gene (β)[31] and a *lox*P site (red triangle), cloned within the inverted or direct repeats of the Sleeping Beauty transposon (white arrowheads)[35]. (**b**) The breeding scheme. We mated males transgenic for both the transposase and transposon transgenes (seed males) to wild-type females to produce F0 animals with new insertions. We performed additional breeding to segregate the different insertions and establish lines with single integrants. (**c**) Distribution of 550 mapped SBlac insertions (blue triangles) on the mouse genome.

distribute SBlac throughout the mouse genome, we cloned it within a Sleeping Beauty transposon[32]. This transposon has been shown to be active in mammals where it transposes in a 'cut and paste' manner, leaving a minimal footprint behind[33,34]. We generated transgenic mouse lines containing several copies of the SBlac transposon inserted as a concatemer (**Supplementary Note**). We also produced transgenic lines expressing a hyperactive form of the Sleeping Beauty transposase (HSB16)[35] under the control of a mouse *Prm1* promoter fragment, the activity of which is restricted to haploid spermatids[36] (**Fig. 1**). This *Prm1*∷HSB16 transgene should trigger remobilization of the transposon in the spermatids of males harboring both the transposon and transposase constructs (so-called 'seed males'). To assess the efficiency of the approach, we crossed seed males with wild-type females and analyzed their progeny (**Fig. 1b**). We identified insertions of the transposon at new genomic positions from all combinations of transposon and transposase lines tested. The transposition frequency was very high (from one to six remobilization events in each transgenic offspring of seed males with 8 or 20 copies of the transposon). Consequently, additional breeding was often necessary to segregate the different insertions and to establish mouse lines with a single insertion. Importantly, in F0 mice, the new insertions systematically co-segregated with the starting concatemer, showing that the transposase was only active after meiosis and was not active in somatic tissues (**Supplementary Note**). Thus, F0 animals were not mosaic for the insertions they carried.

A collection of mouse lines with single-copy insertions

We determined the insertion sites of the transposon by asymmetric PCR on DNA from F0 animals using a transposon-specific primer and a partially random primer. We obtained flanking sequences for 569 insertion sites and mapped 550 of these to unique loci in the mouse genome (**Fig. 1c**; see URLs for the Transposon and Recombinase-Associated Chromosomal Engineering Resource (TRACER) database). The vast majority of these insertions contained a single transposon, but we identified rare instances of more complex events (**Supplementary Note**).

As previously described by others[33,34], we found that many insertions clustered around their starting site, reflecting Sleeping Beauty's tendency for local hopping (**Supplementary Fig. 1**). For subsequent analyses, we excluded these 'local' insertions (within ~2 Mb of the starting sites) because they biased the overall distribution of the insertions, could not be easily segregated from the starting sites and might be associated with chromosomal rearrangements caused by repeated transposition from Sleeping Beauty concatemers[37]. The remaining insertions—more than half of the original set—are distributed across all chromosomes and follow a random distribution with respect to genes or transcriptional start sites (TSS) (**Fig. 2a**,**b** and **Supplementary Fig. 2**). A subset of 165 insertions ('expression tested') was further analyzed for expression of the reporter gene. These insertions (124 from established mouse lines and 41 from embryos with single insertions)

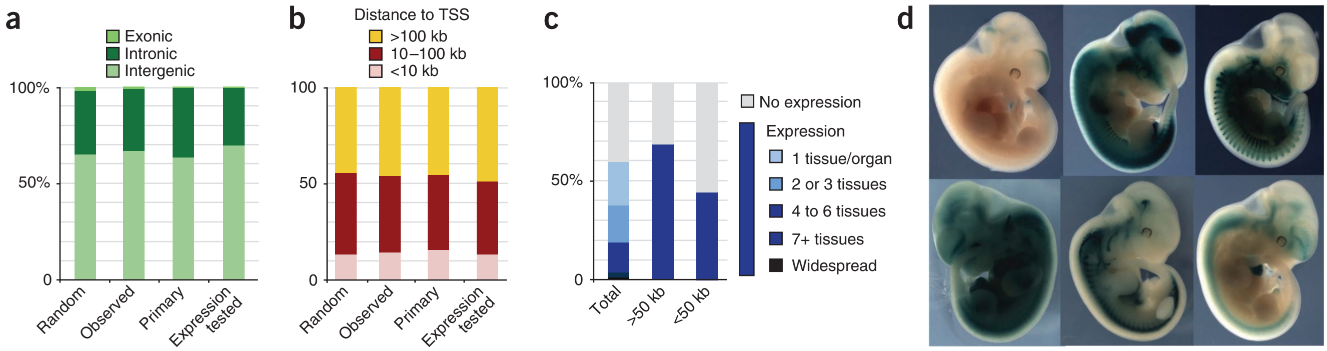


Figure 2 Genomic distribution and transcriptional activity of the different insertions. (**a**) Fraction of insertions within exons, introns or intergenic intervals (based on RefSeq genes). (**b**) Proportion of insertions at different distances from TSS. The observed dataset contains all insertions mapped to single loci, excluding the ones corresponding to the local hotspots associated with the starting concatemers A and C or subsequent local remobilizations. We analyzed 165 of these insertions for reporter activity, forming the 'expression tested' dataset. We obtained random distributions by analyzing independent randomizations with the same sample size as in the observed dataset. (**c**) Expression status of the insertions for all expression-tested insertions or for different subsets according to their distance relative to the nearest TSS. Insertions within 50 kb of a TSS in E11.5 embryos showed expression less frequently than insertions farther away from a TSS ($P = 0.00739$, significance level 0.025), but the complexity of the corresponding patterns was similar with mostly tissue-restricted expression (data not shown). (**d**) Examples of LacZ expression from different insertions in E11.5 embryos.

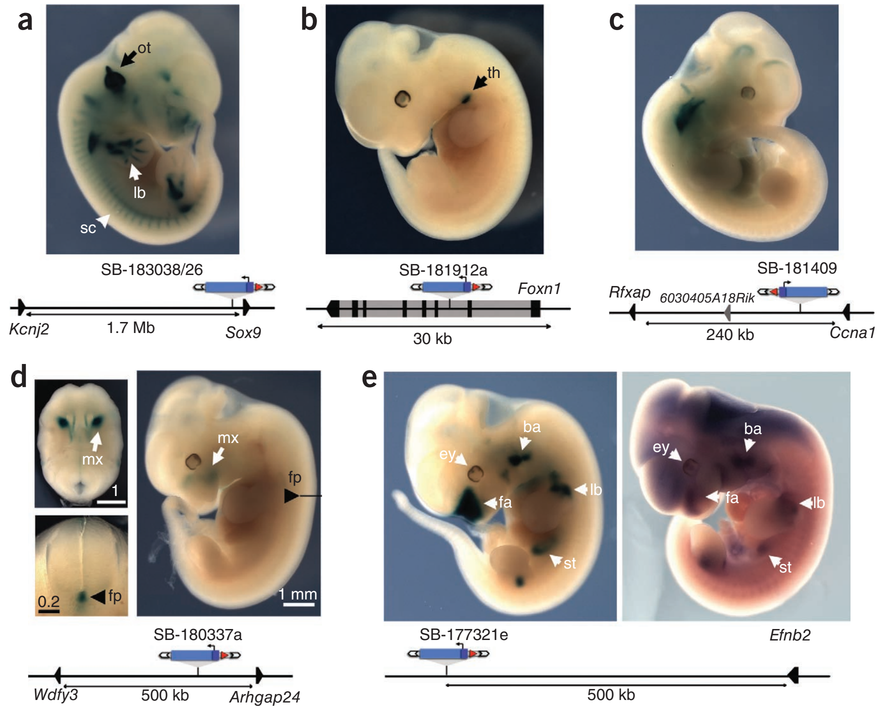


Figure 3 Examples of patterns of activities associated with different insertions. (**a**–**e**) We performed LacZ staining on E11.5 embryos heterozygous for the different insertions of the regulatory sensor. For each insertion, a schematic representation of the locus is shown. Genes are represented by black arrowheads, pointing in the direction of their transcription except for in **b**, where black bars correspond to the different exons and the gray block corresponds to the gene body. (**a**) A transposon 80 kb from *Sox9* showed strong LacZ staining in known *Sox9* expression domains[47] in the limb condensing mesenchyme (lb), the sclerotome (sc) and the developing ear (ot). However, we detected no expression in the neural tissues, where this gene is also strongly expressed[47]. (**b**) *LacZ* expression in the third pharyngal pouch (black arrow, th) of a transposon inserted into the gene body of *Foxn1* mimics endogenous gene expression in the developing thymus primordium[48]. (**c**,**d**) LacZ staining of insertions that are not in the vicinity of developmental regulators. Insets in **d** are sectioned embryos showing expression in the floor plate of the neural tube (arrowhead, fp) and in the maxillary (arrow, mx). Scale bars are indicated in mm. (**e**) Comparison of the transposon (left) and endogenous gene (right) expression in E11.5 embryos revealed shared domains (ey, anterior-most part of the eye; ba, second branchial arch; fa, face; lb, proximal limb; st, stomach) between *Efnb2* and the SB-177321e insertion about 500 kb away. However, strong expression of *Efnb2* in the brain and developing vascular system was not recapitulated by the transposon.

are representative of the diverse situations found in the mouse genome, covering both gene deserts and gene-dense regions.

Pervasive but highly tissue-specific transcriptional regulation

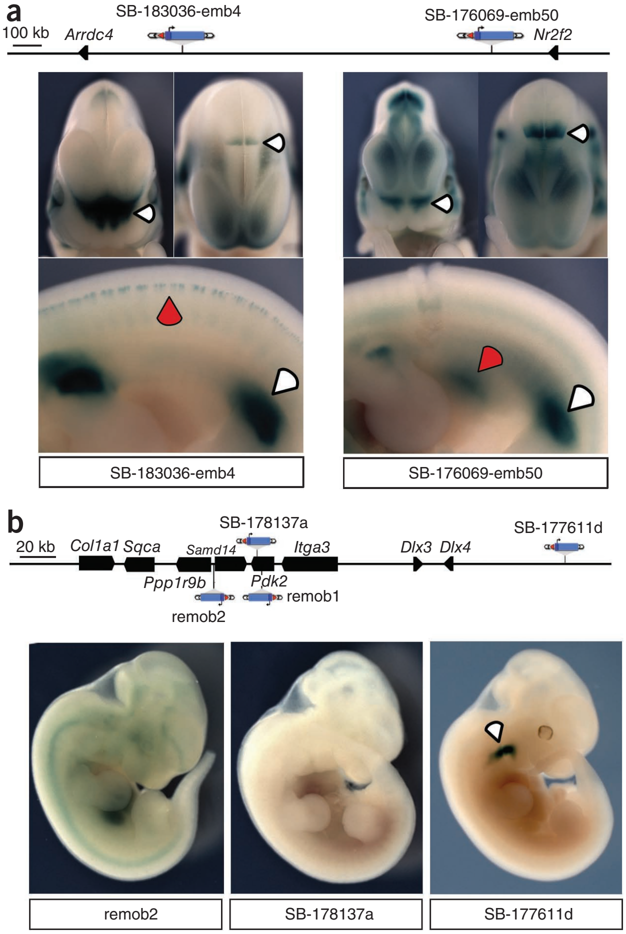
For each of these insertions, we assessed the activity of the reporter gene associated with the SBlac transposon in whole-mount embryonic day (E) 11.5 mouse embryos. We did not detect any reproducible pattern shared by a majority of insertions, confirming the absence of autonomous activity of our reporter gene. The expression patterns observed were independent of the activity of the transposon at its initial position (**Supplementary Fig. 3**), were reproducible between littermates and were stable across multiple generations (only two lines showed variegated expression, one of which was on the X chromosome).

Remarkably, 98 of 165 insertions (~60%) showed tissue-specific expression of the reporter gene at this embryonic stage (**Fig. 2c** and see URLs for the TRACER database). These patterns were very diverse, and some insertions showed almost ubiquitous expression (**Fig. 2d**, lower left), illustrating that the reporter sensor can be expressed in any tissue. However, almost all expression patterns (96 of 98) were restricted to a few organs or embryonic territories. Notably, this predominance of highly tissue-specific activities is in sharp contrast with previous analyses of endogenous gene activity that concluded that most genes (30–70%) had widespread expression at similar embryonic stages[38,39] (**Supplementary Note**).

We found multiple cases where *LacZ* expression of the transposon matched the expression pattern of a neighboring gene (**Figs. 3**–**6** and **Supplementary Figs. 4**–**6**), showing that most observed patterns correspond to biologically meaningful activities. Very frequently we found co-expression with an endogenous gene or another insertion located hundreds to thousands of kb away, highlighting that long-range regulation is a very widespread phenomenon. Such situations were common around known key developmental genes but were also observed frequently in regions without any such kind of gene nearby (**Fig. 3c**,**d** and see URLs for the TRACER database). In many instances, however, the regulatory sensor only recapitulated part of the expression of the flanking genes (**Fig. 3a**–**e**, **Fig. 4b** and **Supplementary Fig. 6**). Importantly, transposons inserted within the same locus could have specific expression domains in addition to shared ones (**Fig. 4** and **Supplementary Figs. 4**,**5**). Thus, depending on its genomic position, the transposon can reveal different subsets of the expression domains of the flanking gene(s). These observations suggest that the differences observed between insertions and flanking genes reflect the distinct range of action of the surrounding regulatory elements and not an intrinsic inability of the regulatory sensor to respond to some regulatory inputs. Notably, the insertion of the transposon in a locus did not modify the expression level of the neighboring genes with which it shared expression patterns (**Supplementary Fig. 7**), indicating that it is acting more as a sensor than as an enhancer trap.

The regulatory potential of the genome, as shown by SBlac activity, did not appear to be linked specifically to any genomic feature. Transposons inserted within genes showed *LacZ* expression in similar proportions as intergenic ones (30 out of 50 for intragenic insertions; 68 out of 115 for intergenic ones; χ^2 tests gave P values > 0.49 for any significance level α below 0.1). However, and maybe counterintuitively, transposons away from TSS showed expression more frequently than the ones located more closely (**Fig. 2c**). This difference was also significant after removing intragenic insertions ($P = 0.034$ for $\alpha = 0.025$), ruling out transcriptional interference as its major cause. This bias could be due to a variety of factors which could either reduce the chance of expression (for example, competition with endogenous genes) or increase silencing (for example, spreading of repressive chromatin around silent genes[40]). However, it is not an absolute rule, and several insertions close to TSS showed expression patterns similar to the neighboring gene.

Figure 4 Short- and long-range activities detected by transposons inserted into the same loci. (**a**) We obtained two insertions (SB-183036-emb4 and SB-176069-emb50) in the 2-Mb gene desert between *Arrdc4* and *Nr2f2*. These insertions shared expression domains (white arrows) in the face and midbrain and showed insertion-specific expression (red arrows), illustrating the existence of overlapping but distinct regulatory landscapes with their own tissue specificity. (**b**) Insertions in the *Col1a1-Dlx4* interval showed different activities. The insertion 70 kb upstream of *Dlx4* (SB-177611d) showed LacZ activity in the visceral arches (arrow), partially mirroring *Dlx3* and *Dlx4* expression, but showed only weak staining in the limb apical ectodermal ridge and no staining in the mandibular arch, compared to the *Dlx3*, and *Dlx4* endogenous genes[49]. An insertion in an intron of *Pdk2* (SB-178137a) showed no expression. Upon remobilization of this one insertion, we discovered two new insertions in the same neighborhood. A very local insertion (remob1, 0.6 kb away) showed also no expression, whereas another (remob2, 17 kb away) was expressed more broadly, notably in the neural tube and brain, matching the flanking genes' activities.



Altogether, GROMIT revealed that regulatory activities are not focused toward gene promoter regions, but that chromosomes are covered with dense 'regulatory jungles'. These domains are mostly associated with tissue-restricted and patterned activities and are particularly present away from genes, whereas their influence appears to be attenuated or more controlled in the vicinity of promoter regions.

Enhancers and repressors define regulatory domains

To compare the regulatory potential of a locus with the intrinsic and autonomous activity of neighboring enhancers, we took advantage of the Vista Enhancer Database. This database comprises about 1,200 genomic elements, many of them showing tissue-specific enhancer activity in E11.5 transgenic mouse embryos[2]. As most of these elements show extreme evolutionary conservation, it is reasonable to expect that the mouse sequences have the same activity as the tested human orthologs. Transposons inserted in the vicinity of a characterized Vista enhancer usually have expression patterns corresponding to the contribution of these adjacent elements (**Fig. 5** and **Supplementary Figs. 6,8**). However, in each case, we found that the expression of the transposon was more restricted than the intrinsic activity of the neighboring enhancers, although it nevertheless recapitulated the precise expression boundaries of the endogenous gene situated several hundred kb away. These recurrent differences seen between the activity of enhancers, determined outside of their normal genomic context, and the expression of endogenous genes or transposons inserted nearby emphasize that the overall regulatory activities associated with a locus are not defined by the mere addition of enhancers but rather are further refined by the action of distinct highly tissue-specific repressor elements.

Fine mapping of genomic regulatory landscapes

So far, a major limitation with Sleeping Beauty has been the difficulty to remobilize single transposons from their new positions[41,42].

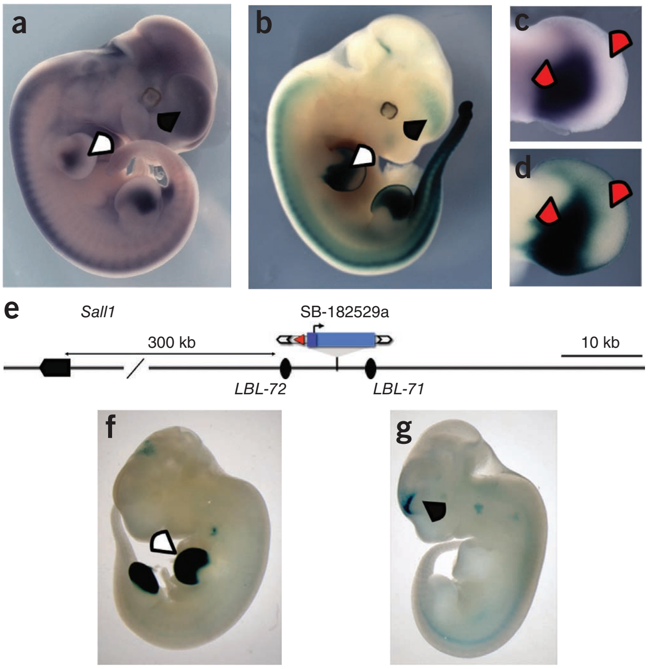


Figure 5 Comparison between the genomic regulatory potential and intrinsic activities of nearby enhancers. (**a**–**g**) An insertion 300 kb upstream of *Sall1* is localized between two evolutionarily conserved enhancer elements as schematized in **e**. The reporter gene (**b**) recapitulates most expression domains of the endogenous gene detected by *in situ* hybridization (**a**). Expression in the limbs (white arrowhead) and forebrain (black arrowhead) overlaps with activity determined for the two flanking Vista enhancers mapped in the region (LBL-72 (**f**) and LBL-71 (**g**)). However, although LBL-72 can drive expression of a reporter gene throughout the whole autopod in a transgenic assay (**f**), in the endogenous context, this activity is silenced both in the most distal mesenchyme and in a proximal anterior region of the autopod (**c**,**d**, red arrows). For simplicity, additional enhancers with activities corresponding to other *Sall1* and SB-182529a expression domains, which had been previously identified[2] but which are localized further away from SB-182529a, are not shown. Photos of the transgenic embryos for Vista enhancers (**f**,**g**) were taken from the Vista Enhancer Browser[2].

Table 1 Remobilization frequency of single-copy transposon insertions

Donor								
Insertion	Type	Chr.	F1	SBlac	Remob[b]	Mapped	Intrachrom[c]	Local[d]
177184c	Het	5	71	25	9	9	2	1
183038	Het	7	52	33	7	4	0	0
183036	Het	1	237	90	44	37	12	5
176599b	Het	2	104	46	11	10	4	3
176599bc[a]	Het	1, 2	53	41	19	17	7	2
176599b	Hom	2	61	57	12	11	2	1
176148b	Hom	16	24	22	7	3	0	0
178235	Hom	16	15	15	3	3	0	0
183041a	Het	3	261	106	41	24	5	4
178137a	Het	11	145	60	11	8	2	2
Total			**1,023**	**495**	**164 (33%)**	**126**	**34 (27%)**	**18 (14%)**

Chr., chromosome; het, heterozygous; hom, homozygous; remob, remobilized; intrachrom, intrachromosomal.
[a]Male with two unlinked insertions (176599b and 176599c). The number of remobilizations is an underestimate as only F1 individuals, which were negative for both 176599b and 176599c, were considered as 'remobilized'. [b]Number of animals with a SBlac transposon in the genome but not at the starting position. [c]New insertions on the same chromosome than the donor site. [d]New insertions within 2 Mb of the position of the donor site.

To evaluate the efficiency of our system, we reintroduced the transposase transgene into different mouse lines carrying single-copy transposons. In all cases, remobilization was extremely efficient, and 17% to 50% of the transgenic F1 mice had the transposon inserted in a new location (**Table 1**). Six out of one hundred offspring from seed males homozygous for the transposon did not inherit the transposon, implying that only three-quarters of the excised transposons reinsert in the genome (accordingly, we saw a deficit of transmission of the transposon (401 out of 923 offspring) from heterozygous seed males). Importantly, more than one-fourth of the remobilization events from these insertions were intrachromosomal and about 15% occurred within 2 Mb of the initial position of the transposon. Such a high rate of local transposition enables systematic exploration of regions of interest.

As a proof of principle, we selected a few insertions and remobilized them to more finely map the surrounding regulatory domains. Insertions obtained within a given locus often shared similar expression profiles, reflecting their association with long-range enhancers acting over several hundred kb (**Fig. 4a** and **Fig. 6c**–**e**). But in other cases, insertions just a few kb apart had different expression features, showing the existence of rather abrupt regulatory transitions both within gene-dense regions (**Fig. 4b**) and in gene deserts (**Supplementary Figs. 4,5**). Many transitions were tissue-specific and involved only some of the regulatory influences detected in the region. These examples illustrate the potential of GROMIT to functionally identify regions that could correspond to insulator elements and restrict the range of action of regulatory enhancers *in vivo*.

Furthermore, because the SBlac transposon contains a *lox*P site, it is possible to generate a series of overlapping rearrangements over a region of interest by combining *in vivo* transposition from one insertion, as described above, with *in vivo* recombineering approaches[28–30]. This could be achieved simply by breeding and, given the efficiency of the tranposition and recombination, only a few cages. As an example, we used this approach to investigate a large gene desert starting from an insertion located 1.75 Mb away from *Myc* (**Fig. 6** and **Supplementary Fig. 9**). In the first two litters screened for remobilization, we identified one animal with a local event, carrying the transposon 0.85 Mb from *Myc*. *Trans*-allelic *Cre*-mediated recombination between the *lox*P sites associated with the two insertions produced embryos carrying either deletion or duplication of the 0.9-Mb intervening region (seven deletions and four duplications out of 107 embryos) together with the regulatory sensor at the recombination point. Notably, whereas both insertions were specifically expressed in the somites and in the face (corresponding to regions where *Myc* is expressed more strongly), these domains were lost or greatly reduced after deletion of the intervening region, suggesting that it contained the corresponding enhancer(s) (**Fig. 6**).

Altogether, these examples demonstrate that GROMIT is a simple and efficient strategy to explore the regulatory architecture of the genome. Coupled with Cre-*loxP* recombination, it enables the identification of potentially important features (insulators and enhancers) as well as the creation of mouse lines with corresponding targeted chromosomal rearrangements.

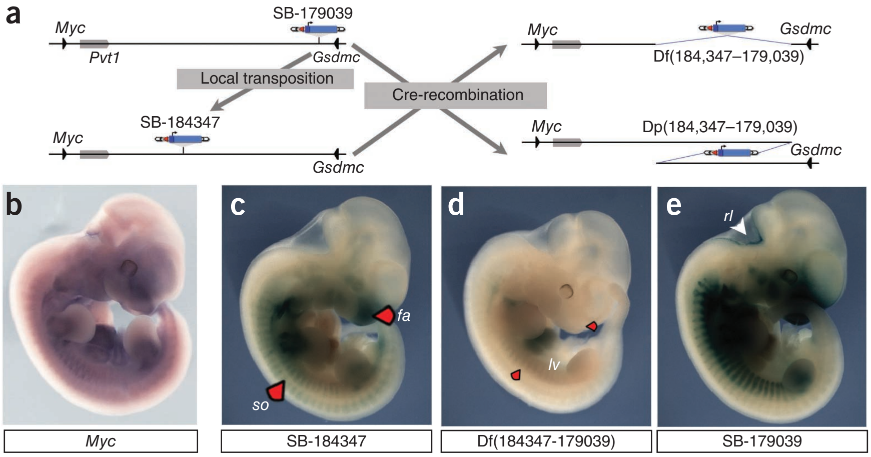


Figure 6 Mapping genomic regulatory domains with sequential tranposition and/or recombination. (**a**) Schematic representation of the *Myc*-*Gsdmc* interval, including the non-coding *Pvt1* gene and the initial insertion from the SB-179039 line. Upon local transposition from SB-179039, we obtained SB-184347, located ~890 kb away from SB-179039. *Cre*-mediated recombination in *trans* between the *lox*P sites produced chromosomes with the deletion or the duplication of the intervening region. The breeding strategy is detailed in **Supplementary Figure 9**. (**b**–**e**) *In situ* hybridization of the endogenous *Myc* gene showed enrichment of expression notably in the face (fa) and somites (so) coinciding with LacZ expression domains observed both with SB-184347 (**c**) and SB-179039 (**e**) in E11.5 embryos. In SB-179039, LacZ expression was stronger than in SB-184347 and included a specific rhombic lip domain (rl) not observed for *Myc* or SB-184347. Both domains were lost when the 890-kb region was deleted (**d**), whereas the duplication led only to quantitative differences when compared to SB-179039 (data not shown).

DISCUSSION

The tools we developed for GROMIT allow remobilization of a single-copy transposon with a high frequency, one to two orders of magnitude higher than what has been reported previously[41,42]. The use of a transposase source expressed only transiently in late spermatogenesis could enhance transposition efficiency directly by virtue of the intense chromatin remodeling that occurs in spermatids and indirectly by preserving remobilization-prone insertions that would be lost with a constitutively expressed transposase. The efficiency and the simplicity of the system represent an important advance that will enable the functional exploration of the mouse genome in many ways. This system enables the production of mouse lines with single and non-mosaic integration sites. It minimizes the risk of rearrangements and other events associated with remobilization from multicopy concatemers[37]. It also opens the possibility of reusing insertions obtained previously and, thanks to the local hopping of Sleeping Beauty, allows systematic analyses of regions of interest, a feature that is not possible with non-locally biased transposon systems such as Piggybac. In this aspect, GROMIT is similar to the recently described LHED (Local Hopping Enhancer Detector) strategy[24]. LHED is based on transposition in embryonic stem cells, which could have advantages. However, given the time, effort and expertise needed to make a mouse from embryonic stem cells, the efficient *in vivo* system we propose is much more cost and time effective, particularly for medium- to large-scale approaches. Furthermore, whereas the LHED transposon is limited to creating nested deletions or inversions with a fixed breakpoint, GROMIT is much more flexible, as the different insertions can be combined independently to generate a series of overlapping chromosomal rearrangements.

Here we used GROMIT to characterize the distribution of the regulatory information present in the mouse genome. The underlying principle is somehow similar to 'enhancer-trap' screens, but it differs from them in essence. GROMIT does not favor expression-competent positions as compared to other approaches that pre-select insertions either directly for expression or indirectly through a built-in selection marker[43,44]. Also, in GROMIT, the reporter gene is driven by a minimal neutral promoter, which is not restricted to specific tissues but which readily interacts with available regulatory elements, provided it is within their range of action. Notably, the strength and reproducibility of *LacZ* expression and the frequent overlap with the expression patterns of flanking genes shows that the transposon reports genuine and biologically meaningful activities and not just transcriptional noise. However, the transposon does not seem to 'trap' enhancers, as its effect on endogenous genes appears to be very limited. Its weak promoter may miss activities that are short range or very tightly associated with a specific gene, but it is otherwise finely regulated within the context of its genomic insertion site and has minimal impact on it, acting therefore as a regulatory sensor. Hence, the expression pattern of this sensor reflects the normal regulatory organization of a locus (integrating regulatory elements according to their range of action and the strength of their specific interactions) and not the artificial situation created by true enhancer-trap constructs.

The large number of lines produced and analyzed with GROMIT provides an unprecedented view of the regulatory organization of the mouse genome. The frequency and the diversity of the activities detected by the regulatory sensor show that a very large part of the genome—and not only in the vicinity of key developmental regulators—is associated with tissue-specific and spatially restricted expression potential. This pervasive presence of regulatory influences with high tissue specificity along mammalian chromosomes contrasts with the rather widespread expression of most genes in mid-gestation embryos[38,39]. Consistently, previous reports have also suggested that regions harboring chromatin marks associated with enhancer activity are mostly cell-type specific[12], and most if not all experimentally tested enhancers show a narrow pattern of activity[2]. Thus, it is likely that most of gene basal and widespread expression is achieved through proximal promoter elements. As discussed above, even though the sensor could capture regulatory input in any tissue, this eventuality depends on the range of action of the responsible element. The paucity of insertions showing widespread expression of SBlac suggests that promoter regions associated with widespread gene activity have very short-range action or are more tightly directed to the neighboring TSS. To report on their influence, the transposon would have to be inserted close-by and to compete for them with the endogenous gene. These two parameters may explain the under-representation of ubiquitous-like activities in the patterns shown by the regulatory sensor and the lower expression frequency of transposons inserted in the vicinity of promoters.

On the other hand, the general presence of activities captured by our transposon system throughout the genome emphasizes that, besides gene promoters, a multitude of other elements contribute to gene regulation as well without being intrinsically bound to a given gene but which potentially act on any gene present within their range of action. In particular, gene deserts appear to be regulatory jungles filled with broadly distributed tissue-specific expression potential. The expression differences between neighboring insertions or with endogenous genes expose the presence of specific regulatory boundaries. These differences could correspond to insulators or reflect that the associated regulatory modules are positioned at different locations and have different ranges of action. Altogether, this reveals the subdivision of the genome into distinct but overlapping regulatory domains with their own specific activities independent of genes. The elements defining these domains seem intrinsically promiscuous: they act in a relatively flexible manner over large distances rather than being intrinsically bound to activate specific endogenous target genes.

For genes already broadly expressed, these highly tissue-specific regulatory inputs may have no major impact. They are, however, in agreement with the predominant presence of cell-type–specific expression quantitative trait loci away from genes[45] and could explain why genome-wide association studies repeatedly point to non-genic intervals[9]. Fine tuning of gene activity by remote tissue-specific modulators could be more common than usually considered and could represent the result of an extensive evolutionary regulatory tinkering to adjust basic biological functions to developmental processes. Nonetheless, some of the activities detected by our regulatory sensor may not be relevant for any endogenous gene, notably the ones found in large regions with little evolutionary conservation or recent expansion. These activities could correspond to cryptic and evolutionary labile enhancers, the existence of which would help the rapid evolutionary turnover of gene regulatory elements recently proposed[46].

Our survey also highlights the contribution of tissue-specific repressive elements. Notably, their action is not limited to preventing the inappropriate activation of a gene by unrelated neighboring enhancers or to silence it in broad domains. Instead, they repress the activity of enhancers in a very precise and controlled manner to further refine the overall regulatory output. Evolving specific gene expression patterns may be simpler by intersecting positive and negative regulatory elements as compared to developing complex modules with highly restricted activities. Such situations could lead to abrupt regulatory changes by unmasking or extending the range of activities already present but normally repressed or hidden from genes. Consequently, the impact of genomic variation should be considered in the holistic context of these dense and complex regulatory interactions.

As shown here, GROMIT provides useful regulatory information about a given region in a gene-independent manner. Such functional information can help make better biological sense of the ever-growing catalog of chromatin profiling and transcription factor binding sites, especially for the ones detected away from genes and for which the relationship to gene expression is rather elusive. The comparison of GROMIT data with other genomic information (chromatin and transcription factor profiling, and chromosome conformation capture) will also provide new and important insights regarding the functional organization of the genome. Importantly, with ongoing remobilization projects, the number and density of insertions should grow steadily, providing information about the regulatory architecture of more regions with increased resolution.

Furthermore, with increasing numbers of genome-wide association studies pointing toward the importance of non-genic intervals, we need to develop robust strategies to evaluate the role of these regions and to understand the impact of structural variants. As we illustrated here, with local hopping and recombineering, GROMIT offers a simple and efficient alternative to the current embryonic stem cell–based approaches for the generation of models of human aneuploidies and other large structural variants[24,26,27]. The 150 lines already available provide starting sites and access to almost 10% of the mouse genome, and this coverage will expand with the additional random insertions generated as byproducts of each remobilization project. Homologous recombination in embryonic stem cells can also be used to target a (eventually modified) Sleeping Beauty transposon to a precise genomic position, with GROMIT subsequently enabling the generation of nested rearrangements in a time- and cost-effective manner. By virtue of its simplicity and high efficiency, the GROMIT toolkit opens, with the expanding TRACER resource, a new level to functional investigations of the mammalian non-coding genome.

URLs. Transposon and Recombinase-Associated Chromosomal Engineering Resource (TRACER) database, http://tracerdatabase.embl.de; Blat, http://genome.ucsc.edu/cgi-bin/hgBlat; Vista Enhancer Browser, http://enhancer.lbl.gov/frnt_page.shtml.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank all the members of the EMBL mouse and transgenic facilities for pronuclear injections of the transgenes and mouse husbandry, particularly C. Klasen, A. Schultz and S. Feller; B. Fletcher (University of Florida) for the Sleeping Beauty transposon and transposase plasmids; A. Ephrussi and A. Aulehla for discussions and comments on the manuscript; and A. Hermelin and EMBL information technology service for helping with the TRACER website. We are grateful to all the members of the Spitz lab and students of the EMBL PhD classes of 2008 and 2009 who contributed to embryo dissection and staining. O.S. is funded by a Louis-Jeantet Foundation PhD fellowship. This work was supported by funding from the European Molecular Biology Laboratory (to F.S.) and from the European Community's Seventh Framework Programme (FP7/ 2007-2013) under the grant agreement n°FP7-211868 (to L.E. and F.S.).

AUTHOR CONTRIBUTIONS

F.S. conceived and designed the GROMIT strategy. S.R., O.S., V.V.U., C.H. and F.S. performed the experiments. O.S., D.D. and L.E. performed the statistical analyses. All authors discussed the results and contributed to the writing of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. ENCODE Project Consortium. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Pennacchio, L.A. *et al. In vivo* enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).
3. Kleinjan, D.A. *et al.* Long-range downstream enhancers are essential for *Pax6* expression. *Dev. Biol.* **299**, 563–581 (2006).
4. Spitz, F., Gonzalez, F. & Duboule, D. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113**, 405–417 (2003).
5. Nobrega, M.A., Ovcharenko, I., Afzal, V. & Rubin, E.M. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
6. Sagai, T. *et al.* A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. *Development* **136**, 1665–1674 (2009).
7. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).
8. Wang, K. *et al.* Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* **459**, 528–533 (2009).
9. Visel, A., Rubin, E.M. & Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature* **461**, 199–205 (2009).
10. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
11. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
12. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
13. Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41**, 885–890 (2009).
14. Lettice, L.A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
15. Jeong, Y. *et al.* Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat. Genet.* **40**, 1348–1353 (2008).
16. Benko, S. *et al.* Highly conserved non-coding elements on either side of *SOX9* associated with Pierre Robin sequence. *Nat. Genet.* **41**, 359–364 (2009).
17. D'haene, B. *et al.* Disease-causing 7.4 kb *cis*-regulatory deletion disrupting conserved non-coding sequences and their interaction with the FOXL2 promotor: implications for mutation screening. *PLoS Genet.* **5**, e1000522 (2009).
18. Kurth, I. *et al.* Duplications of noncoding elements 5′ of *SOX9* are associated with brachydactyly-anonychia. *Nat. Genet.* **41**, 862–863 (2009).
19. Klopocki, E. *et al.* A microduplication of the long range SHH limb regulator (ZRS) is associated with triphalangeal thumb-polysyndactyly syndrome. *J. Med. Genet.* **45**, 370–375 (2008).
20. Dathe, K. *et al.* Duplications involving a conserved regulatory element downstream of *BMP2* are associated with brachydactyly type A2. *Am. J. Hum. Genet.* **84**, 483–492 (2009).
21. Sun, M. *et al.* Triphalangeal thumb-polysyndactyly syndrome and syndactyly type IV are caused by genomic duplications involving the long-range, limb-specific SHH enhancer. *J. Med. Genet.* **45**, 589–595 (2008).
22. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
23. Kidd, J.M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
24. Kokubu, C. *et al.* A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nat. Genet.* **41**, 946–952 (2009).
25. Kondo, T. & Duboule, D. Breaking colinearity in the mouse HoxD complex. *Cell* **97**, 407–417 (1999).
26. Nakatani, J. *et al.* Abnormal behavior in a chromosome-engineered mouse model for human 15q11–13 duplication seen in autism. *Cell* **137**, 1235–1246 (2009).
27. Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409–412 (2010).
28. Spitz, F., Herkenne, C., Morris, M.A. & Duboule, D. Inversion-induced disruption of the Hoxd cluster leads to the partition of regulatory landscapes. *Nat. Genet.* **37**, 889–893 (2005).
29. Hérault, Y., Rassoulzadegan, M., Cuzin, F. & Duboule, D. Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nat. Genet.* **20**, 381–384 (1998).
30. Wu, S., Ying, G., Wu, Q. & Capecchi, M.R. Toward simpler and faster genome-wide mutagenesis in mice. *Nat. Genet.* **39**, 922–930 (2007).
31. Yee, S.P. & Rigby, P.W. The regulation of myogenin gene expression during the embryonic development of the mouse. *Genes Dev.* **7**, 1277–1289 (1993).

32. Ivics, Z., Hackett, P.B., Plasterk, R.H. & Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **91**, 501–510 (1997).
33. Luo, G., Ivics, Z., Izsvák, Z. & Bradley, A. Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **95**, 10769–10773 (1998).
34. Horie, K. *et al.* Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Mol. Cell. Biol.* **23**, 9189–9207 (2003).
35. Baus, J., Liu, L., Heggestad, A.D., Sanz, S. & Fletcher, B.S. Hyperactive transposase mutants of the Sleeping Beauty transposon. *Mol. Ther.* **12**, 1148–1156 (2005).
36. Peschon, J.J., Behringer, R.R., Palmiter, R.D. & Brinster, R.L. Expression of mouse protamine 1 genes in transgenic mice. *Ann. NY Acad. Sci.* **564**, 186–197 (1989).
37. Geurts, A.M. *et al.* Gene mutations and genomic rearrangements in the mouse as a result of transposon mobilization from chromosomal concatemers. *PLoS Genet.* **2**, e156 (2006).
38. Reymond, A. *et al.* Human chromosome 21 gene expression atlas in the mouse. *Nature* **420**, 582–586 (2002).
39. Wurst, W. *et al.* A large-scale gene-trap screen for insertional mutations in developmentally regulated genes in mice. *Genetics* **139**, 889–899 (1995).
40. Pauler, F.M. *et al.* H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome Res.* **19**, 221–233 (2009).
41. Geurts, A.M. *et al.* Conditional gene expression in the mouse using a Sleeping Beauty gene-trap transposon. *BMC Biotechnol.* **6**, 30 (2006).
42. Yae, K. *et al.* Sleeping Beauty transposon-based phenotypic analysis of mice: lack of Arpc3 results in defective trophoblast outgrowth. *Mol. Cell. Biol.* **26**, 6185–6196 (2006).
43. Korn, R. *et al.* Enhancer trap integrations in mouse embryonic stem cells give rise to staining patterns in chimaeric embryos with a high frequency and detect endogenous genes. *Mech. Dev.* **39**, 95–109 (1992).
44. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
45. Dimas, A.S. *et al.* Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science* **325**, 1246–1250 (2009).
46. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **21**, 1036–1040 (2010).
47. Wunderle, V.M., Critcher, R., Hastie, N., Goodfellow, P.N. & Schedl, A. Deletion of long-range regulatory elements upstream of *SOX9* causes campomelic dysplasia. *Proc. Natl. Acad. Sci. USA* **95**, 10649–10654 (1998).
48. Gordon, J., Bennett, A.R., Blackburn, C.C. & Manley, N.R. *Gcm2* and *Foxn1* mark early parathyroid- and thymus-specific domains in the developing third pharyngeal pouch. *Mech. Dev.* **103**, 141–143 (2001).
49. Sumiyama, K. *et al.* Genomic structure and functional control of the *Dlx3–7* bigene cluster. *Proc. Natl. Acad. Sci. USA* **99**, 780–785 (2002).

ONLINE METHODS

Transgenes and transgenic mice. The SBlac regulatory sensor was constructed by cloning the human β-globin minimal promoter-driven *LacZ* reporter gene[31] and a *lox*P site between the terminal inverted repeats of the Sleeping Beauty transposon[35] (kindly provided by B.S. Fletcher, University of Florida). The SBlac transgene was released (together with a portion of a plasmid polylinker and an additional *lox*P site) as a linear fragment by digesting plasmid pSB11 with *Xho*1. Four independent lines were generated (SBlac-A, -B, -C and -E) by microinjection of this fragment in mouse fertilized oocytes. Copy number was determined (copy numbers of ~20, 2, 8 and 1, respectively) by comparison on protein (Southern) blot with known amount of the transgene (data not shown).

The spermatid-specific transposase transgene consists of the mouse *Prm1* promoter region (PCR amplified from genomic DNA with primers Prm1xho5 and Prm1bh3) cloned upstream of the coding region of the hyperactive Sleeping Beauty transposase *HSB16* (ref. 35) (provided by B.S. Fletcher, University of Florida) coupled with the 3′ untranslated region of the rabbit *β-globin* gene. This *Prm1*∷HSB16 transgene was released by *Acc65*I digestion, gel purified and injected in the pronuclei of fertilized mouse oocytes using standard procedures. Three lines were established (*Prm1*∷HSB16-A, -B and -C). All showed similar activities with respect to transposition efficiency.

Transgenic lines were maintained by breeding with C57Bl/6J mice. Embryos were produced by mating transgenic males with CD1 females. For genotyping, tail or embryonic membrane biopsies were lysed in lysis buffer (10 mM Tris-HCl pH 8, 50 mM KCl, 2 mM $MgCl_2$, 0.1 mg/ml gelatin, 0.45% NP-40 and 0.45% Tween-20) supplemented with proteinase K (100 μg/ml) and heat inactivated. Genotypes were determined by PCR using transgene-specific primers (*Prm1*∷HSB16, primers Prm1_tg1 and HSB16_tg2, product = 474 bp; SBlac. primers SBlac_tg1 and SBlac_tg2, product = 800 bp). Sequences of the different primers are given in **Supplementary Table 1**. After mapping, SBlac insertions were genotyped with one primer specific to the given insertion and one generic transposon (details and primers available on the TRACER database, see URLs). Animal experiments were conducted in accordance with the principles and guidelines defined by the Laboratory Animal Resources of the European Molecular Biology Laboratory.

Identification of transposon insertion sites. To identify transposon insertion sites, we applied a nested asymmetric PCR strategy using Platinum Taq (Invitrogen). A first round of PCR amplification was performed on a genomic DNA template from tail or yolk sac biopsies with a transposon-specific primer pointing outward from the transposon (SB-R1 for the right end, SB-L1 for the left end, 20 μM) and a random primer with a 5-bp 3′ anchor (KmonP-N7-ctcag or KmonP-N7-tcctg, 100 μM). A second round of amplification was carried out on 1 μl of a 1/100 dilution of the first PCR reaction using SB-R2 (right) or SB-L2 (left) and KmonP, both at 20 μM. Primers and PCR programs are detailed in **Supplementary Table 1** and **2**, respectively. The final PCR product was run on a gel for quality testing and sequenced using SB-R3 for the right end or SB-L3 for the left end. The sequences obtained for bona fide transposon insertion contained about 80 bp of one transposon end before reading into the flanking mouse genomic sequence. These flanking sequences were aligned to the mouse genome sequence (Build 37, mm9, using the UCSC Blat webpage, see URLs). Most sequences were long enough to enable unequivocal mapping of the transposon to a unique position (match >99%, next best matches <95%).

LacZ staining. E11.5 mouse embryos were dissected in cold PBS, fixed in PBS with 4% PFA on ice for 30 min, washed twice with ice cold PBS and once at room temperature (19–24 °C), and then stained overnight for β-galactosidase activity in a humid chamber at 37 °C as previously described[4]. After staining, embryos were washed in PBS and stored at 4 °C.

Whole-mount *in situ* hybridization. Mouse embryos were collected at 11.5 days post coitum (dpc), and whole-mount *in situ* hybridization was performed with DIG-labeled gene-specific antisense probes in accordance with established protocols[4]. Probes were generated by SP6 or T7 transcription on linearized templates obtained after cloning a partial complementary DNA (gene-specific primers are available in **Supplementary Table 1**) in pGEM-T-Easy.

Identification of insertional hotspots. We calculated 'l', the interval between a given pair of primary insertions (on chromosomes 9 and 18 for SBlac-A and SBlac-C, respectively), and 'm', the number of insertions within interval 'l'. We computed the *P* value as shown in equations (1) and (2):

$$P_m(l) = \frac{(Rl)^m}{m!} \times e^{-Rl} \tag{1}$$

$$R = \frac{M}{L} \tag{2}$$

with *M* being all primary insertions on the given chromosomes and *L* being the total length of that chromosome**.** This procedure was repeated for each possible pair of primary insertions, and the interval showing the most significant *P* value defines the hotspot. This method identified two local hot spots, chr9: 96,021,889–97,203,726 with 39 insertions for start site A and chr18: 27,967,813–30,846,559 with 27 insertions for start site C. Insertions in a hotspot that came from the corresponding start site were removed to generate the 'hotspot-free' insertions.

Correlation between insertions and other genomic features. The gene list comprises all the genes with full coordinates and an annotated coding region start and end from RefSeq (UCSC Table Browser, mouse genome mm9, RefSeq release 39, as on 1 February 2010). From this list, we generated: (i) a TSS list based on the coordinates of the gene start sites and (ii) an exon list using the genomic positions of the start and end coordinates of all exons. Sperm-active genes were defined from the ArrayExpress datasets E-TABM-412 and E-TABM-130 (ref. 50). These datasets were RMA normalized, and the mean expression level per probe set was computed. We kept probes with expression higher than log2(100) and converted their Affymetrix probeset IDs into Ensembl gene IDs via Ensembl BioMart (v56, NCBIM37). The sperm-active gene list corresponds to the intersection of the Ensembl gene IDs obtained by this process from both datasets. The number of insertions overlapping intragenic regions, exons, introns and intergenic regions was calculated for each of the transposon insertion subsets. The significance (*P* value) of the overlaps was assessed by repeating the analysis with 100 independent randomizations. Randomizations were generated by computing a set of random insertion sites with a sample size equal to the analyzed insertion set. The distances between the insertions and the closest TSS were calculated, and the results were binned into the following categories: <10 kb, 10–100 kb and >100 kb from the TSS. To determine if insertion sets showed an intrinsic bias, we compared their distributions to the distribution obtained from ten independent randomized insertion sets of the same size.

We compared the distribution of insertions between the categories 'expressed' and 'not expressed' when subdivided in two groups according to their distance to the nearest TSS (that is, below or above a given threshold). A χ^2 test was used to determine the significance and to calculate the corresponding *P* value. We considered different thresholds from 10 kb to 500 kb, with 10-kb steps. No significance could be reached for threshold >50 kb. However, for each threshold below 50 kb, the distribution in 'expressed' and 'not expressed' categories of the insertions close to or further away from the TSS were different at a significance level of $\alpha < 0.05$.

Quantitative PCR to measure gene expression. To assess the possible influence of Sleeping Beauty insertions on the mRNA expression of flanking genes, we selected four lines representing intragenic (SB-183610) and gene proximal (SB-177627a) insertions, as well as insertions in gene deserts (SB-183382 and 178318b). Mice were bred to generate wild-type embryos, heterozygous or homozygous for a given insertion, except for SB-177627a, where only heterozygous and wild-type insertions were generated (**Supplementary Table 3**). All embryos were collected at 11.5 dpc and the tissues of interest were dissected, frozen separately in liquid N_2 and stored at −80 °C. Embryonic membranes were collected and used to genotype the embryos (details of primers online through the TRACER database, see URLs). Subsequently, total RNA was isolated from the samples with the required genotype using PureLink RNA Mini Kit (Invitrogen). We subjected 300 ng to 1 μg of isolated RNA to reverse transcription with

ProtoScript M-MuLV First Strand cDNA Synthesis Kit (New England Biolabs). Quantitative PCR was performed on an ABI7500 system with SYBR Green (Applied Biosystems), with every sample being processed in triplicate. Specific primers were used for the genes of interest (**Supplementary Table 1**). For SB-183382 and SB-178318b, we also measured *LacZ* mRNA expression. *GusB* was used as an internal control gene, and expression levels were compared using the ΔΔCp method. No statistically significant differences were found (two-tailed Student's *t*-test with unequal variance) for any of the genes.

50. Lefèvre, C. & Mann, J.R. RNA expression microarray analysis in mouse prospermatogonia: identification of candidate epigenetic modifiers. *Dev. Dyn.* **237**, 1082–1089 (2008).