Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

M.Sc. Lisa Häfele

born in: Heilbronn, Germany

Oral examination: 20th of September, 2024

Profiling pathogenicity of Bovine Meat and Milk Factors in cancer by genome and transcriptome analysis

Referees: Prof. Dr. Benedikt Brors

Prof. Dr. Stefan Wiemann

Abstract

Bovine Meat and Milk factors are circular DNA sequences isolated from bovine milk and serum samples, which have been proposed to contribute to cancer development of different cancer types by inducing chronic inflammation in exposed tissues. While experimental analyses indicated the presence of certain BMMF sequences in different tumor types, only specific BMMF genomes and cancer types have been targeted in experiments so far.

For this reason, I screened multiple publicly available high-throughput sequencing data sets for a comprehensive library of BMMF genomes using the D-ViSioN algorithm to fill this knowledge gap by *in silico* analysis. With this, I managed to prove the feasibility of BMMF detection via computational tools in RNA, WGS and single cell sequencing data and developed processing steps to filter, normalize and characterize the BMMF signal. I screened WGS and RNA sequencing samples of 29 and 25 different cancer cohorts of the PCAWG project, RNA sequencing data of five cancer types provided by the TCGA project, as well as, 15 healthy tissue cohorts derived from healthy donors included in the GTEx project. Additionally, I analyzed cell line data of the DepMap project and a single cell data set of metastatic lung cancer.

I detected BMMF sequences on the RNA and even stronger on the DNA level in tumor and non-tumor samples of patients with a wide range of different cancer types as well as in samples of healthy donors. The detection of BMMF group 1 targets outnumbered the detection of BMMF group 2, 3 and 4 targets by far both on DNA and RNA level. The comparison of BMMF detection in a set of cancer and tissue types across five different data sets revealed the highest percentage of BMMF positive samples for ovarian, stomach and uterine cancer in RNA sequencing data as well as for breast, kidney, lung, pancreas, prostate and stomach cancer in WGS data. For further subtyping of the reported BMMF hits, I defined in total 26 BMMF subgroups spanning the four main BMMF groups. Detailed analysis of BMMF detection at subgroup level showed that for a broad set of BMMF subgroups and a broad range of different cancer cohorts a lower BMMF signal was found in the RNA data of tumor samples compared to matched healthy tissue samples. These findings would indicate no increased cancer risk upon detection of these BMMF types. On the contrary, BMMF subgroup 6 in acute myeloid leukemia and ovarian cancer, subgroup 5 in stomach cancer, subgroup 8 in uterine cancer and subgroup 21 in acute myeloid leukemia were found to be increased in the case versus control comparison of RNA data, which are thus candidates for investigating potential high-risk patterns. While WGS data of early-onset prostate cancer patients exhibited a higher BMMF signal in non-tumor samples than in tumor samples of the same patients, a kidney, lung, pancreatic and prostate

cancer cohort each included two or more BMMF subgroups with increased BMMF detection in tumor samples compared to non-tumor samples of the same patients. These analyses highlighted the importance of BMMF subgroups 1, 5, 6, 7, 10 and 21, which frequently stood out in different data sets analyzed. In addition, I characterized the specific coverage of BMMF reads on the respective BMMF templates for the BMMF genomes C1MI.3M.1, H1MSB.1, C1MI.2 and C1HB.4, which showed that either the entire sequence or large parts of it are covered by BMMF reads indicating a specific detection.

With these analyses, I identified new cancer types-of-interest as well as new target BMMF genomes for further BMMF research. The definition and characterization of BMMF-positive cohorts and subgroups might help to understand the pathogenic phenotype of BMMFs and to establish BMMF detection workflows helpful in diagnostic and therapeutic setup.

Zusammenfassung

Bovine Meat and Milk Faktoren sind zirkuläre DNA Sequenzen, die aus Rindermilch und Rinderserumproben isoliert wurden und die im Verdacht stehen durch das Auslösen chronischer Entzündungen als mögliche Risikofaktoren zu der Entstehung bestimmter Krebstypen beizutragen. Experimentelle Analysen konnten bestimmte BMMFs in verschiedenen Krebsproben nachweisen, jedoch wurde bisher nur eine kleine Gruppe an BMMF Genomen in wenigen verschiedenen Krebsarten experimentell analysiert.

Um diese Wissenslücke zu füllen, habe ich in meiner Doktorarbeit öffentlich verfügbare Sequenzierungsdatensätze mit dem D-ViSioN Algorithmus auf eine Bibliothek an BMMF-Genomen untersucht. Dabei konnte ich zeigen, dass BMMF-Sequenzen in RNA-, Ganz-Genom- und Einzelzellsequenzierungsdaten nachgewiesen werden können, und verschiedene Schritte für das Filtern, Normalisieren und die genauere Charakterisierung der detektierten BMMF-Sequenzen entwickeln. Insgesamt habe ich Ganz-Genom und RNA Sequenzierungsdaten von 29 beziehungsweise 25 verschiedenen Krebskohorten des PCAWG Projektes, wie auch RNA Sequenzierungsdaten fünf verschiedener Krebstypen aus dem TCGA Projekt und 15 Gewebekohorten gesunder Spender aus dem GTEx Projekt auf BMMF-Sequenzen untersucht. Neben den Daten dieser Sequenzierungsprojekte habe ich außerdem Zellliniendaten des DepMap Projektes und einen Datensatz von Einzelzellsequenzierungsdaten aus Patienten mit metastasierendem Lungenkrebs auf BMMFs analysiert.

Sowohl auf RNA- als auch noch stärker auf DNA-Ebene konnte ich BMMF-Sequenzen in Tumor-, in Blut- oder in Gewebeproben außerhalb des Ausgangstumors von Krebspatienten mit einer großen Spannbreite verschiedener Krebstypen detektieren. Die Anzahl der detektieren Reads von BMMF Gruppe 1 übertraf dabei die Detektion von den Gruppen 2, 3 und 4 bei weitem. Der Vergleich der BMMF-Detektion in Tumor- und Gewebeproben fünf verschiedener Datensätze, zeigte, dass die RNA-Daten von Eierstock-, Magen- und Gebärmutterkrebspatienten wie auch die WGS-Daten von Brust-, Nieren-, Lungen-, Bauchspeicheldrüsen-, Prostata- und Magenkrebs die höchsten prozentualen Anteile an BMMF-positiven Proben enthalten. Für eine genauere Klassifizierung der detektierten BMMF-Sequenzen habe ich 26 verschiedene Untergruppen innerhalb der vier BMMF Hauptgruppen definiert. Die detaillierte Analyse der detektierten Untergruppen zeigte für viele verschiedene Untergruppen und Krebstypen ein stärkeres BMMF-Signal in RNA-Daten von gesunden Gewebeproben als in Tumorproben. Dies würde darauf hindeuten, dass kein Zusammenhang zwischen der Anwesenheit dieser BMMF-Untergruppen und diesen Krebsarten besteht. Einige

Tumor-Kohorten des PCAWG RNA-Datensatzes enthielten jedoch auch Untergruppen mit erhöhter Detektion in den Tumorproben verglichen mit gesunden Gewebeproben. Dies betrifft Untergruppe 6 in akuter myeloischer Leukämie und Eierstockkrebs, wie auch Untergruppe 5 in Magenkrebs, Untergruppe 8 in Gebärmutterkrebs und Untergruppe 21 in akuter myeloischer Leukämie. Dementsprechend stellen diese Untergruppen Kandidaten für die weitere Suche nach Hochrisiko-BMMF-Sequenzen dar. Während die WGS-Daten von Prostatakrebs bei jungen Patienten ein höheres BMMF-Signal in Blutproben als in Tumorproben aufweisen, ist die BMMF-Detektion in jeweils einer Nieren-, Lungen-, Bauchspeicheldrüsen- und Prostatakrebs-Kohorte für mindestens zwei Untergruppen erhöht. Dabei stechen immer wieder die Untergruppen 1, 5, 6, 7, 10 und 21 heraus. Deswegen, habe ich im Anschluss die Read-Abdeckung der BMMF-Genome C1MI.3M.1, H1MSB.1, C1MI.2 und C1HB.4 analysiert. Die Abdeckung der gesamten Genome beziehungsweise zumindest weiter Teile davon weist auf eine spezifische Detektion dieser Genome hin.

Mit diesen Analysen konnte ich neue relevante Krebstypen für die BMMF-Forschung identifizieren wie auch mehrere potentielle Hochrisiko-BMMF-Genome. Die Definition and Charakterisierung von BMMF-positiven Kohorten und Untergruppen kann dazu beitragen, die Detektion von BMMFs in Experimenten zu verbessern und die Bedeutung und Rolle von BMMFs in Krankheiten wie Krebs besser zu verstehen.

Acknowledgement

First of all, I want to thank Prof. Dr. Benedikt Brors and Dr. Timo Bund for providing me with the opportunity to be part of their research groups and to write my thesis as a collaboration project between the division of Applied Bioinformatics and the division of Episomal-Persistent DNA in Cancer and Chronic Diseases.

I would also like to thank Prof. Dr. Benedikt Brors and Prof. Dr. Stefan Wiemann for reviewing my thesis as well as Prof. Dr. Gert Fricker and PD Dr. Axel Mogk for joining the Examination Commission for the defense of my thesis.

My sincere thanks go to Dr. Lars Feuerbach and Dr. Timo Bund for the supervision of my thesis, for all of their input and support for this project as well as for their advice during the writing process of my thesis. Additionally, I would like to thank Dr. Prakash Balasubramanian and Kai Horny for their previous work on the D-ViSioN workflow, which allowed the analyses performed in this thesis.

Furthermore, I want to thank all of my colleagues in the groups of Applied Bioinformatics and of Episomal-Persistent DNA in Cancer and Chronic Diseases for their support and the great atmosphere within both groups. Special thanks go to the colleagues in my office for all of the enjoyable and helpful conversations as well as to Corinna Sprengart for her help with all kinds of organizational matters. Finally, I want to thank my family and friends for their support, patience and for believing in me and cheering me up during the entire process of working on this thesis.

Tables of Contents

List of Abbreviations

BMMF	Bovine Meat and Milk Factor
DepMap	Cancer Dependency Map
D-ViSioN	Detection of Integrated Viral Sequences by Singletons
GTEx	Genotype Tissue Expression
IARC	International Agency for Research on Cancer
IHC	immunohistochemistry
MS	multiple sclerosis
NSAID	nonsteroidal anti-inflammatory drugs
ORF	open reading frame
PCAWG	Pan-Cancer Analysis of Whole Genomes
Rep	replication initiator protein
RNS	reactive nitrogen species
ROS	reactive oxygen species
TCGA	The Cancer Genome Atlas
WGS	whole-genome sequencing
ZIR	Zero-Inflated Rank Test

List of figures

Supplementary Figures

List of tables

Supplementary Tables

1. Introduction

Cancer rates are on the rise worldwide, between 2020 and 2040 the number of cancer cases diagnosed per year is predicted to increase by 47 % (Sung et al., 2021). This can be partly attributed to the global population growth and the increasing life expectancies in many regions of the world, but also to the exposure to risk factors and changing lifestyle habits such as smoking, alcohol consumption, sun exposure, nutrition or physical exercise (Brown et al., 2018; Clinton et al., 2020; Key et al., 2020; Sung *et al.*, 2021). Currently, each individual has an average lifetime risk of developing cancer of about 20%, with 8.86% of all women and 12.59% of all men dying from cancer. The five most frequently diagnosed cancer types worldwide include breast, lung, colorectal, prostate and stomach cancer (Sung *et al.*, 2021). For all of these cancer types as well as for several other cancer types such as liver cancer, pancreatic cancer or esophageal cancer lifestyle-related risk factors have been proposed (World Cancer Research Fund/American Institute for Cancer Research, 2018). Besides of diet, lack of physical activity or exposure to UV light, there are also other preventable cancer causes, such as infectious diseases. Infectious agents are reported to cause about 15% of all cancer cases (Plummer et al., 2016). Chronic inflammation is another known risk factor for carcinogenesis, which can be induced by a range of different conditions as well as dietary and environmental factors. Inflammation has been proposed as explanation for observed correlations between certain diets, obesity and cancer (Steck and Murphy, 2020; World Cancer Research Fund/American Institute for Cancer Research, 2018). Furthermore, chronic inflammation caused by infectious agents is involved in several cancer types (Lu and Li, 2014; White et al., 2014). Additionally, inflammatory diseases causing chronic inflammation are associated with multiple cancer types, such as pancreatic cancer or stomach cancer, where pancreatitis and gastritis are known to precede cancer development (Farrow and Evers, 2002; Lowenfels et al., 1993; World Cancer Research Fund/American Institute for Cancer Research, 2018).

In this thesis I aim to investigate the role of a group of proposed infectious agents called Bovine Meat and Milk Factors (BMMF) in a broad range of different cancer types. BMMFs have been discovered in bovine milk and serum samples and are therefore expected to be taken up by humans via the consumption of bovine milk and meat products throughout their entire lives. BMMFs have been postulated to cause chronic inflammation in exposed tissues, which can result in carcinogenesis after long periods of latency (zur Hausen et al., 2019; zur Hausen and de Villiers, 2015a). To analyze the impact of BMMFs for cancer development, I used the D-

ViSioN workflow to screen high-throughput sequencing data of cancer patients and healthy tissue samples for the presence of a library of BMMF sequences.

1.1 Diet, nutrition and cancer

In the past decades, the interest in the association of lifestyle factors with cancer has continuously grown, since this information is crucial for cancer prevention strategies. Several of the most common cancer types such as breast, colorectal and prostate cancer have been investigated in the context of diet and nutrition, but also regarding other factors such as body weight, exercise and breastfeeding (Steck and Murphy, 2020). The analysis of lifestyle habits is complex, since it includes combinations of different factors and nutrients that often interact with each other as well as with the gut microbiome. Additionally, in many cases there is a high heterogeneity within the study populations, because the analyses often rely on self-report of the study subjects and the definitions of the dietary patterns investigated vary in different study designs. These problems might explain, why for some cancer types the risk assessment for certain dietary patterns varies between different studies with different setups (Steck and Murphy, 2020). The most consistent association of certain dietary patterns with cancer risk, is reported for colorectal cancer. Unhealthy diets with high levels of red and processed meat consumption are described to increase the risk, whereas plant-based diets are reported to reduce colorectal cancer risk (Grosso et al., 2017; Steck and Murphy, 2020). Similar patterns – but with less consistent evidence – have been discussed for a number of other cancer types such as breast, lung or ovarian cancer (Steck and Murphy, 2020). In case of breast and lung cancer, healthy nutrition has been described to reduce cancer risk significantly, while on the other hand some, but not all studies indicate an association of unhealthy diet with an increased risk for these two cancer types (Grosso *et al.*, 2017). Besides of different diets, alcohol consumption is also listed as risk factor for several cancer types including colorectal, liver, breast and esophageal cancer due to the carcinogenic intermediate product acetaldehyde, that is generated during ethanol metabolism in the human body (Clinton *et al.*, 2020; Key *et al.*, 2020; Pöschl and Seitz, 2004; Steck and Murphy, 2020). Based on the data of the 2012 GLOBOCAN report, alcohol consumption is to be blamed for 5.5% of all cancer cases (LoConte et al., 2018; Praud et al., 2016).

1.1.1 Carcinogenicity of red and processed meat

Red meat is generally defined as unprocessed, usually cooked meat including for example pork, beef, lamb or goat, whereas processed meat refers to meat treated with for example salting,

curing or smoking (Bouvard et al., 2015). Colorectal cancer has been most intensively investigated for the association with red and processed meat consumption with a broad range of epidemiological studies from different origins. When analyzing studies for red and processed meat separately, the International Agency for Research on Cancer (IARC) figured, that the majority of both cohort studies and case-control studies indicated a correlation of colorectal cancer with high levels of consumption of both red and processed meat (Bouvard *et al.*, 2015). The IARC also reported a positive association for the consumption of red meat and pancreatic and prostate cancer, whereas stomach cancer was found to correlate with the consumption of processed meat. After the assessment of more than 800 studies, the IARC reported sufficient evidence to consider processed meat as carcinogenic and additionally stated, that there is as well "limited evidence (…) for the carcinogenicity of the consumption of red meat" (Bouvard *et al.*, 2015). There are different possibilities, how red and processed meat might contribute to the increased cancer risks. During processing steps of meat, there are often potentially carcinogenic chemical compounds generated (Bouvard *et al.*, 2015). The consumption of these chemicals might explain the increased cancer risks.

1.1.2 Dairy product and their role in cancer

Dairy products contain several micronutrients such as calcium or vitamin D, which are beneficial for general health. However, many dairy products have high fat contents. Furthermore, dairy products can also be contaminated by pesticides and growth factors, which could promote cancer (Moorman and Terry, 2004). Dairy products have been repeatedly investigated for their association with cancer, however with opposing results. In some studies, dairy products have been reported as risk factors, whereas in other cases they have been found to decrease cancer risk (Moorman and Terry, 2004; Steck and Murphy, 2020). For colon cancer some studies indicated a reduced cancer risk for higher levels of milk or dairy products consumption. This might be caused by the intake of calcium, which is thought to bind secondary bile acids and ionized fatty acids as well as to interact with multiple signaling pathways (Aune et al., 2012; Clinton *et al.*, 2020). On the other hand, high levels of dairy consumption in childhood have been reported in association with increased colorectal cancer rates (van der Pols et al., 2007). Dairy product consumption has also been linked with an increased risk in case of prostate cancer (Clinton *et al.*, 2020). In case of breast cancer, there are conflicting reports. While some studies found no association between the consumption of dairy products, another study reported an increased risk for breast cancer due to milk and dairy products as well as a beneficial effect when replacing dairy milk with soy milk (Fraser et al., 2020; Wu et al., 2021).

Studies regarding lactose intolerance showed a significantly reduced breast, lung and ovarian cancer risk for lactose intolerant individuals compared to their siblings and parents. Since people affected by lactose intolerance are expected to consume low amounts of dairy products, this observation suggests a potential link between dairy products and these three cancer types (Ji et al., 2015; zur Hausen and de Villiers, 2015a).

1.1.3 Link between diet and inflammation

Inflammation has been associated with cancer since the 19th century. Acute inflammation is a crucial part of our body's defense against infectious agents as well as of the wound healing process (Murphy, 2012). Chronic inflammation however is linked to a number of different cancer types, since tissues affected by chronic inflammation are continuously exposed to reactive nitrogen and oxygen species, which are produced as part of the inflammation reaction and cause DNA mutations (Todoric et al., 2016). Inflammation plays a role in the progression of most solid cancers, however in about one in five cancers the tumor initiation can be directly linked to chronic inflammation (Steck and Murphy, 2020; Todoric *et al.*, 2016). The link between cancer and chronic inflammation is further confirmed by observations of long-term use of nonsteroidal anti-inflammatory drugs (NSAID), which results in a significant reduction of cancer risk and mortality for several cancer types (Rothwell et al., 2011; Todoric *et al.*, 2016). The anti-inflammatory effect of NSAID has been described to reduce the risk of dying because of esophagus, colorectal, stomach, breast, lung, ovarian and prostate cancers (Rothwell *et al.*, 2011; Todoric *et al.*, 2016; zur Hausen *et al.*, 2019).

The chronic inflammation resulting in cancer development can be induced by oncogenic infectious agents or chronic diseases accompanied by chronic inflammation such as gastritis, pancreatitis and inflammatory bowel disease (Harmon et al., 2017; Todoric *et al.*, 2016; World Cancer Research Fund/American Institute for Cancer Research, 2018; zur Hausen *et al.*, 2019). Besides of these causations for chronic inflammation, a number of lifestyle factors have also been linked with inflammation (Todoric *et al.*, 2016). Higher incidence rates of diseases linked to chronic inflammation in western countries compared to other countries and regions indicate a link between lifestyle factors and chronic inflammation (Furman et al., 2019). Unhealthy diets have been described to have a positive association with inflammation due to causing higher presence of inflammatory markers such as cytokines and C-reactive protein (Ahluwalia et al., 2013; Harmon *et al.*, 2017). So-called proinflammatory diets include red and processed meat as well as on the one hand high contents of sugar and saturated fatty acids and on the other hand low contents of fiber and certain groups of vegetables, that have been described to be inversely

correlated with inflammation (Liu et al., 2017; Tabung et al., 2018). While dairy products have been frequently investigated regarding their effect on inflammation, most studies in this field indicate either no positive association between the consumption of dairy products and inflammation or even a reducing effect of dietary products on inflammation (Hess et al., 2021; Nieman et al., 2021). Proinflammatory diets increase the risk for colorectal cancer (Harmon *et al.*, 2017; Tabung *et al.*, 2018). Additionally, obesity is linked with inflammation and has been described as risk factor for several different cancer types (Grosso *et al.*, 2017; Todoric *et al.*, 2016; World Cancer Research Fund/American Institute for Cancer Research, 2018). Metabolic stimuli such as salt or saturated fatty acids are able to activate macrophages, which leads to cytokine expression and consequently to inflammation (Furman *et al.*, 2019; Hess *et al.*, 2021). Adipose tissue macrophages have been reported to accumulate in adipose tissue in case of obesity where they facilitate inflammation of the adipose tissue as well as lipid homeostasis (Chavakis et al., 2023; Kratz et al., 2014). Consequently, the combination of nutrition and obesity can cause chronic inflammation in adipose tissue, which is believed to explain the correlation between obesity and cancer risk.

1.1.4 Epidemiological observations

Lifestyle habits and dietary patterns vary geographically and culturally. Thus, epidemiological observations of cancer incidence rates in different geographical regions and populations can provide insights in potential links between nutrition and cancer. Besides of lifestyle factors, potential genetic differences between populations could also contribute to the different geographic patterns observed for some cancer types. However, migrant studies showed for breast and colorectal cancer, that the cancer risk increases for migrants and their descendants after moving from a country with low breast and colorectal cancer incidence rates to one with high breast and colorectal prevalence (Moorman and Terry, 2004; zur Hausen and de Villiers, 2015a). This indicates, that lifestyle factors such as nutrition play an important role for the differences in the geographical patterns of these cancer types (Moorman and Terry, 2004).

Breast cancer accounts for about 25 % of all cancer diagnoses in women, however the incidence rates differ significantly between different regions of the world. Europe, North America, Australia and New Zealand have the highest breast cancer incidence rates worldwide (Fig. 1.1). Countries with a higher Human Development Index generally exhibit higher breast cancer incidence rates compared to transitioning countries (Sung *et al.*, 2021). However, cancer rates are increasing in both transitioning countries as well as in developed countries such as Japan, which used to have a lower breast cancer prevalence than many other transitioned countries

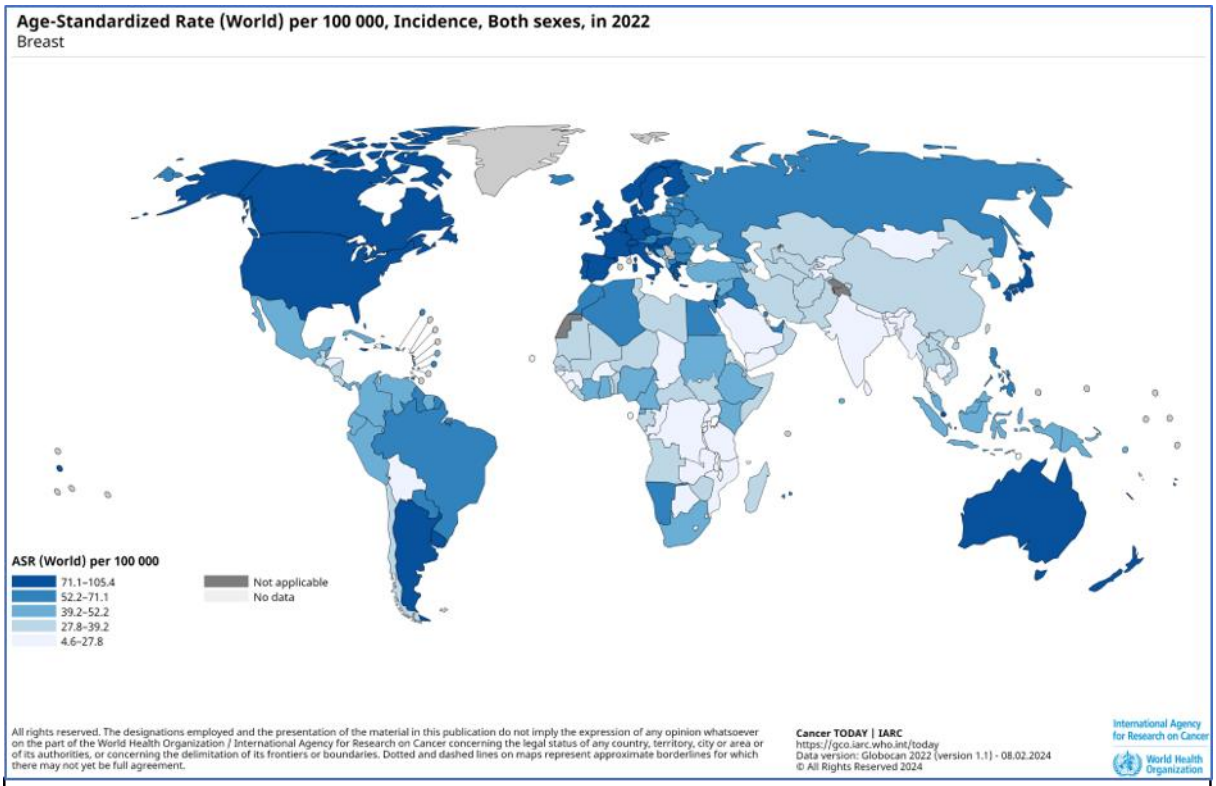


Fig. 1.1: Global breast cancer incidence rates: Map with worldwide breast cancer incidence rates based on Globocan 2022 data, generated using the IARC/WHO dataviz online tool. Link: https://gco.iarc.fr/today/en/dataviz/maps-heatmap?mode=population&cancers=20

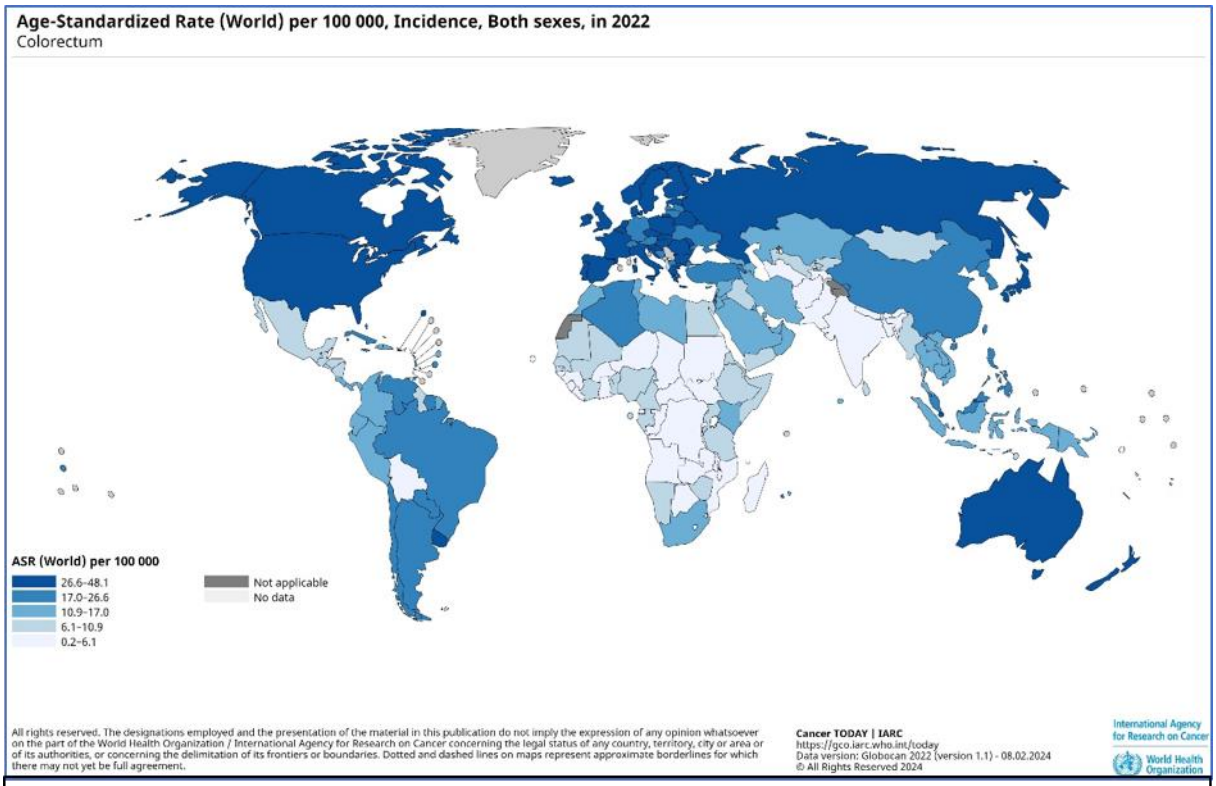


Fig. 1.2: Global colorectal cancer incidence rates: Map with worldwide colorectal cancer incidence rates based on Globocan 2022 data, generated using the IARC/WHO dataviz online tool. Link: https://gco.iarc.fr/today/en/dataviz/maps-heatmap?mode=population&cancers=41

(Li et al., 2003; Sung *et al.*, 2021). These increases are attributed to changes in lifestyle habits. The Japanese breast cancer rates for example used to be very low, but have been growing during the past decades. The beginning of the rise of breast cancer in Japan can be traced back to the end of World War II and the subsequent changes in lifestyle and nutrition (Li *et al.*, 2003; Sung *et al.*, 2021). The tremendous increases in the consumption of meat, dairy products and eggs are discussed as potential causes for the shift in breast cancer incidence (Li *et al.*, 2003). Besides of dietary patterns there are also several other lifestyle factors that have to be considered for the increased breast cancer rates such as later childbirth, breast feeding or exposure to hormones (Li *et al.*, 2003; Sung *et al.*, 2021).

In case of colorectal cancer, the observed geographical distributions and developments are similar to the ones previously described for breast cancer. Colorectal cancer incidence is four times higher in developed countries than in transitioning countries and regions (Sung *et al.*, 2021). Just as breast cancer, colorectal cancer occurs most frequently in Europe, North America, Australia and New Zealand (Fig. 1.2). However, colorectal cancer rates are also increasing in other regions due to changes regarding to dietary patterns and other lifestyle factors such as physical activity and body weight (Sung *et al.*, 2021).

1.1.5 Link between epidemiological observations and potential infectious agents

Epidemiological observations of breast and colorectal cancer incidence rates, indicate a correlation between lifestyle factors and the prevalence of these two cancer types. The IARC reported a positive association between colorectal cancer and the consumption of red and processed meat, however the reason for this link remains unclear (Bouvard *et al.*, 2015). While chemical generated during the processing of meat have been discussed as potential explanation for this, these chemicals are also produced during the processing steps of chicken or fish, which have not been linked to an increased cancer risk (Bouvard *et al.*, 2015; Nayak et al., 2009; zur Hausen, 2012).

N-glycolylneuraminic acid (Neu5Gc) has been additionally been suggested as potential reason for the correlation between consumption of red meat and colorectal cancer (Alisson-Silva et al., 2016; zur Hausen *et al.*, 2019). Neu5Gc is a sialic acid, that cannot be produced by humans, but by certain animals. Consequently, after the breast-feeding period Neu5Gc is ingested via red meat and dairy products and subsequently incorporated in human glycoproteins and gangliosides (Alisson-Silva *et al.*, 2016; zur Hausen *et al.*, 2019). Due to these changes to proteins and lipids of human cellular membranes, infectious agents can bind to receptors to

which they were unable to bind previously during the breast-feeding period (zur Hausen *et al.*, 2019). After the incorporation at the cell surface, Neu5Gc can act as antigen. This can result in the formation of antibodies, which cause an immune reaction when encountering Neu5Gc at cell surfaces leading to inflammation (Alisson-Silva *et al.*, 2016). Interestingly, Neu5Gc is enriched in red meat, but absent in chicken and either not present or present in alternative variants in fish (zur Hausen *et al.*, 2019). Thus, Neu5Gc consumption could explain the increased colorectal cancer risk due to red meat consumption. However, epidemiological observations show, that countries such as Mongolia exhibit low colorectal and breast cancer rates in spite of high levels of red meat consumption (Fig. 1.1, Fig 1.2) (zur Hausen *et al.*, 2019).

Furthermore, epidemiological observations indicate a correlation specifically between the consumption of beef and colorectal cancer. Countries such as Japan and Korea reported increased colorectal cancer rates after the consumption of beef, pork and dairy products gained popularity (zur Hausen, 2012; zur Hausen and de Villiers, 2015a). India on the other hand, where the majority of the population does not consume beef, exhibits very low colorectal cancer incidence rates (Fig. 1.2). The only Indian regions with higher colorectal cancer prevalence are inhabited by minority groups, that consume beef, which highlights a potential connection between beef consumption and colorectal cancer risk (Nayak *et al.*, 2009; zur Hausen, 2012; zur Hausen and de Villiers, 2015a).

Just as India, Mongolia also has low colorectal cancer and breast cancer incidence rates however with high levels of red meat consumption (Fig. 1.1, Fig. 1.2). In contrast to India beef consumption accounts for 40-50% of the total red meat consumption (zur Hausen and de Villiers, 2015a). While in most regions of the world *Bos taurus*-derived cattle is most prevalent and makes up for the majority of bovine meat and milk consumption, Mongolian cattle mostly includes species adapted to the climate such as Yaks, local *Bos taurus* sub-species and cattle cross-breeded with Zebus. Just as Mongolia, Bolivia also stands out with low incidence rates for colorectal and breast cancer (Fig. 1.1, Fig. 1.2) as well as with a cattle population mainly comprised of Zebus instead of Eurasian *Bos taurus*-derived cattle. Consequently, meat and milk consumption of Yaks or Zebu-derived cattle species, seems to correlate with lower colorectal and breast cancer rates compared to countries with *Bos taurus* based bovine meat and milk consumption (zur Hausen and de Villiers, 2015a). Thus, epidemiological observations hint, that meat and milk products of *Bos taurus*-derived cattle might contain a specific factor causing increased colorectal cancer risks. This factor could be an infectious agent present in Bos taurus-derived cattle, which might be involved in carcinogenesis of colorectal cancer and potentially

as well of breast cancer, which shows very similar epidemiological patterns as colorectal cancer (Fig. 1.1, Fig. 1.2) (zur Hausen and de Villiers, 2015a).

For breast cancer, a correlation between breast-feeding periods of women and their breast cancer risk has been reported for a long time (zur Hausen *et al.*, 2019). In women with multiple periods of breast-feeding, the breast cancer risk was found to be reduced by up to 36% (Faupel-Badger et al., 2013; zur Hausen *et al.*, 2019). Breast-feeding is known to prevent a number of viral infections in babies due to the characteristic set-up of oligosaccharides in human milk, which block cell surface receptors of infectious agents and thus prohibit infectious agents from binding to them (zur Hausen *et al.*, 2019). Bovine milk contains a different composition of oligosaccharides, which might have developed specifically for protection against bovine pathogens (zur Hausen, 2017a). During lactation the protective oligosaccharides are not only present in human milk, but also in the breast tissue. Since these oligosaccharides are also detectable in the blood or urine of breast-feeding women, it is also likely that other tissues are exposed to them at lower rates (de Villiers and zur Hausen, 2021; zur Hausen *et al.*, 2019). Consequently, the protective effect of breast-feeding might also stretch to other cancer types albeit with lower effects. Some studies also for example show reduced colorectal cancer risks linked to multiple breast-feeding periods and a young age at the first childbirth (zur Hausen *et al.*, 2019). Since the characteristic sugars present in human milk protect against viral infections, the observations about reduced breast and to some extent also colorectal cancer rates in women with long-term breast-feeding periods could also indicate an infectious agent playing a role in breast and colorectal cancer (zur Hausen *et al.*, 2019). Synergy effects between such an infectious agent of bovine origin with chemicals produced during meat processing or Neu5Gc containing cell membranes seem possible (zur Hausen, 2012; zur Hausen *et al.*, 2019; zur Hausen, 2017b).

1.2 Infectious agents and cancer

Infectious agents are estimated to cause about 15-20% of cancer cases (White *et al.*, 2014; zur Hausen, 2001). Based on the GLOBOCAN database, 2.2 million cancer diagnoses in 2018 can be linked to carcinogenic pathogens, which comprises about 13% of the newly diagnosed cancer cases in that year (de Martel et al., 2020). In the 2012 GLOBOCAN database, 15.4% of the new cancer diagnoses were attributed to carcinogenic infectious agents (Plummer *et al.*, 2016). The identification of infectious agents involved in cancer development is complex, since there can be decades of latency between the first contact with the infectious agent and the cancer diagnosis (zur Hausen and de Villiers, 2015b). Currently, there are 11 infectious agents defined

as oncogenic (Tab. 1.1). The largest fraction of the cancer cases attributed to infectious agents is caused by seven oncogenic viruses, followed by the cancer cases attributed to bacteria. *Helicobacter pylori* is so far the only bacterium classified as oncogenic by the IARC. *H. pylori* causes about 36% of gastric cancer cases (de Martel *et al.*, 2020; Hansen et al., 2021; Plummer *et al.*, 2016). The infection with *H. pylori* can result in the manifestation of chronic gastritis, which can be a precursor to stomach cancer (Correa, 1992; Lu and Li, 2014; Parsonnet et al., 1991). Besides of oncogenic viruses and *H. pylori*, the three parasites *Opisthorchis viverrini*, *Clonorchis sinensis* – which are both causing cholangiocarcinoma – and *Schistosoma haematobium* – causing bladder carcinoma – are also defined as oncogenic pathogens by the IARC (de Martel *et al.*, 2020; Dheilly et al., 2019).

Tab. 1.1: Overview of oncogenic infectious agents: 11 infectious agents have been classified as oncogenic by the IARC. They are listed with the respective cancer types they are causing and their mechanism of carcinogenicity, if known.

Infectious agent	Cancer type	Promotion of carcinogenesis
Helicobacter pylori	Gastric cancer	Inflammation + chronic gastritis (Todoric *et al.*, 2016)
Human papillomaviruses	Cervical carcinoma, anogenital + oropharyngeal cancers	E6 + E7 oncogenes, integration in host genome
Hepatitis B virus	Hepatocellular carcinoma	Inflammation
Hepatitis C virus	Hepatocellular carcinoma	Inflammation
Epstein-Barr virus	Burkitt's lymphoma, Hodgkin lymphoma	Role in carcinogenesis not completely understood, expression of EBNA1, LMP1 and LMP2
Human T-lymphotropic virus 1	T-cell leukemia	Expression of tax protein, integration in host genome
Kaposi sarcoma-associated herpesvirus	Kaposi's sarcoma	Viral protein expression + adapts immunes signaling
Merkel cell polyomavirus	Merkel cell carcinoma	Integration in host genome
Opisthorchis viverrini	Cholangiocarcinoma	Not understood (Dheilly *et al.*, 2019)
Clonorchis sinensis	Cholangiocarcinoma	Not understood (Na et al., 2020)

Schistosoma haematobium	Bladder carcinoma	Promotes bacterial co-infection, associated with high levels of nitrosamines (Dheilly *et al.*, 2019)

1.2.1 Oncogenic viruses

The first oncogenic virus was discovered in chicken by Peyton Rous in 1911, who reported that a "cell-free filtrate" was sufficient to transmit a sarcoma to a healthy chicken (Rous, 1911). Epstein-Barr virus (human herpesvirus 4, EBV) was the first oncogenic virus described in humans in 1964, when it was discovered in a Burkitt's lymphoma cell culture (Epstein, 2015). While EBV was first linked to Burkitt's lymphoma, it is also discussed in relation to several other cancer types such as gastric cancer, different types of lymphomas and nasopharyngeal carcinoma (Chang et al., 2017; Shannon-Lowe et al., 2017).

Today, seven oncoviruses are known. Hepatitis B virus (HBV) and hepatitis C virus (HCV) both cause hepatocellular carcinoma, HBV accounts for 56 % and HCV for 20 % of liver cancer mortality worldwide (Sung *et al.*, 2021; White *et al.*, 2014). Human papilloma viruses (HPV) are best known for their role in cervical carcinoma, however they are also involved in several anogenital cancers such as penis, vulva, vagina, anal cancer as well as in oropharyngeal cancers (Illah and Olaitan, 2023; White *et al.*, 2014; zur Hausen, 2002). It is estimated that every second cancer caused by infectious agents in women can be attributed to HPV, whereas only 5% of infection-linked cancers in men are linked to HPV (White *et al.*, 2014). Together with *H. pylori*, the three viruses HBV, HCV and HPV account for 90% of all cancer cases attributed to infectious agents (de Martel *et al.*, 2020; Plummer *et al.*, 2016). Additionally, three further oncogenic viruses have been described so far: human T-lymphotropic virus 1 (HTLV-1) was discovered to cause adult T-cell leukemia, Kaposi sarcoma-associated herpesvirus (KSHV/human herpesvirus 8) was found in Kaposi's sarcoma, and Merkel cell polyomavirus (MCV) was reported in Merkel cell carcinoma (Krump and You, 2018; White *et al.*, 2014). Cancer types caused by oncogenic viruses occur at increased levels in HIV patients, which might for example be caused by the immunosuppression of the host organism due to HIV (Proulx et al., 2022; zur Hausen and de Villiers, 2015b). However, HIV patients also exhibit higher rates of several non-infectious cancer types (White *et al.*, 2014). Despite of this, HIV is generally not included in the list of oncogenic viruses, since the mechanisms behind the increased cancer risk are yet to be understood (Proulx *et al.*, 2022; White *et al.*, 2014).

There are two different modes in which oncogenic viruses may cause cancer: On the one hand there are viruses directly promoting carcinogenesis by the expression of viral genes, that can for example integrate into the host genome (de Villiers and zur Hausen, 2021). On the other hand, there are oncogenic viruses that indirectly contribute to cancer by inducing chronic inflammation in affected tissues (de Villiers and zur Hausen, 2021; zur Hausen, 2001). HPV is a prominent example of an oncovirus directly promoting cancer development. High-risk HPVs such as HPV 16 and 18 are able to directly cause cancer in their epithelial target cells due to the viral E6 and E7 oncogenes (Graham, 2017). Both of these genes are able to integrate into the host genome and the expressed oncoproteins can change the host cell's expression pattern (Graham, 2017; White *et al.*, 2014; zur Hausen, 2001). E6 targets several apoptosis regulators such as p53 and as well as modulating cell signaling, gene expression and proliferation, whereas E7 manipulates the cell cycle of keratinocytes and targets a broad set of other proteins such as tumor suppressor protein Rb or STING, which plays an important role in signaling pathways of the immune system (Graham, 2017; Krump and You, 2018; zur Hausen, 2002).

Besides of HPV, EBV is also considered as a direct contributor to carcinogenesis. EBV is a DNA virus with a large genome encoding 85 genes and after the initial infection phase it latently persists as episome in memory B lymphocytes in the tonsils (Krump and You, 2018; Thorley-Lawson, 2015; White *et al.*, 2014). While all Burkitt's lymphomas exhibit a translocation of the oncogene c-MYC, not all Burkitt's lymphomas contain EBV DNA (Chang *et al.*, 2017). Since EBV DNA was most prevalent in endemic Burkitt's lymphomas in malaria regions, it is assumed that these Burkitt's lymphoma cases can be attributed to co-infection of EBV and malaria (Chang *et al.*, 2017; Rochford and Moormann, 2015; Shannon-Lowe *et al.*, 2017). The exact mechanism EBV uses to facilitate carcinogenesis of Burkitt's lymphoma is not clear, however the EBV protein EBNA1 has been reported to affect the regulation of genes concerning apoptosis and cell death. EBV might also increase the likelihood of c-MYC translocation. In Hodgkin lymphoma about 30-50% of the cases are linked to EBV, but it unknown as well how EBV promotes carcinogenesis. However, the expression of different EBV proteins such as EBNA1, LMP1 and LMP2 has been documented for Hodgkin lymphoma and other EBV-linked lymphoma types and the respective proteins are thought to prevent apoptosis and affect signaling pathways (Shannon-Lowe *et al.*, 2017; White *et al.*, 2014). Additionally, some EBV-linked cancer types are also known for their relation to inflammation, which might also contribute to EBV-supported carcinogenesis (Shannon-Lowe *et al.*, 2017).

Furthermore, HTLV-1, KSHV and MCV are also reported to have a direct impact on their host cells. KSHV persists as an episome similar to EBV and expresses viral proteins, that modulate transcription, the cell cycle and tumor suppressor proteins. Additionally, some KSHV mimic interleucins and other proteins involved in immune signaling. The retrovirus HTLV-1 also expresses viral proteins, most importantly the Tax protein. Tax is involved in transforming the host cell by modulating transcription factors to manipulate the host cell's cell cycle and proliferation as well as by promoting DNA damage. Besides of expressing viral proteins, HTLV-1 is also able to clonally integrate into the host genome. The same is the case for MCV, for which clonal integration seems to be causative for its carcinogenic activity. Integration always disables the replication function of MCV's replication protein while retaining its ability to interact with cellular proteins such as the tumor suppressor Rb (White *et al.*, 2014; zur Hausen and de Villiers, 2015b).

On the contrary, HBV and HCB are considered to contribute to carcinogenesis mostly in an indirect way by inducing chronic inflammation (White *et al.*, 2014; zur Hausen and de Villiers, 2015b). However, HBV can also take a more direct role via integrating into the host genome. Integration of HBV can happen at many different places in the human genome and does not always have functional consequences. However, through integration events interrupting genes involved in cellular signaling and regulation, HBV can have a direct role in carcinogenesis as well (Peneau et al., 2022; White *et al.*, 2014). Additionally, HBV expresses the HBx gene, which was shown to have oncogenic properties *in vitro*. Just as HBV, HCV might partially directly cause liver cancer as well, since its core protein is able to interact with different transcription factors including p53 (Heredia-Torres et al., 2022; White *et al.*, 2014).

1.2.2 Prevention and treatment of cancers caused by oncogenic viruses

The discovery of the role of different infectious agents in carcinogenesis opened new paths for treatment and prevention of cancer (zur Hausen, 2001). The development of antiviral therapy for example facilitated progress in the treatment of HCV infections (Krump and You, 2018; White *et al.*, 2014). For other cancer types caused by viruses, the development of vaccines provided a safe tool for cancer prevention. So far, vaccinations against HBV as well as against several HPV types are available, whereas no vaccine could be developed yet for HCV (White *et al.*, 2014). The first cancer preventing vaccination was established in the 1980s, when the first HBV vaccines became commercially available (Emini et al., 1986; Pattyn et al., 2021). Due to the long timespans between the manifestation of chronic HBV infection and the development of hepatocellular carcinoma, which can comprise several decades, the effects of

the HBV vaccine on global liver cancer burden can only be observed with delay (Flores et al., 2022). While the global immunization coverage was at about 85% in 2019, just 30% of all infants were fully vaccinated against HBV in 2000, which means that the effects of the vaccination will not be visible yet everywhere (Pattyn *et al.*, 2021). However, more than three decades after the start of the first HBV vaccination campaigns for infants, countries that started early HBV vaccination programs such as Taiwan show reduced hepatocellular carcinoma incidence rates among children and young adults born after the introduction of the vaccine (Flores *et al.*, 2022; Pattyn *et al.*, 2021).

Vaccination against HPV started about two decades later compared to the HBV vaccination. In 2006 the first HPV vaccine was licensed to immunize against the two most prevalent high-risk HPV types 16 and 18, which cause about 70% of all cervical cancers, as well as against the most common low-risk HPV types 6 and 11, which account for 90% of all genital warts (Cheng et al., 2020; Illah and Olaitan, 2023). Since 2014 a nonavalent vaccine including five more oncogenic HPV types is available as well (Cheng *et al.*, 2020; Illah and Olaitan, 2023). The HPV vaccines currently approved rely on virus-like-particles of the HPV capsid protein L1 for immunization (Cheng *et al.*, 2020; Illah and Olaitan, 2023). While the HPV vaccines were mostly targeted at women at the beginning of their introduction, they are now also approved for men to prevent anogenital and oropharyngeal cancers. First national reports after a decade of HPV vaccination indicate that the vaccines reduce HPV 16 and 18 prevalence as well as cervical cancer rates (Illah and Olaitan, 2023). Since HPV vaccination has only be introduced during the past 15 years, there is still a high HPV prevalence in the general population due to HPV infections prior to the availability of vaccines as well as to too low vaccination rates. Hence, there is still a big necessity for treatment of chronic HPV infections. Currently, there are several clinical trials for therapeutic vaccines in progress, however no therapeutic HPV vaccine has been licensed at this point (Illah and Olaitan, 2023).

1.3 Bovine Meat and Milk Factors

An infectious agent present in bovine meat and milk product of *Bos taurus*-derived cattle was postulated as explanation for the correlation between consumption of *Bos taurus*-derived meat and milk products and colorectal and breast cancer incidence rates (zur Hausen and de Villiers, 2015a). Circular DNA sequences called **B**ovine **M**eat and **M**ilk **F**actors (BMMF) have been proposed as potential infectious agents present in bovine meat and milk products. BMMFs have been first isolated from bovine milk and serum samples as well as from brain lesions of a multiple sclerosis patient (zur Hausen, 2017a). Multiple sclerosis (MS) has also been repeatedly

linked to milk consumption (Morin et al., 2024; zur Hausen, 2015; zur Hausen, 2017a). Most BMMFs known so far are closely related to two sequences known as Sphinx.176 and Sphinx2.36, which have been previously described in context of transmissible spongiform encephalopathy (Manuelidis, 2011).

1.3.1. Sphinx DNAs and transmissible spongiform encephalopathy

Transmissible spongiform encephalopathies (TSEs) are a group of rare neurodegenerative disorders better known as prion diseases (Scheckel and Aguzzi, 2018). Prion diseases are caused by infectious, misfolded proteins called prions, which build aggregates in the central nervous system by assembling healthy correctly folded prion proteins with misfolded proteins (Scheckel and Aguzzi, 2018; Zerr et al., 2024). The incubation times of TSEs are typically very long and can last up to decades (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024). TSEs can be grouped into three different forms, which include genetic, acquired, sporadic TSEs (Scheckel and Aguzzi, 2018). Additionally, polymorphisms in the PRNP gene have been described to cause a predisposition to prion diseases, respectively an early onset of genetic prion diseases, whereas other polymorphisms might even have protective effects (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024). Prion diseases occur not only in humans, but also in several other mammals such as scrapie, which affects sheep and goats, or bovine spongiform encephalopathy (BSE), which has been described in cattle (Scheckel and Aguzzi, 2018).

About 85 % of human prion disease cases are caused by sporadic Creutzfeldt-Jakob disease (Zerr *et al.*, 2024). However, while sporadic TSEs are the most frequent prion diseases, their cause is unknown (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024). Genetic forms such as genetic Creutzfeldt-Jakob disease or fatal familial insomnia on the other hand are always characterized by point mutations in the PRNP gene encoding the prion protein (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024). The transmission of prions between humans or other mammalian species causes acquired TSEs such as Kuru, which was transmitted via cannibalistic rituals, iatrogenic Creutzfeldt-Jakob disease or variant Creutzfeldt-Jakob disease, which is transmitted to humans due to consumption of bovine meat contaminated by BSE (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024; zur Hausen, 2017a).

Prion aggregates follow a cycle of nucleation and fragmentation in which larger aggregates split into new replicating fragments that form new aggregates (Scheckel and Aguzzi, 2018; Zerr *et al.*, 2024). Due to this, the accumulation of protein aggregates in the central nervous system not only prevents the original protein function, it also interferes both with cellular processes and

intercellular communication. It is only poorly understood, how prions affect neurons, but in the end the accumulation of prions leads to neurotoxicity (Scheckel and Aguzzi, 2018). Besides of prion diseases, several other protein misfolding diseases are known including Alzheimer's disease or Parkinson's disease, which share many features with prion diseases. However, in contrast to prion diseases, protein misfolding diseases have not been shown to be transmitted between different organisms (Scheckel and Aguzzi, 2018). Prions are defined as pathogens that are transmissible only due to proteins (Zerr *et al.*, 2024). However, an involvement of circular DNA sequences called **S**low **P**rogressive **H**idden **IN**fections of variable (**X**) (Sphinx) has been discussed recently (Manuelidis, 2011).

The two Sphinx sequences were first found in samples of different TSE animal-models such as hamsters and mice infected with different scrapie strains and Creutzfeldt-Jakob disease (Manuelidis, 2011). Sphinx 1.76 (1.8 kb) and Sphinx 2.36 (2.4 kb) are circular DNAs enriched in infectious preparations of TSE-infected animal cell cultures compared to uninfected samples (Manuelidis, 2011). Both circular DNAs contain an open reading frame similar encoding a replication initiator protein (Fig. 1.3) (Rep protein) related to bacterial replication proteins. The replication of a range of bacteria plasmids as well as of several DNA viruses depends on a Rep protein binding to their DNA (Kilic et al., 2019). In case of Sphinx 1.76, the rep protein showed

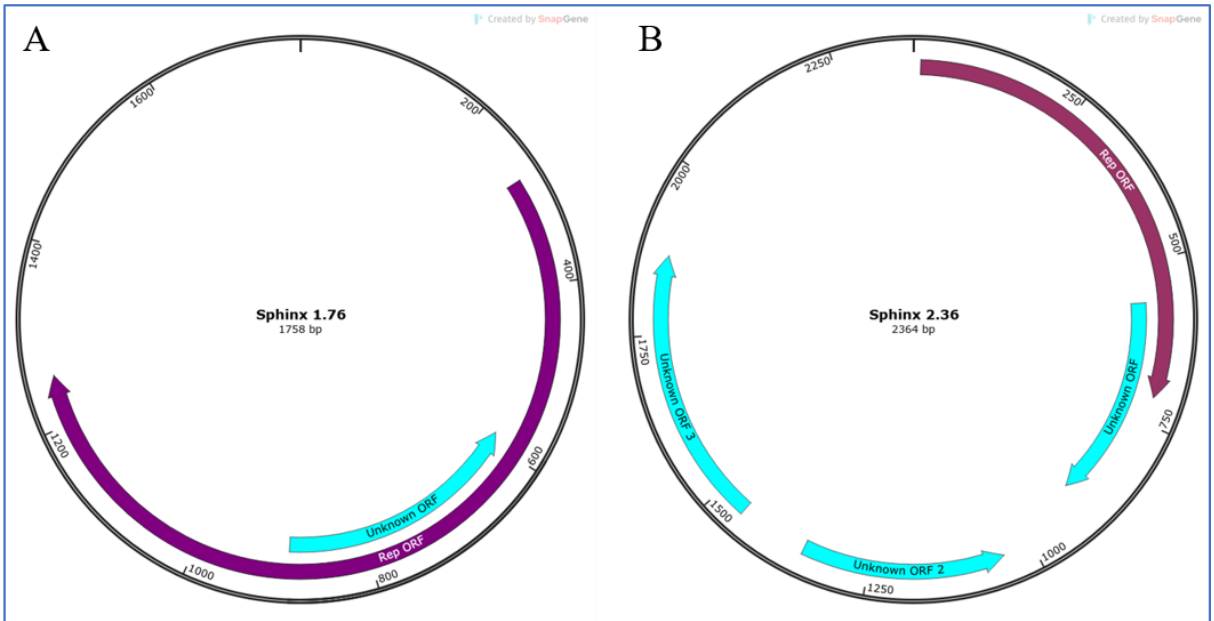


Fig. 1.3: Genome maps of Sphinx 1.76 and Sphinx 2.36: Map of circular DNA sequences Sphinx 1.76 (A) and Sphinx 2.36 (B). The Rep open reading frames are marked in purple, open reading frames with unknown function are pictured in cyan. The open reading frames were predicted using ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/), with the minimal ORF length set as 300 nucleotides, standard genetic code and ATG plus alternative start codons allowed. The sequence maps were generated using SnapGene Viewer (www.snapgene.com).

homologies to a Rep 3 superfamily protein of *Acinetobacter junii* and *Acinetobacter baumannii*

plasmids, whereas in case of Sphinx 2.36 a homologous region to a Rep 1 superfamily protein on a *A. baumannii* plasmid was found (Manuelidis, 2011). There are also additional, smaller open reading frames predicted for the Sphinx sequences, however their function is unknown so far (Fig. 1.3) (Manuelidis, 2011).

The Sphinx molecules co-purify with infectious TSE-particles in animal experiments (Manuelidis, 2011; 2013). Additionally, infectious particles were shown to be still infectious after removal or digestion of prion proteins (Botsios and Manuelidis, 2016). On the other hand, nuclease treatment of TSE-strains to destroy co-purifying DNAs decreased the infectivity by more than 99%. This indicates, that the investigated TSE-strains need nucleic acids for infectivity (Botsios and Manuelidis, 2016). However, since Sphinx molecules were also found in healthy cell culture and brain samples, they are also not solely responsible for infectivity (Botsios and Manuelidis, 2016; Manuelidis, 2019).

1.3.2 Isolation and characterization of BMMFs

BMMFs are episomal DNA molecules that were first isolated and described in 2014, when 11 Sphinx 1.76-related sequences were identified. Five of them were isolated from bovine serum samples, four from milk samples and two from a human MS brain sample (Whitley et al., 2014). These Sphinx 1.76-related sequences were later defined as BMMF group 1. Additionally, three Sphinx 2.36-related BMMFs were isolated from healthy bovine serum samples (Funk et al., 2014). Due to their similarities to Sphinx 2.36, these isolates were assigned to BMMF group 2. Besides of these two groups of Sphinx 1.76 or Sphinx 2.36-related molecules, there were also three sequences isolated that showed no close relation to the two previously described Sphinx-sequences, but to gemycircularviruses. Two of these sequences have been identified in healthy cattle serum samples (HCBI8.215, HCBI9.212), while the third was isolated from a MS serum sample (MSSI2.225) (Lamberto et al., 2014). Since these sequences differed from the sequences in the first two BMMF groups, they were classified as BMMF group 3 isolates. Furthermore, a BMMF distantly related to a *Psychrobacter* plasmid was isolated from a serum sample of a MS patient and named **m**ultiple **s**clerosis **s**erum **i**solate MSSI1.162 (Gunst et al., 2014). This isolate was later assigned to BMMF group 4.

In subsequent analyses in the following years, several other sequences related to either BMMF group 1 or 2 were identified in healthy cattle serum samples or commercially available cow milk as well as in human colorectal cancer or normal colon samples via laser microdissection (de Villiers et al., 2019; Falida et al., 2017). Currently, more than 150 BMMF isolates have

been described with the majority of them belonging to either BMMF group 1 or 2. All BMMF 1 and 2 sequences known so far contain an open reading frame (ORF) encoding for a highly conserved replication (Rep) protein, which also shows homologies to prokaryotic replication proteins (Fig. 1.3) (de Villiers *et al.*, 2019; zur Hausen, 2017a). X-ray crystallography structure analysis of the Rep protein of BMMF group 1 isolate H1MSB.1 showed high similarities to the structure of other bacterial Rep proteins (Kilic *et al.*, 2019). The BMMF1 Rep proteins are related to Rep superfamily 3. BMMF group 1 isolates share an iteron-like tandem repeat region upstream of the Rep protein. This repeat region is additionally preceded in most BMMF1 templates by a palindromic sequence, which constitutes a potential origin of replication (de Villiers *et al.*, 2019). In total, BMMF1 sequences are closely related to *A. baumanii* plasmids with about 70 % sequences identity (zur Hausen, 2017a). Most BMMF1 templates are highly conserved especially within the Rep ORF and range from 1.5 to 2.5 kb. However, there are also outliers within the BMMF1 group such as C1MIs.3M.1 or C1HB.4. C1MIs.3M.1 is the smallest BMMF group 1 member with 461 nt. C1HB.4 on the other hand is the largest BMMF group 1 isolate with almost 3 kb (Fig. 1.4). C1HB.4 contains not only a Rep ORF like all BMMF group 1 members, but also a second larger ORF, which shows similarities to an Acinetobacter mobilization protein and thus might serve for plasmid recombination (de Villiers *et al.*, 2019).

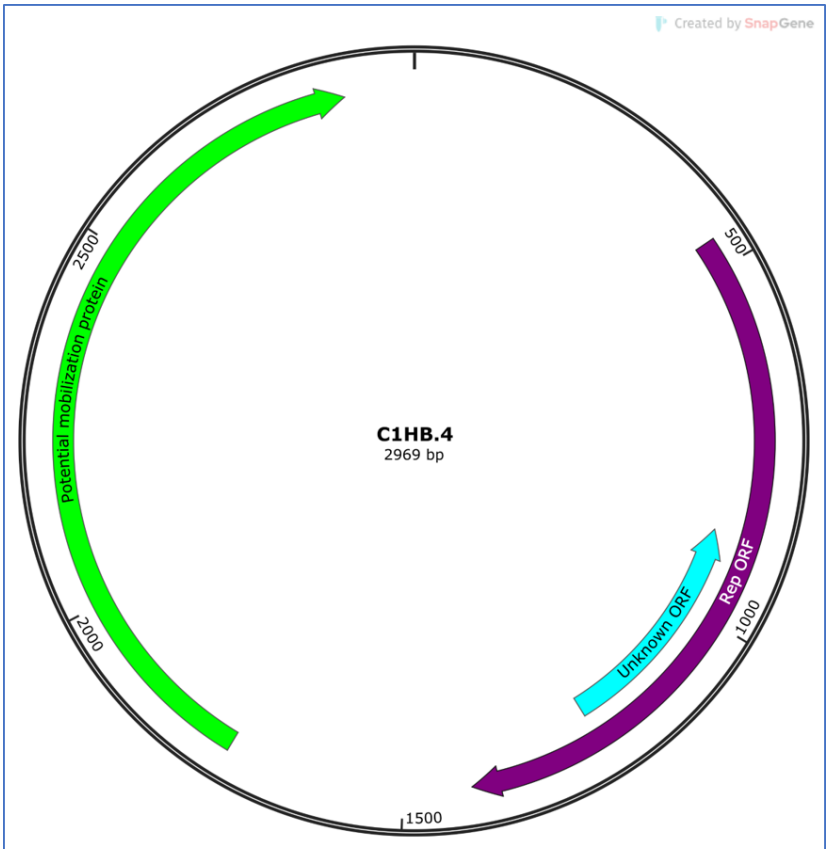


Fig. 1.4: C1HB.4 map: Map of BMMF group 1 sequence C1HB.4 (cattle group 1 healthy bovine isolate 4). The Rep open reading frame is marked in purple, the open reading frame with unknown function is pictured in cyan and the putative mobilization protein is shown in green. The open reading frames were predicted using ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/), with the minimal ORF length set as 300 nucleotides, standard genetic code and ATG plus alternative start codons allowed. The sequence maps were generated using SnapGene Viewer (www.snapgene.com).

BMMF group 2 isolates have additional, smaller ORFs besides of the Rep ORF (Fig. 1.3 B), however the precise functions of the proteins potentially transcribed from these ORFs are not known so far (de Villiers *et al.*, 2019; zur Hausen, 2017a). Just as BMMF1 sequences, BMMF2 templates are also

related to *A. baumanii* plasmids. BMMF group 2 Rep proteins belong to Rep superfamily 1 or 2. In contrast to BMMF group 1, the Rep ORFs are not preceded by an iteron-like tandem repeat region. While most BMMF2 sequences range from 2.1 to 3.1 kb, there are additionally several small BMMF group 2 isolates with less than 1 kb. These small BMMF genomes might act as satellites to the larger BMMF genomes (de Villiers *et al.*, 2019).

Transfection experiments with four different BMMF group 1 isolates, revealed that BMMFs are transcriptionally active, express their Rep protein and are able to replicate in human cell cultures. Human cell culture experiments also suggested that BMMF expression affects the gene expression patterns of the human host cells (Eilebrecht et al., 2018). Additionally, antibodies against a BMMF Rep protein were detected in two out of 30 human plasma samples, which supports a human immune response to BMMFs and thus indicates the potential immunogenicity of BMMFs (Eilebrecht *et al.*, 2018).

The origin and evolution of BMMF sequences are unknown so far, however BMMFs share common features with both bacterial plasmids and viruses. The sequences of BMMF1 and BMMF2 templates show homologies to *A. baumanii* plasmids (de Villiers *et al.*, 2019; Manuelidis, 2011). The Rep proteins encoded by BMMF group 1 and 2 isolates exhibit sequences similarities and structural homologies to bacterial Rep proteins (de Villiers *et al.*, 2019; Kilic *et al.*, 2019; Manuelidis, 2011). The BMMF1 iteron-like tandem repeat region upstream of the Rep ORF additionally resembles bacterial structures. The palindromic sequences of BMMF1 isolates are however similar to the origin of replication of a range of circular single stranded DNA viruses (de Villiers *et al.*, 2019).

1.3.3 Isolation of BMMFs from non-taurine origin

Recently, milk samples of other, non-taurine cattle species were analyzed for the presence of BMMFs as well. In a study of milk samples of two water buffalo herds domiciled in Germany, full-length circular DNA sequences could be found in 53 % of the samples (König et al., 2021a). 20 of the 21 sequences identified in these samples could be assigned to BMMF group 1 and 2 with nine of these sequences sharing up to 100 % sequence identity to previously described BMMF templates. One sequence was classified as BMMF group 3 isolate (König *et al.*, 2021a). Water buffaloes are rarely present in Europe or North America, where bovine meat and milk production is mostly based on *Bos taurus*-derived cattle, however they are used for milk production in several Asian countries, most prominently in India (König *et al.*, 2021a). In addition to milk samples of European aurochs-derived taurine cattle and water buffaloes, recent

publications also include sheep and goat milk in the investigation for BMMFs. Just as *Bos Taurus*-derived cattle and water buffaloes, goat and sheep belong as well to the *Bovidae* family, albeit not to the *Bovinae* sub-family (König *et al.*, 2021a; b). The analysis of 73 sheep milk samples from five different farms and 40 goat milk samples from three different farms as well as of six commercially available sheep and goat milk samples each, resulted in the identification of six BMMF1-related circular DNAs (König *et al.*, 2021b).

Besides of milk samples, fecal and blood samples of yaks, zebus, water buffaloes and watusi cattle living in German zoos were analyzed for BMMFs (König et al., 2023). Nine BMMF1-related and 14 BMMF2-related circular DNA-sequences were detected in these samples. Interestingly, the majority of these isolates was found in fecal samples. Consequently, BMMFs are present in a wide range of different species part of the *Bovidae* family. This includes not only European-based cattle, goats and sheep, but also African and Asian cattle species (de Villiers *et al.*, 2019; König *et al.*, 2023; König *et al.*, 2021a; b). Furthermore, recent studies examine samples of food of non-bovine origin for detection of BMMF DNA fragments. Such fragments were isolated from different sources including meat, seafood, dairy products, vegetables, fruits and grain (Pohl et al., 2022). Besides of food samples, BMMF fragments have also been found in saliva samples of pigs and feces of pigs and chicken (Pohl *et al.*, 2022). In a follow-up publication 16 full-length BMMF sequences of either BMMF group 1 or 2 were isolated from meat, seafood, fruit and vegetable samples (Habermann et al., 2023). Consequently, BMMFs can be detected at a range of different food and animal samples not only confined to the *Bovidae* family. Since the origin and transmission of BMMFs are not clearly known so far, it cannot be determined for sure, if BMMF sequences are generally widely distributed in the environment or if they are introduced to food products during processing and storage steps (König *et al.*, 2023; Pohl *et al.*, 2022). However, the detection of BMMFs in animal milk, serum and fecal samples as well as in human tissue samples, indicates an uptake of BMMFs by different living organisms via nutrition as well as a link to the digestive system and milk production.

1.3.4 Putative role of BMMFs in carcinogenesis

Oncogenic infectious agents can have latencies ranging from several years to decades after the first infections (zur Hausen and de Villiers, 2014; zur Hausen, 2017a). In case of BMMFs, the first contact likely happens early in life due to the consumption of bovine meat and dairy products, since the isolation of BMMFs from milk and bovine serum samples indicated a presence of BMMFs in these products (zur Hausen, 2017a). The first infection at a young age

leads to an activation of the immune system due to the BMMF molecules (zur Hausen, 2017a). While in case of colorectal cancer BMMFs reach the affected tissue directly via food consumption, BMMFs might travel to other target tissues via blood, lymphatic system or aerosols in case of lung cancer (zur Hausen *et al.*, 2019; zur Hausen, 2017a). The consumption of either red meat or milk has not only been discussed in the context of different cancer types such as colorectal cancer, but also in relation to several other diseases such as a group of neurodegenerative diseases including TSE, MS, Parkison's disease and Alzheimer's disease as well as several cardiovascular and autoimmune diseases (Morin *et al.*, 2024; zur Hausen, 2017a). Since two BMMF isolates originate from the brain lesion of a MS patient, BMMFs might also play a role in MS, for which milk has been repeatedly discussed as potential risk factor (Morin *et al.*, 2024; zur Hausen, 2017a).

The role of BMMFs in cancer was investigated by using mouse monoclonal antibodies against the Rep protein of the BMMF 1 genome H1MSB.1 to study the presence and impact of BMMFs in cancer tissue samples. Analyses of colorectal cancer samples of tumor and peritumor tissue showed BMMF1 Rep detection in the peritumor using immunohistochemistry (IHC), but not in the tumor samples (Bund et al., 2021). Furthermore, BMMF DNA was successfully isolated from peritumor samples by laser microdissection, whereas no DNA could be retrieved from paired tumor tissue samples (Bund *et al.*, 2021). These findings indicate, that BMMFs are present in the peritumor of colorectal cancer patients, but not or at lower frequencies in paired tumor samples. Subsequent IHC analyses of colorectal tissue samples confirmed that the BMMF1 Rep protein is expressed in significantly more cells in peritumor tissue compared to not only tumor samples but also to healthy colon samples of age-matched donors (Nikitina et al., 2023c). Looking at different developmental stages of colorectal cancer, Rep expression was detected in the adjacent tissue of low-grade dysplasia, high-grade dysplasia and colorectal cancer in higher levels than in healthy colorectal samples (Nikitina *et al.*, 2023c). The higher Rep-positivity in adjacent tissues of different stages of the development of colorectal cancer compared to healthy control samples supports a role of BMMF in carcinogenesis of colorectal cancer.

Further IHC analysis of Rep antibody staining in combination with CD68 staining, revealed a colocalization of Rep with $CD68^+$ macrophages in the cytoplasm of cells of the lamina propria in both peritumor samples of colorectal cancer patients and in control tissues of healthy donors (Bund *et al.*, 2021; Nikitina *et al.*, 2023c). However, peritumor tissues showed significantly higher levels of combined Rep and $CD68^+$ positivity compared to cancer-free tissue samples

(Bund *et al.*, 2021). Higher Rep and CD68-positivity in the peritumor were additionally reported to be linked to worse survival outcomes, however the correlation was not strong enough to be statistically significant (Nikitina *et al.*, 2023c). There was no correlation between Rep- and CD68-positivity in tumor tissues, since Rep expression and Rep-positivity in macrophages decreased in the dysplastic tissues from low-grade to high-grade dysplasia to colorectal cancer, where barely any Rep expression could be detected (Nikitina *et al.*, 2023c). Since IHC staining of $CD3^+$ T-cells and $CD20^+$ B-cells did not show an association of T- and B-cells with Rep enriched regions, the BMMF1 Rep localization seems to be specifically linked to $CD68^+$ macrophage infiltration (Bund *et al.*, 2021). Increased presence of macrophages is characteristic for local chronic inflammatory processes. Consequently, the colocalization of BMMF1 Rep positivity with macrophages indicates an association of BMMFs with chronic inflammation (Bund *et al.*, 2021; zur Hausen *et al.*, 2019). This link was confirmed by the detection of increased levels of DNA oxidation products caused by radicals in Rep-enriched regions via IHC staining (Bund *et al.*, 2021). These results show the increased presence of reactive oxygen and nitrogen species in the Rep-positive foci. Due to their close proximity to the epithelial cells of crypts of Lieberkühn in the colon, these rapidly replicating cells are also subjected to radicals, which after longer periods of exposure results in the accumulations of DNA mutations (Fig. 1.5) (Bund *et al.*, 2021; zur Hausen *et al.*, 2019).

In conclusion, BMMFs are expected to be consumed via bovine meat and milk products starting from an early age. The presence of BMMFs is proposed to induce chronic inflammation in exposed tissues, which is accompanied by the infiltration of macrophages and production of reactive oxygen and nitrogen species (ROS/RNS) (Fig. 1.5). The adjacent cells are consequently subjected to diffusing radicals, which results in the accumulation of mutations in tissues exposed to BMMFs. After longer periods of latency, the continuous exposure to BMMFs leads to malignant transformations in the affected cells and to the formation cancer progenitor cells (Fig. 1.5) (de Villiers and zur Hausen, 2021). In case of colorectal cancer this causes the development of polyps, which will eventually develop into colorectal cancer (Bund *et al.*, 2021; zur Hausen *et al.*, 2019). The lack of BMMF positivity in tumor samples compared to peritumor samples indicates that BMMF-promoted indirect carcinogenesis via chronic inflammation does not require the ongoing presence of the carcinogenic agents within or at the tumor cell (de Villiers and zur Hausen, 2021; zur Hausen *et al.*, 2019).

The proposed model of indirect carcinogenesis caused by chronic inflammation in BMMF-exposed tissues also could explain why long-term intake of NSAID reduces the risk for several

cancer types including colorectal, stomach, breast, lung, ovarian, esophagus and prostate cancer (Rothwell *et al.*, 2011; Todoric *et al.*, 2016; zur Hausen *et al.*, 2019).

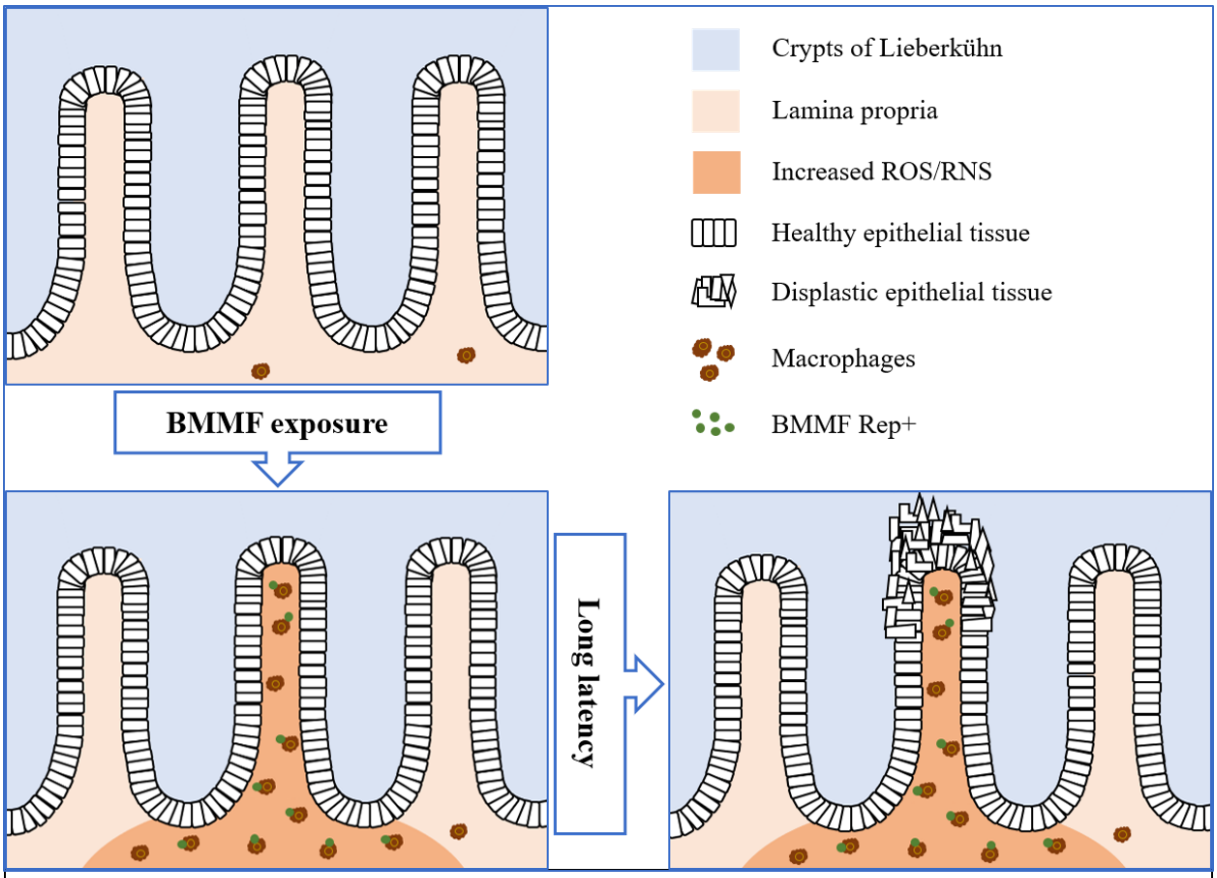


Fig. 1.5: Modell of BMMF-mediated indirect carcinogenesis: Schematic overview of indirect cancer modell and the perceived effect of the presence of BMMFs in colorectal tissue. The exposure of previously healthy tissue to BMMFs results in the accumulation of macrophages and reactive oxygen and nitrogen species. These radicals diffuse and cause mutations in the DNA of cells in proxmity to the BMMF exposed tissue. After long periods of latency, the accumulation of mutation leads to displastic transformation of the respective tissue region and eventually to cancer (figure based on/adapted from Nikitina et al., 2023, fig.5).

These cancer types linked with chronic inflammation consequently constitute the list of cancer types-of-interest for future BMMF research. Thus, subsequent IHC analyses of lung adenocarcinoma (LUAD) and pancreatic ductal adenocarcinoma (PDAC) tissue samples with BMMF1 Rep antibodies showed Rep positivity associated with $CD68^{+}$ macrophages (Nikitina et al., 2023b). Immunogold electron microscopy additionally yielded the detection of Rep-antibody marked vesicular structures in colorectal cancer, lung adenocarcinoma and pancreatic ductal adenocarcinoma (Nikitina et al., 2023a). Consequently, BMMF-mediated indirect carcinogenesis might not only play a role in colorectal cancer, but also in several cancer types linked to chronic inflammation.

1.4 Aims of the thesis

BMMFs are potential infectious agents, that have been identified in meat, milk and serum samples of animals belonging to the Bovidae family amongst others. Epidemiological observations support a connection between the incidence rates of certain cancer types and specific nutritional patterns – such as the consumption of bovine meat and milk. BMMFs might be the explanation for this, since they are consumed by human via food starting from an early age due to their presence in meat and milk products of these species. BMMFs have been proposed to indirectly promote carcinogenesis of a range of different cancer types after decades of latency by inducing chronic inflammation in exposed tissues. Studies about long-term use of NSAIDs showed that anti-inflammatory medication reduces the risk for several cancer types including colorectal, breast, lung, prostate, stomach, esophagus and ovarian cancer. Since these cancer types seem to have a joint link to chronic inflammation, they are consequently of interest for BMMF research. For several cancer types, there has been experimental evidence for the presence of increased amounts of BMMF Rep protein in peritumor tissue as well as for an association of the Rep protein with $CD68^+$ macrophages and increased presence of ROS/RNS in colorectal cancer patients when compared to healthy individuals. However not all cancer types potentially relevant for BMMF-facilitated carcinogenesis have been investigated in detail by wet-lab analyses. Additionally, the number of tissues samples available for experimental analysis of cancer types-of-interest is limited. If BMMFs indeed contribute to indirect carcinogenesis, BMMF RNA and DNA should be detectable in high-throughput sequencing data of tumor or preferably peritumor samples in cancer types linked to chronic inflammation. Consequently, *in silico* analysis of existing, publicly available sequencing data provides an opportunity to add information about cancer types not investigated so far and to expand the number of cancer tissue samples screened for the presence of BMMFs.

For this reason, I analyzed for this thesis publicly available high-throughput sequencing data of cancer sequencing projects such as The Cancer Genome Atlas (TCGA) or Pan-Cancer Analysis of Whole Genomes (PCAWG) for Bovine Meat and Milk Factors to support wet-lab results of BMMF detection in certain cancer types. Additionally, I screened sequencing data of healthy tissue samples provided by the GTEx project as case versus control comparison. In case of cancer types linked to chronic inflammation, elevated BMMF detection in cancer would support a contribution of BMMFs to indirect carcinogenesis. Analysis for BMMF detection on both RNA and DNA level might provide new insights in the transcription or bioactivity of BMMFs. This thesis aimed to screen a broad range of different cancer and tissue types for BMMF reads

to identify cancer types and tissues exposed to BMMFs to obtain a better understanding of the role of BMMFs in carcinogenesis and of the prevalence and spread of BMMFs in healthy tissues.

I also aimed to examine potential differences in detection levels between the four main BMMF groups as well as between the individual genomes within these groups to examine if the detection patterns point at putative high-risk BMMF groups and genomes. *In silico* analysis allows screening for a comprehensive library of BMMF sequence, whereas wet-lab analyses target only specific BMMF genomes, which are covered by primers and antibodies. A comprehensive wet-lab analysis of all of the more than 150 BMMF genomes described so far would be expensive and extremely time-intensive and thus hardly feasible. Consequently, the *in silico* analysis performed during this thesis aims to identify, which BMMF isolates are most frequently detected and thus relevant for in-detail experimental analyses performed by my wet-lab colleagues. Additionally, I aim to examine, if there are any specific BMMF groups, subgroups or genomes, which are characteristic for certain tissue or cancer types to provide the basis for future, more specific BMMF research both in wet-lab and *in silico.*

2. Methods

2.1 Software

All analyses were performed within the DKFZ cluster structure. Table 2.1 lists the software and pipelines used for data processing, analysis and visualization.

Tab. 2.1 Pipelines and software

Name	(Cluster-)Version	References
Anaconda	anaconda3/2021.05	Anaconda Software Distribution, 2021
AWK	GNU Awk 4.0.2	Aho, 1987
Bash	GNU bash, version 4.2.46(2)-release (x86_64-redhat-linux-gnu)	Ramey, 2003
BEDTools	bedtools/2.24.0	Quinlan and Hall, 2010
BioPerl	Bio 1.7.5	Stajich et al., 2002
BLAST	ncbi-blast/2.7.1	Altschul et al., 1990
bwa	bwa/0.7.15	Li and Durbin, 2009; Li, 2013
D-ViSioN	Updated version for LFS cluster	Horak et al., 2020
gen3-client	gen3-client 2022.04	https://gen3.org/resources/user/gen3-client/; https://github.com/uc-cdis/cdis-data-client/tree/1.2.1
Jalview	Jalview version 2.11.3.2	Waterhouse et al., 2009; Troshin et al., 2011; Troshin et al., 2018
Jupyter lab	jupyter lab, version 3.3.2	Kluyver, 2016
MEGA	MEGA version 11.0.13	Tamura et al., 2021
MUSCLE	https://www.ebi.ac.uk/jdispatcher/msa/muscle	Madeira et al., 2024
Nano	GNU nano version 2.3.1	https://nano-editor.org/docs.php
ORFfinder	Web version	https://www.ncbi.nlm.nih.gov/orffinder/
Perl	perl/5.20.2	https://www.perl.org/; Christiansen, 2012
picard	picard/1.61	https://broadinstitute.github.io/picard/

Pysam	pysam 0.15.2	https://github.com/pysam-developers/pysam?tab=readme-ov-file
Python	python/3.7.0	van Rossum, 2009
R	For D-ViSioN: R/3.3.1 For D-ViSioN output analysis + visualization: R/4.3.0 + R/4.4.0	R Core Team, 2024
samtools	samtools/1.6	Danecek et al., 2021
SnapGene Viewer	SnapGene Viewer version 7.2	SnapGene software (www.snapgene.com)
sratools	sratools/3.0.2	https://github.com/ncbi/sra-tools
STAR	STAR/2.5.3a	Dobin et al., 2013

2.2. R packages

The data presented in this thesis was visualized using R 4.4.0. The R packages used are listed in table 2.2.

Tab. 2.2 R packages

Name	Version	References
bio3d	bio3d 2.4-4	Grant et al., 2006
BiocParallel	BiocParallel 1.38.0	Morgan, 2024b
Biostrings	Biostrings 2.72.0	Pagès, 2024
Circlize	circlize 0.4.16	Gu et al., 2014
colorRamp2	colorRamp2 0.1.0	Gu, 2022
ComplexHeatmap	ComplexHeatmap 2.20.0	Gu et al., 2016
dplyr	dplyr 1.1.4	Wickham, 2023c
GenomicAlignments	GenomicAlignments 1.40.0	Lawrence et al., 2013
GenomicFeatures	GenomicFeatures 1.56.0	Lawrence et al., 2013
ggbreak	ggbreak 0.1.2	Xu et al., 2021
ggplot2	ggplot2 3.5.1	Wickham, 2016
ggpubr	ggpubr 0.6.0	Kassambara, 2023

ggtree	ggtree 3.12.0	Yu et al., 2016; Yu, 2023
patchwork	patchwork 1.2.0	Pedersen, 2024
readr	readr 2.1.5	Wickham, 2024
readxl	readxl 1.4.3	Wickham, 2023b
seqinr	seqinr 4.2-36	Charif, 2007
stringr	stringr 1.5.1	Wickham, 2023a
SummarizedExperiment	SummarizedExperiment 1.34.0	Morgan, 2024a
tibble	tibble 3.2.1	Müller, 2023
tidytree	tidytree 0.4.6	Yu, 2023
treeio	treeio 1.28.0	Wang et al., 2020; Yu, 2023
ZIR	ZIR_1.0.0	Wang, 2021, https://github.com/PennChopMicrobiomeProgram/ZIR

2.3 D-ViSioN

The D-ViSioN pipeline (**D**etection of Integrated **Vi**ral Sequences by **Si**nglet**on**s) was designed by Dr. Gnana Prakash Balasubramanian to detect viral reads and integration sites in high-throughput sequencing data (Horak *et al.*, 2020). In the following, the D-ViSioN pipeline was updated by Hamza Khan and Kai Horny (Horny, 2018). Finally, D-ViSioN was further adapted by me during my master thesis for using different input data formats as well as for performing D-ViSioN analyses within the ODCF LSF cluster structure of the German Cancer Research center (DKFZ) (Häfele, 2020). D-ViSioN uses two subsequent alignment steps performed by BWA mem and BLAST. The fast BWA mem alignment is performed first and the reads extracted from the BWA mem output are submitted as input to the more sensitive second alignment step conducted via BLAST (Fig. 2.2) (Horny, 2018).

In this thesis, I applied D-ViSioN to detect BMMF reads in bulk and single cell RNA sequencing as well as in whole-genome sequencing (WGS) data. For this, D-ViSioN was provided with a library of BMMF sequences (see chapter 2.3.1.). I analyzed single-end sequencing data and paired-end RNA sequencing data using the “Infectome Mode” of D-ViSioN to extract reads mapping to the BMMF database (Fig. 2.1). In case of paired-end WGS

data I combined the “Infectome Mode” with the “Integrome Mode” (Fig. 2.1). The integrome mode relies on the extraction of so-called singletons for the investigation of integration events (Fig. 2.1).

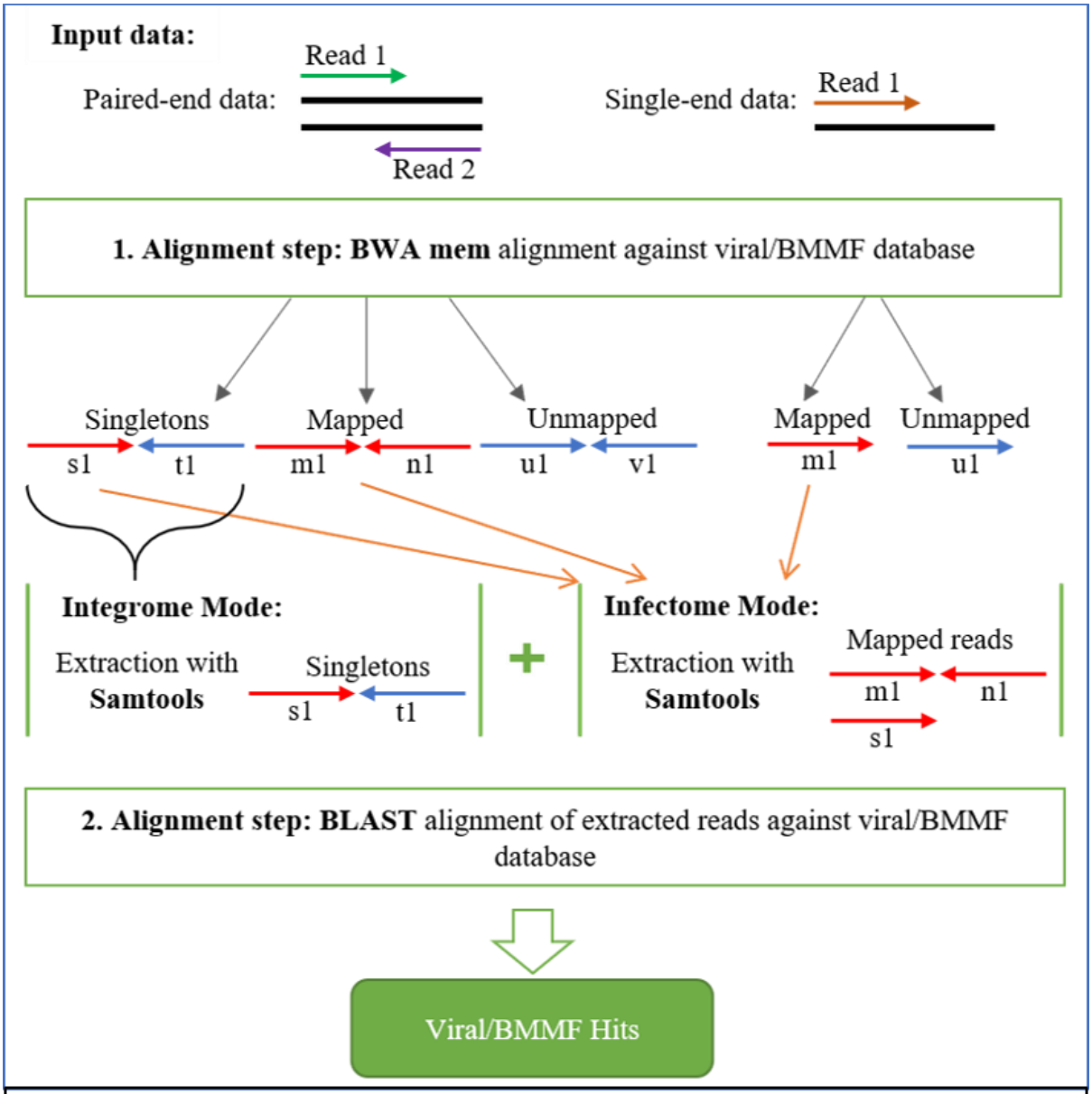


Fig. 2.1: D-ViSioN algorithm: The D-ViSioN pipeline detects viral reads and integration sites. D-ViSioN uses BWA mem and BLAST for two subsequent alignment steps to screen input sequencing data for viral/BMMF sequences provided in a library. The integrome mode extracts singletons to identify integration sites. The infectome mode extracts all reads mapped to the viral/BMMF database to detect the provided sequences in the input data. The integrome and infectome mode can be performed in parallel in the same run. D-ViSioN provides additional modes applying verification steps against the human genome or a nucleotide database. Since these steps were not performed in the analyses, the verification modes are not included in this schematic description of the algorithm (Horak et al., 2020; Horny, 2018).

Singletons are in this context defined as read-mate pair, where one read maps to the BMMF

library, whereas the matching read is aligned to the human reference genome (Fig. 2.2). The combination of both D-ViSioN modes thus enables the detection of BMMF reads as well as of integration sites in parallel in one run. Since BMMFs are not expected to integrate into the human genome, the main focus in this thesis will be on the output of the BMMF read detection of the infectome mode.

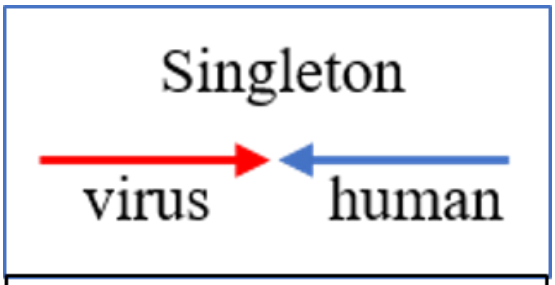


Fig. 2.2: Singleton: Read pair in which on read maps to a viral sequence and the mate mappes on the human genome.

2.3.1 BMMF Library

The BMMF library I used for the D-ViSioN analyses was provided by Dr. Timo Bund from the "Episomal-persistent DNA in Cancer- and Chronic Diseases" division at the DKFZ. This library contains 174 sequences which range from a length of 461 nt to 3090 nt. Besides of 167 sequences, that have been isolated and identified as BMMF sequences, the library also contains 6 BMMF-like plasmid sequences and one control sequence unrelated to BMMFs (Ac.Bau.DS002.16S.rRNA). Of the 6 BMMF-like plasmids, one sequence can be added to BMMF group 1, whereas the other five sequences can be assigned to BMMF group 2. In total, the library contains 57 BMMF group 1 sequences, 112 BMMF group 2 sequences, 3 BMMF group 3 sequences and one BMMF group 4 sequence. The sequences included in the BMMF library are listed with their sequence lengths, their origin and their GenBank accession number in supplementary table S1.

2.3.2 Processing and analysis of D-ViSioN results

D-ViSioN reports the detected reads after the second alignment step in different output files. While there are results summary files such as the file "BMMF_result_table_transposed.txt", this files just lists the names of the detected sequences of the submitted library together with the number of reads detected for this sequence. However, for a more detailed analysis of the D-ViSioN output, there is additional information needed such as the sequence identity between the detected reads and the library entry, the alignment length and positions. This missing information can be obtained from an intermediate output file called "BMMFf1_bestHits.out". This, as well as all of the subsequent output analysis and processing steps are conducted in R. The sequences of the detected reads can be retrieved from the file "BMMFm1n1s1.fasta" using the read identifiers. In case of the PCAWG WGS data, there are sometimes read identifiers with

more than one read assigned to it. Consequently, these “fake duplicates” have to be removed from further D-ViSioN output analysis steps, since not all reads assigned to the same identifier always match the BMMF library.

For the D-ViSioN output analysis, I decided to introduce a so-called “3-hits threshold” to filter out samples with a very low number of BMMF hits detected. According to this threshold, samples are only defined as BMMF-positive if at least 3 BMMF hits of the same BMMF group have been detected for this sample. This threshold is supposed to ensure, that only samples with a reliable BMMF signal are taken into account for further analysis steps. In subsequent analysis steps the BMMF positivity is also examined at BMMF group level and split between the four different BMMF groups. In this case, the 3-hits threshold is applied to all subgroups separately to determine if a sample is positive for the respective subgroup.

For further verification of the detected reads, the alignment positions of the reads are used to mapped the reads to the library entry they are assigned to. The *circlize* R package was used to visualize the alignments of the detected reads to the BMMF library sequences. The *ORF finder* webtool was used to obtain start and end positions of predicted open reading frames for BMMF templates-of-interest. For this, ORF finder was run with the minimal ORF length set to 150 nt using the standard genetic code and allowing “ATG” and alternative initiation codons as start codons while ignoring nested ORFs. The predicted open reading frame positions were also visualized using *circlize* to compare their location to the detected reads. Reads mapping to so-called peak regions covered by high numbers of BMMF reads, were additionally analyzed by aligning them to the respective library sequence using MUSCLE.

2.3.3. Phylogenetic analysis of BMMF library

An additional problem occurred during the assignment of reads to specific entries of the BMMF library. The results summary files only report the top hit of the output files of the BLAST alignment step. However, due to the high sequence identity between many of the BMMF isolates included in the BMMF library, detected reads can in many cases not be unambiguously assigned to a specific BMMF sequence. Frequently, the extracted BMMF reads map to more than one BMMF isolate with a very high – or even up to 100 % – sequence identity. To address this problem, I performed a phylogenetic analysis of all BMMF group 1 and BMMF group 2 sequences to split these two big BMMF groups into subgroups. For this, I applied the MEGA 11 software. First, I used MEGA to align the sequences within those two BMMF groups with a MUSCLE alignment performed with the standard settings suggested by MEGA. Second, I

used the alignment result to construct an UPGMA tree with 1000 bootstrap replications and a maximum composite likelihood nucleotide substitution model. The condensed bootstrap consensus tree is calculated with a cut-off value of 50 %. The timetree for BMMF group 1 is determined using the BMMF group 2 isolate Sphinx 2.36 as outgroup, whereas for BMMF group 2 the BMMF group 1 isolate Sphinx 1.76 is introduced as outgroup. The MEGA output is visualized in R using the packages *ggplot2, treeio, ggtree* and *tidytree*. Based on the created phylogenetic timetrees, I decided to split BMMF group 1 in 11 subgroups and BMMF group 2 in 12 subgroups. The two small BMMF1 templates C1MIs.3M.1 and C1HB.6.2 as well as the six small BMMF2 templates C2MI.10As.1, C2MI.5As.1, C2MI.7As.1, C2MI.9As.2, C2MI.9Bs.4 and C2MI.9Bs.6 were excluded from the phylogenetic analyses, since the size difference between these isolates and the larger BMMF genomes leads to unreliable alignment results. Consequently, the small isolates were later manually assigned to subgroups based on Nucleotide BLAST results and MUSCLE alignments to the larger genomes included in these subgroups. To investigate the similarity and sequence identity between the genomes of the different BMMF subgroups, I exported the MUSCLE alignments of MEGA for both BMMF1 and BMMF2 and calculated the sequence identity in R using *bio3d*. I visualized the calculated sequence identities using *ComplexHeatmap*.

In the next step, I used the BMMF subgroups to assign the detected reads to these subgroups. For this, the BLAST hits with the lowest e-value and the highest sequence identity are taken into account. For each read these top-ranked BLAST hits are examined to determine, if the BMMF sequences listed among the best BLAST hits belong to the same BMMF subgroup. If all best matches comprise BMMF sequences within one subgroup, the read gets assigned to this subgroup. If a read cannot be unambiguously assigned to one subgroup, the read gets classified as “unclear”.

2.3.4 Analysis of sequencing depth and normalization

I determined the sequencing depth using samtools. The samtools view command was used to count (-c) the reads of the sequencing data files. I focused on the mapped read only, when determining the sequencing depth. Consequently, the -F 0x4 flag was set to exclude unmapped reads from this. This was performed for all samples, besides of the DepMap cohort and the TCGA COAD data from the legacy TCGA portal. Here, I only determined the sequencing depth for samples passing the 3-hits threshold to be defined as BMMF positive. I assessed the impact of the sequencing depth, by comparing the sequencing depth of BMMF negative and BMMF positive samples. I generated Q-Q plots and histograms in R, to determine if the sequencing

depth of the investigated samples is normally distributed. In the next step, I compared positive and negative samples using the Mann-Whitney-U-Test in R. Since some of the data sets showed differences between the sequencing depth in positive and negative samples, I used the sequencing depth to normalize the BMMF hits detected in different samples and cohorts.

For the normalization of the BMMF reads detected for each sample using D-ViSioN, the number of BMMF reads was normalized against the total number of mapped reads. While usually the scaling factor 10^6 is used to calculate the normalized counts/reads per million reads, I accounted for the low hit numbers typically found for BMMFs and instead calculated counts/reads per billion reads with a scaling factor of 10^9 (formula I).

I: $$Reads\ per\ billion\ reads = \frac{Number\ of\ BMMF\ reads\ detected\ for\ sampe * 10^9}{Total\ number\ of\ mapped\ reads}$$

In the next step, the BMMF reads per billion mapped reads were normalized for cohort size by dividing by cohort size and multiplying with factor 100 to calculate the BMMF reads per billion mapped reads for a cohort size of 100 samples. The normalized read numbers obtained with this normalization method were subsequently used to compare the levels of BMMF detection between different datasets and cohorts. The statistical comparisons are conducted using the Zero-Inflated Rank (ZIR) Test provided by the ZIR R package (Wang, 2021). The ZIR test is a modified Wilcoxon rank sum test designed for comparing data containing a high number of zeros and adapted for both equal and unequal sample sizes (Wang, 2021).

2.4 Sequencing data

The input data for the D-ViSioN algorithm was derived from four different large-scale sequencing projects such as PCAWG, TCGA, GTEx and DepMap and one single cell dataset published in 2020 by Maynard and colleagues (Maynard et al., 2020).

2.4.1 PCAWG

The Pan-Cancer Analysis of Whole Genomes (PCAWG) project has been established to follow-up and build on the previous large cancer sequencing studies of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). The PCAWG consortium aimed to further investigate the features and mutations cancer cells use to proliferate and to escape cellular control mechanisms and the immune system (The ICGC/TCGA Pan-Cancer Analysis

of Whole Genomes Consortium, 2020). For this thesis, I analyzed PCAWG RNA and whole genome sequencing samples downloaded to the DKFZ ODCF cluster. Of 1009 RNA sequencing data samples available there, I screened 987 samples of 25 different cancer types using D-ViSioN and R as described in chapter 2.3. The remaining samples were omitted from further analysis, since they were not included in the samples deposited at the ICGC Data Portal (https://dcc.icgc.org/) – most likely they were excluded during quality assurance steps due to lack of sample quality (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). The cancer cohorts and the number of samples included in the PCAWG RNA sequencing data analyzed are listed in the supplementary table S2.

In case of the PCAWG WGS data stored at the DKFZ ODCF cluster, I focused on cohorts of cancer types and tissues relevant for BMMF research. In total I analyzed 3611 PCAWG WGS samples of 29 different cancer cohorts using D-ViSioN and R (see 2.3). For the majority of the 1792 donors these samples were derived from, there is both one tumor and one normal tissue sample included in the analysis. However, in case of some sub-cohorts such as the UK prostate cancer cohort (PRAD-UK) there are more than two sample files available for the same patient. Consequently, the number of tumor files analyzed is not always identical with the number of normal tissue files analyzed. Additionally, not all files available could be successfully analyzed. The Japanese liver cancer cohort LIRI-JP included several samples, that could not be successfully analyzed using D-ViSioN in repeated attempts and that were consequently excluded from further analyses.

1820 of the screened samples are tumor samples, which are mostly classified as primary tumor samples from either solid tumor, bone marrow, peripheral blood or lymph nodes. 3.4 % of all tumor samples analyzed are derived from either recurrent or metastases, whereas one single sample was taken from a tumor-derived cell line. In addition to the tumor samples, I also analyzed 1792 normal tissue samples. 72.5 % of the normal tissue samples are blood-derived samples, whereas 21.6 % of the normal tissue samples are derived from solid tissue distant to the primary tumor (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). Only 1.7 % normal tissue samples are originating from tissue adjacent to the primary tumor. The remaining normal tissue samples are taken for example from bone marrow, lymph nodes and EBV-immortalized cell lines. The different cancer cohorts included in the WGS primary and normal tissue data analyzed are listed in detail in supplementary table S3.

2.4.2 TCGA

The Cancer Genome Atlas (TCGA) project is an American cancer genomics projects conducted as cooperation of the National Cancer Institute and the National Human Genome Research Institute (Hutter and Zenklusen, 2018). The TCGA project is one of the large-scale cancer sequencing studies investigating the landscape of molecular changes and aberrations in cancer preceding the PCAWG project (The Cancer Genome Atlas Research Network et al., 2013; The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020). For this thesis, I analyzed RNA sequencing data of the TCGA project downloaded and stored at the DKFZ ODCF cluster. The available data included breast cancer, colorectal cancer, liver cancer, lung adenocarcinoma and pancreatic cancer RNA sequencing data (supplementary materials, S4). In total I analyzed 2969 TCGA RNA sequencing samples as described in chapter 2.3 (S4).

While the data of four of the five cancer cohorts were obtained from the harmonized TCGA data, the colorectal cancer data available and downloaded to the DKFZ ODCF cluster was downloaded from the legacy portal. The harmonized and legacy TCGA data differ regarding the human reference genome used for alignment (Gao et al., 2019). The TCGA legacy data was generated using GRCh37, whereas the harmonized data set was aligned against GRCh38 (Gao *et al.*, 2019). Besides of the reference genome, the sequencing techniques and bioinformatics workflows progressed between the generation of the TCGA legacy and harmonized data. Consequently, the differences in data generation and processing can generate a bias, especially in terms of gene counts (Gao *et al.*, 2019). Comparisons between the two data sets showed, that the differential gene expression increases from the legacy to the harmonized data set, whereas the relative detection and the ranking of the detected genes was mostly preserved between both analysis methods (Gao *et al.*, 2019).

2.4.3 GTEx

The Genotype Tissue Expression (GTEx) project intends to study genetic variants and their impact on gene regulation and expression in different tissues (The GTEx Consortium, 2013). The postmortem donors included in the GTEx project died aged between 21 and 70 and underwent autopsy for samples collection within 24 h of death (The GTEx Consortium, 2013; 2020). Since the main objective of GTEx was to collect “healthy” samples without a disease-background, individuals with certain conditions such as specific viral infections, metastatic cancer or cancer treatment within the past 2 years are excluded from the data set (GTEx Consortium, 2017; The GTEx Consortium, 2013). For this thesis, I screened 5561 GTEx paired-

end RNA sequencing samples of 15 different cohorts derived from 12 different tissues (supplementary materials, S5). The GTEx project aims to have an average sequencing depth of 50 million aligned reads per RNA sequencing sample (The GTEx Consortium, 2013). I used the gen3-client to download the GTEx samples of tissues-of-interest to the DKFZ ODCF cluster after data access was obtained via dbGaP. I downloaded the gen3-client with help of a conda environment, which was subsequently used for GTEx data download. To start the GTEx download an up-to-date credentials file generated with API is required. The API key has to be renewed every month. Before starting the data download, a manifest with information about the files-of-interest needs to be generated. Together with the credentials key, the manifest is used to start the download. After the download is completed, the md5 sums were calculated and the data was structured as D-ViSioN input data using softlinks. For several steps of the GTEx download, I applied scripts written by Andrej Vondran (Vondran, 2022). Subsequently, the GTEx RNA sequencing data samples were analyzed using the D-ViSioN pipeline as described in chapter 2.3

2.4.4 DepMap

As control to the cancer sequencing data sets, I analyzed sequencing data of cell lines obtained from the Cancer Dependency Map (DepMap) project (Tsherniak et al., 2017). The DepMap project is a successor of the Cancer Cell Line Encyclopedia project, which collected and characterized a broad library of cancer cell lines used to model different cancer types in scientific research (Barretina et al., 2012). The DepMap project aims to build on the Cancer Cell Line Encyclopedia project to identify genetic features linked to the proliferation and survival of cancer cell lines (Tsherniak *et al.*, 2017). For this purpose, DepMap provides an updated and processed version of the cell line library. I used D-ViSioN to analyze WGS data of 256 cell lines of 22 different cancer sites part of the DepMap project. The cancer sites and the number of cell line covering them are listed in supplementary table S6. The D-ViSioN pipeline runs and the subsequent output analyses were performed as described in chapter 2.3.

2.5.5 Single cell data

The single cell RNA sequencing data set I analyzed with D-ViSioN is publicly available under the NCBI BioProject accession number PRJNA591860 and as study SRP238929 via the SRA Run selector (Maynard *et al.*, 2020). The single cell data set comprises more than 23000 cells derived from 49 biopsies of 30 patients with metastatic lung cancer (Maynard *et al.*, 2020). The samples originate not only from primary tumors, but also from metastases and in three cases

from tumor-adjacent tissue. Consequently, the biopsies were not only taken from lung tissue, but also from tissues with metastases including pleura, brain, liver, lymph node and adrenal tissue (Maynard *et al.*, 2020). I used the SRA Run Selector to obtain the list of accession numbers of all cells included in this study as well as the metadata of the data set. In the next step, I downloaded the sequence read archive data (SRA) and extracted them in fastq format using the SRA Toolkit version sratools/3.0.2 installed at the cluster. Subsequently, I analyzed the single cell data with D-ViSioN as described in chapter 2.3. The metadata table as well as the scripts published part of the publication by Maynard et al., 2020 were used to identify the biopsy sites and the cell types of BMMF positive cells.

3. Results

3.1 Databases and samples analyzed using D-ViSioN

For the examination of the BMMF signal in a range of different cancer types and tissue sites, I analyzed data from the PCAWG, TCGA, GTEx and DepMap projects. In my analyses I divide these projects into 6 different main data sets – 3 cohorts with RNA data: PCAWG RNA, TCGA RNA, GTEx RNA – and 3 cohorts with WGS data: PCAWG WGS tumor, PCAWG WGS normal tissue/blood and DepMap WGS. The TCGA and PCAWG projects provide RNA tumor samples of cancer patients. Additionally, the PCAWG project also includes WGS data of both tumor samples and matching normal tissue/blood samples, which allows the comparison of the BMMF detection in a cancer sample and a non-cancer sample for the same patient. Since the RNA data sets do not include normal tissue samples of the same patients, RNA data of the GTEx project was analyzed for a case versus control comparison. The GTEx project provides a data set of healthy tissue data of donors, who died without the context of an acute disease (GTEx Consortium, 2017; The GTEx Consortium, 2013). Finally, I analyzed WGS data of cancer cell lines collected by the DepMap project for BMMFs to estimate the frequency of contaminations or endogenous positivity.

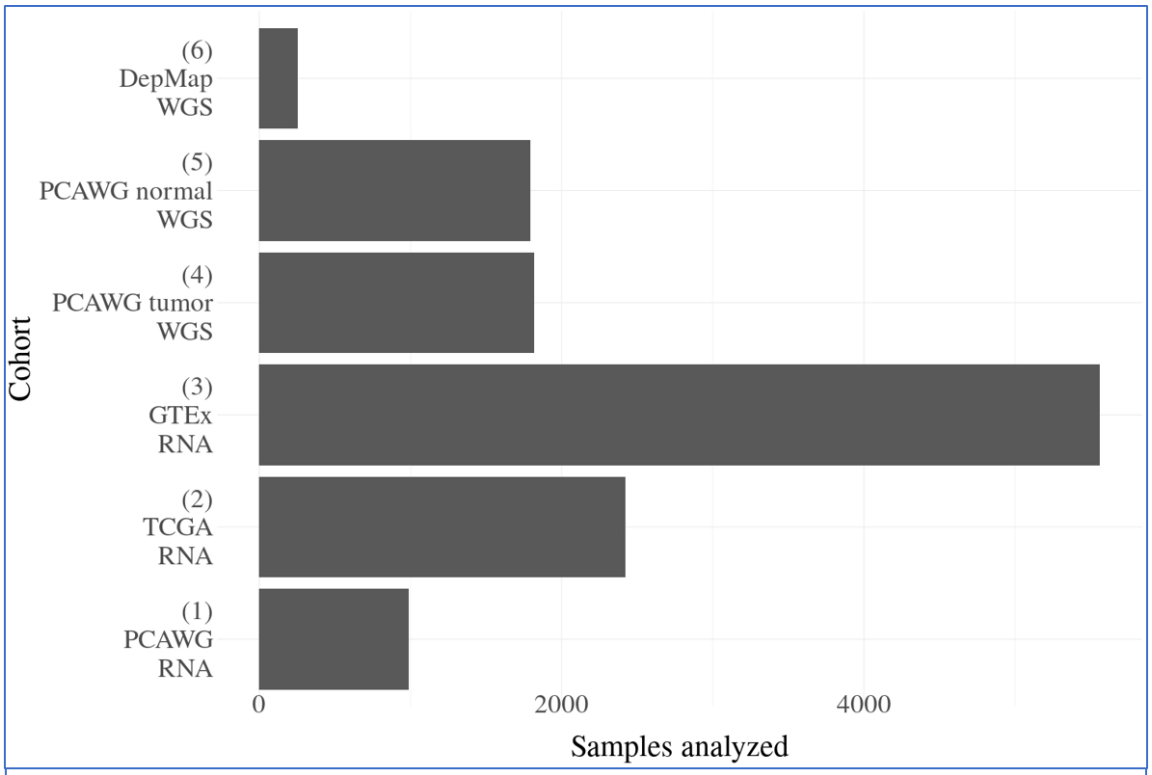


Fig. 3.1: 6 high-throughput sequencing data sets analyzed: The samples investigated with D-ViSioN are derived from different databases: DepMap, PCAWG, TCGA and GTEx. In case of PCAWG the data analyzed is divided in three different categories: PCAWG RNA, PCAWG WGS tumor, PCAWG WGS normal.

In subsequent analysis steps, the results will always be categorized using these 6 data sets. In case of the 3 RNA data cohorts, the highest number of samples was available and analyzed for the GTEx database (fig. 3.1). The cancer RNA cohorts of the TCGA and PCAWG projects have much lower total sample numbers. In case of the PCAWG WGS cohorts, the PCAWG tumor and normal cohorts comprise almost the same number of samples (PCAWG WGS tumor: 1820, PCAWG WGS normal: 1791), since there is both a tumor sample and a negative control available for most patients included in the PCAWG database (fig. 3.1). The DepMap cohort of cell line sequencing data used as negative control is the smallest of the 6 data sets including 256 samples (fig. 3.1).

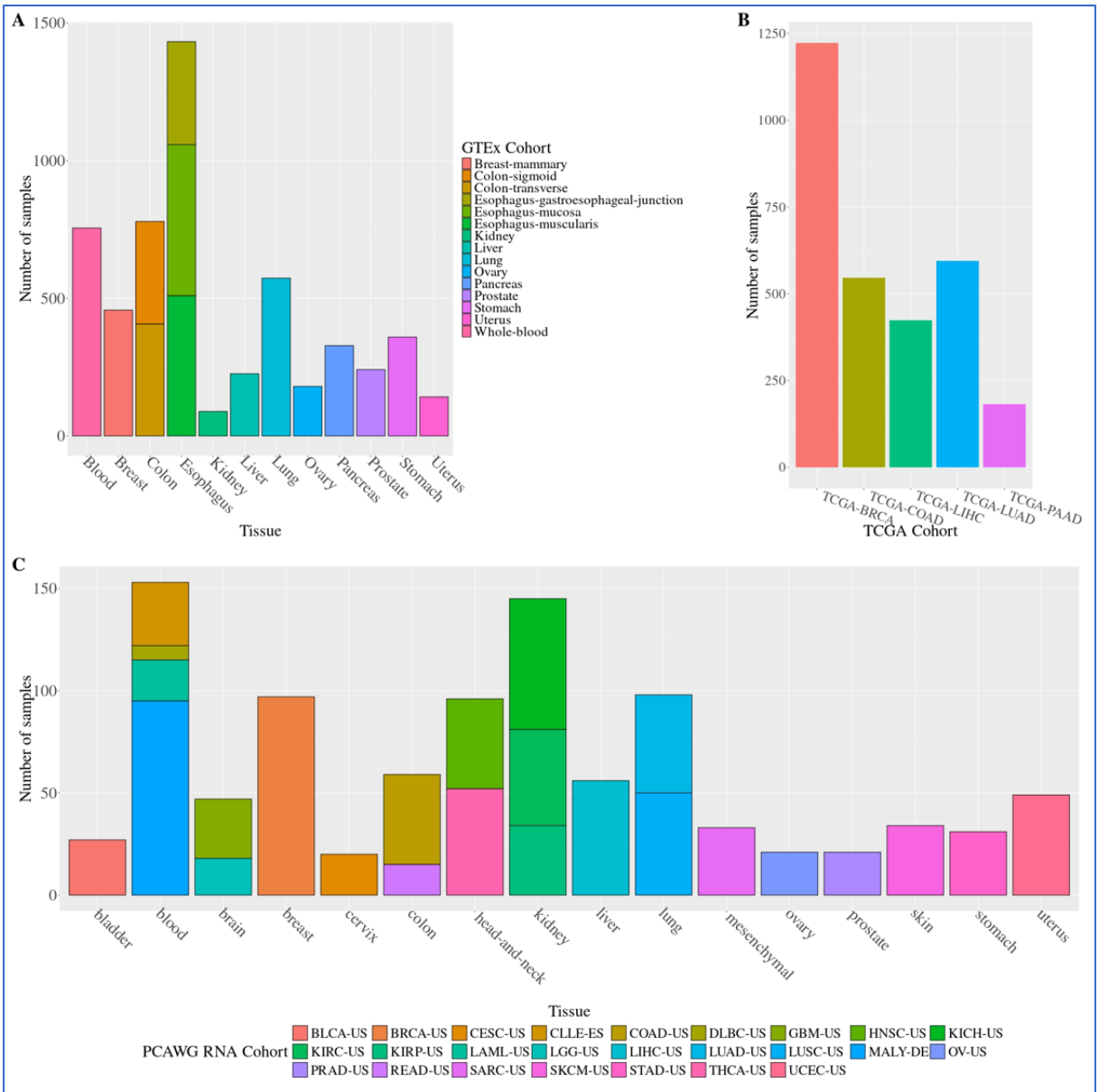


Fig. 3.2: Cohorts included in RNA sequencing data sets: A) GTEx RNA data set: The analyzed GTEx cohorts cover 12 different tissues. B) TCGA RNA data set: 5 cohorts of the TCGA project were screened using D-ViSioN: breast cancer, colorectal cancer, liver cancer, lung cancer and pancreatic cancer. C) PCAWG RNA data set: The PCAWG RNA data set contains 25 cohorts covering 16 different cancer sites.

All data sets contain cell line, tumor and tissue data of different cancer sites or tissues. I selected 12 tissues-of-interest of the GTEx data set for D-ViSioN analysis based on the inflammation context of matching cancer types as well as because of the *in silico* results of the corresponding cancer cohorts of the TCGA and PCAWG data sets (fig. 3.2A). The samples sizes available for the respective tissues vary. While there are only 89 kidney samples included in the GTEx data set, there are 755 whole blood samples available for D-ViSioN analysis (fig. 3.2A). In case of colon and esophagus, there are 2, respectively 3 cohorts included with samples of different regions of the respective organ (fig. 3.2A). I analyzed TCGA data of breast cancer (TCGA-BRCA), colorectal cancer (TCGA-COAD), liver cancer (TCGA-LIHC), lung adenocarcinoma (TCGA-LUAD) and pancreatic adenocarcinoma (TCGA-PAAD) (fig. 3.2B). The colorectal cancer RNA data was retrieved from the TCGA legacy portal, whereas the other four cohorts were downloaded from the harmonized TCGA portal. The TCGA breast cancer data is by far the largest of the TCGA cohorts with nearly 1250 samples (fig. 3.2B). The PCAWG RNA cancer cohorts contain less patients than the TCGA RNA cohorts, the smallest PCAWG RNA cohort (DLBC-US – lymphoid neoplasm diffuse large B-cell lymphoma) comprises for example just 7 patients. In contrast to the TCGA data set there are however 25 different cancer types of 16 different cancer sites covered by the PCAWG RNA data available for D-ViSioN analysis (fig. 3.2C).

The DepMap data set consists of cell line data of 22 different cancer sites (fig. 3.3A). The number of cell lines differs between the different data sets. There are for example more than 80 lung and about 30 breast cancer cell lines available, whereas there are less than 5 prostate cancer cell lines included (fig. 3.3A). The PCAWG WGS data analyzed using D-ViSioN includes both tumor and normal tissue/blood data of 29 different cancer cohorts belonging to 13 different cancer sites (fig. 3.3B). While there would have been PCAWG WGS data of more cancer types available, I focused on the cohorts shown in figure 3.3B for D-ViSioN analysis based on the D-ViSioN results of the PCWG-RNA data. Most analyzed PCAWG-WGS cancer cohorts contain the same number of tumor and normal tissue/blood samples – one each per patient. However, the prostate cancer cohort from the UK (PRAD-UK) provides not only primary tumor samples, but also samples of metastases (fig. 3.3B). The Japanese liver cancer cohort LIRI-JP also shows discrepancies between the tumor and normal tissue sample numbers. In case of this cohort, the differences regarding the sample number are caused by the uncomplete analysis due to repeated failed D-ViSioN runs. Since this problem only occurred for this single sub-cohort, the interruptions of the D-ViSioN analysis might be caused by technical issues with the raw data files of this cohort downloaded to the DKFZ internal ODCF cluster structure.

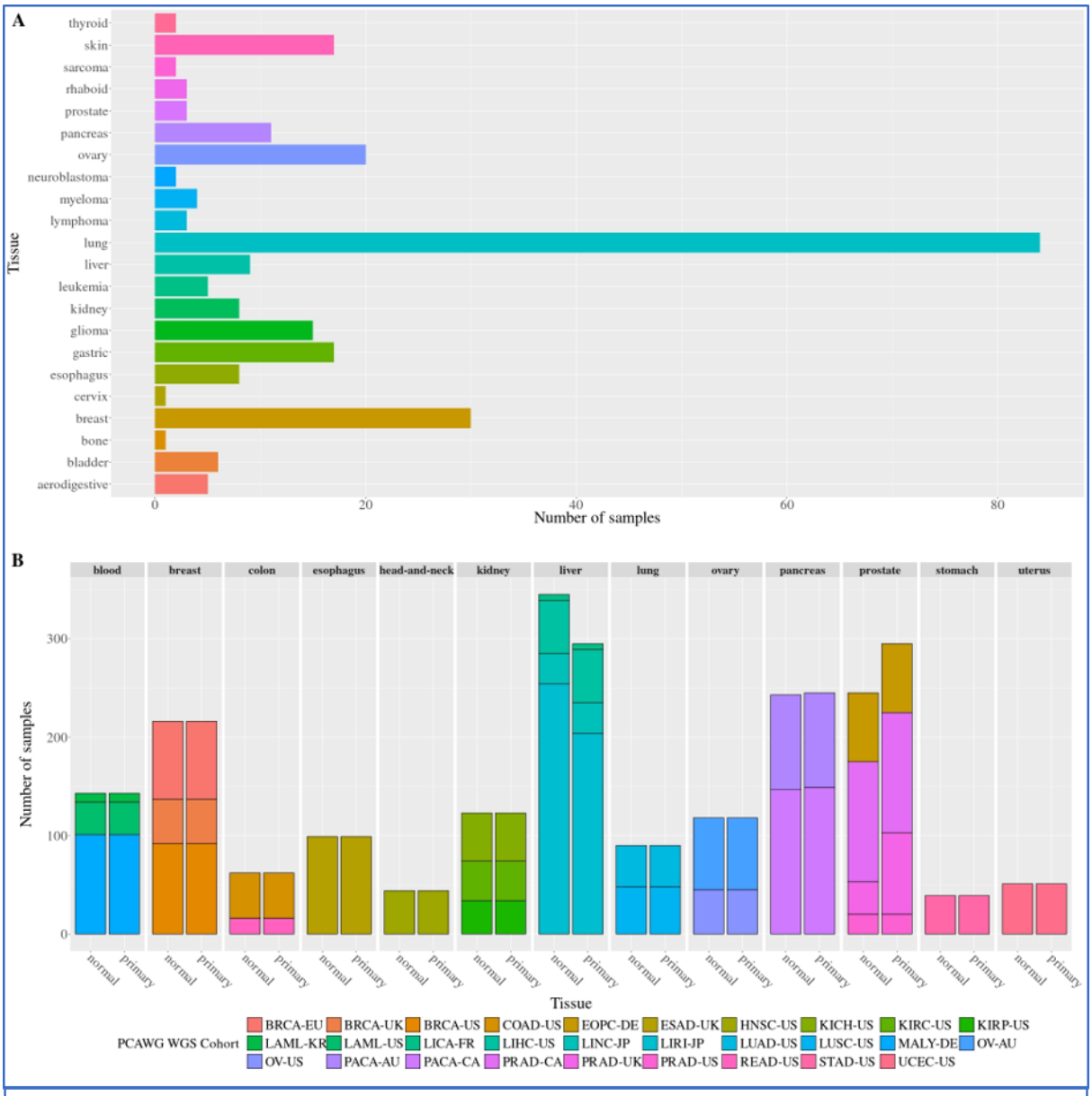


Fig. 3.3: Cohorts included in WGS data sets: A) DepMap data set: The screened cell lines cover 22 different tissues. B) PCAWG WGS data set: The PCAWG WGS data set contains for most samples one tumor sample and one normal tissue/blood sample. WGS data of 13 different cancer sites was analyzed. There are from one to four different cancer cohorts available per cancer site, which is shown with stacked bar plots.

3.1.1 Definition of BMMF positivity

In general, BMMF reads were detected in all 6 data sets analyzed using D-ViSioN, albeit at different quantities. In all data sets screened, the number of samples without BMMF reads detected is exceeding the number of samples with BMMF reads detected by far (fig. 3.4, fig. 3.5). The majority of RNA samples with BMMF reads detected exhibits very low BMMF read numbers in all three RNA data sets (fig. 3.4). Two TCGA-BRCA samples as well as 5 PCAWG-RNA samples of different cancer cohorts are the only outliers with more than 10

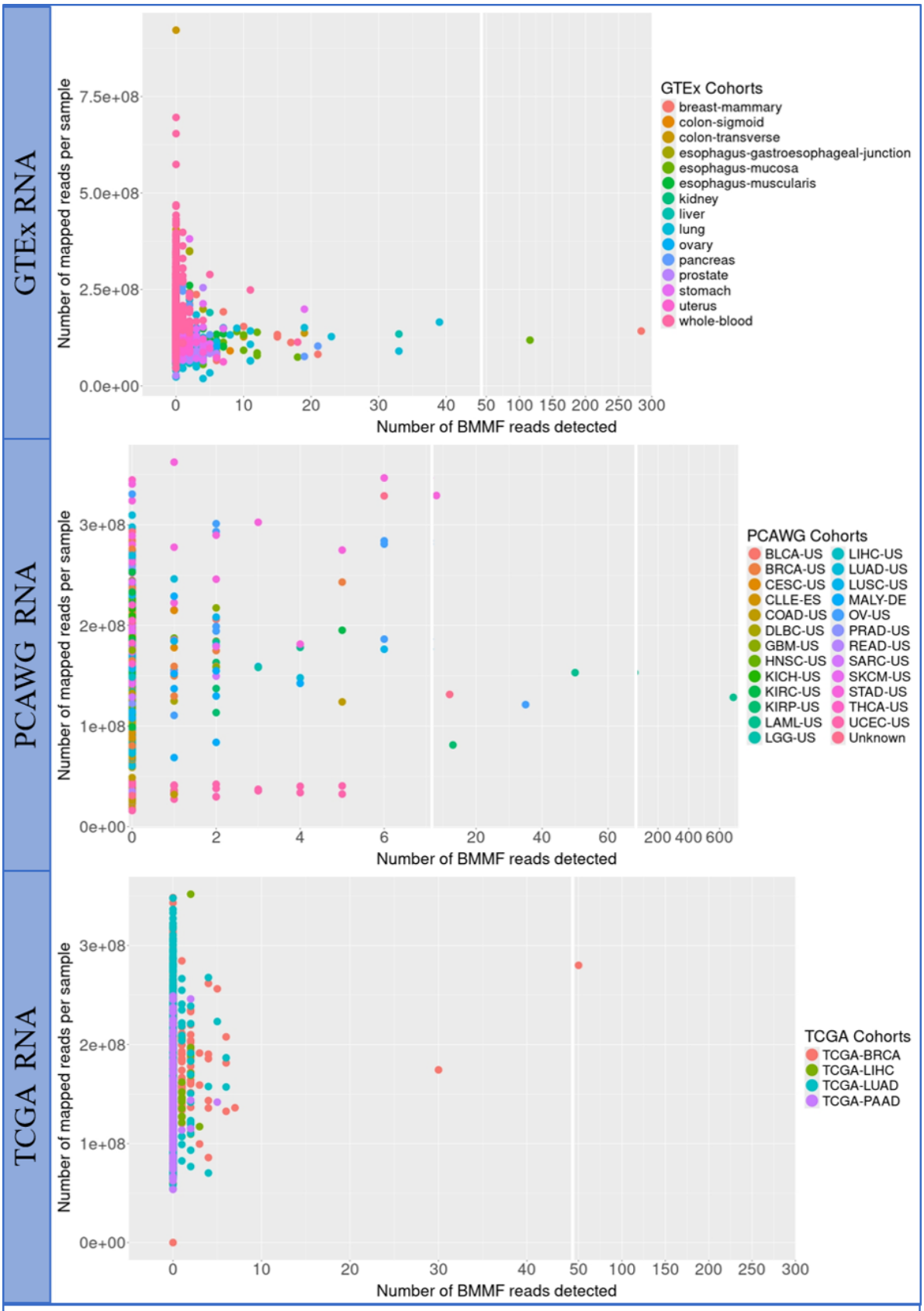


Fig. 3.4: BMMF reads detected in RNA-seq data sets: BMMF reads detected per sample plotted against sequencing depth for the RNA sequencing data sets GTEx RNA, PCAWG RNA, TCGA RNA. Each dot represents one sample. The colors of the dots indicate the tissue/cancer cohort of the sample. The reads were detected with D-ViSioN, the sequencing depth was calculated with samtools. The plot was generated using R.

BMMF reads detected per sample among the cancer RNA sequencing data samples (fig. 3.4). Looking at the GTEx RNA data set, more than 20 different samples have been found with at least 10 BMMF reads detected (fig. 3.4).

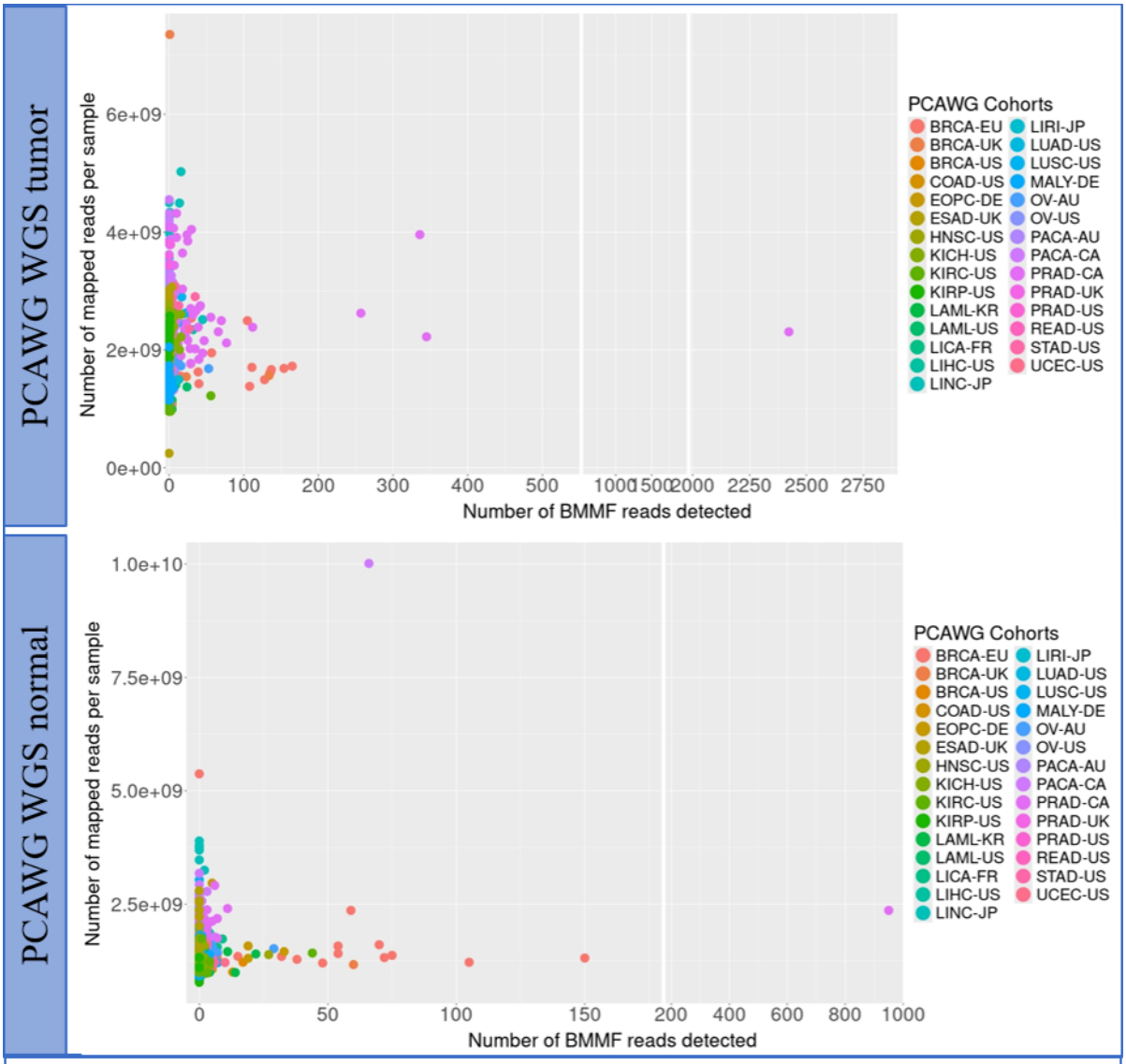


Fig. 3.5: BMMF reads detected in PCAWG WGS data: BMMF reads detected per sample plotted against sequencing depth for the PCAWG WGS tumor and PCAWG WGS normal tissue/blood data sets. Each dot represents on sample. The colors of the dot indicate the tissue/cancer cohort of the sample. The reads were detected with D-ViSioN, the sequencing depth was calculated with samtools. The plot was generated using R.

The tumor and normal tissue/blood PCAWG WGS data sets show on average higher numbers of BMMF reads detected per sample compared to the RNA data sets (fig. 3.4, fig. 3.5). The samples with higher numbers of BMMF reads detected are mostly derived from corresponding primary and normal tissue cohorts of the same cancer types such as the EU and UK breast cancer cohorts or the Canadian prostate cancer cohort. However, just as in the RNA data sets, there are still high numbers of PCAWG WGS samples with very low amounts of BMMF reads

(fig. 3.5). In case of the DepMap data base, D-ViSioN detected only one BMMF read in the ovarian carcinoma cell line COV644 (fig. 3.6).

The observation of generally low BMMF read numbers in the majority of RNA and WGS samples with BMMF reads, raised the question how many reads are needed to consider a sample as "BMMF positive". A single BMMF read in a sample does not constitute a reliable BMMF signal. Defining files with very low read number as positive poses the risk to count samples as positive with single reads caused by noise, contaminations or spill-over during sample sequencing. Even though the mostly negative DepMap cell line data seems to indicate, that contaminations with BMMF sequences do not seem to be a widespread problem, I still decided to exclude samples with reads counts lower than three from subsequent analysis steps for definition of BMMF positive samples. Additionally, the so-called three-hits threshold I introduced only defines samples as positive if there have been at least three BMMF reads of the same BMMF group reported for the respective sample. Since the BMMF isolates within the four major BMMF groups are highly conserved, there must be at least three BMMF reads per sample of closely related sequences to consider the sample as BMMF positive.

The introduction of the three-hits threshold impacts the number of samples counted as positive for data sets analyzed using D-ViSioN. In the PCAWG RNA data set, there has been at least one BMMF read detected in 8.92% of the samples analyzed. Taking into account the three-hits threshold, 2.33% of the samples can be considered as BMMF positive (fig. 3.6). In case of the TCGA RNA cohorts, there is one or more BMMF reads reported for 5.7% of the samples. However, only 0.83% of the TCGA samples meets the three-hits threshold, which means that more than 80% of the TCGA RNA samples with BMMF reads detected will not be defined as positive in subsequent analysis steps (fig 3.6). 13.04% of all samples in the GTEx RNA data set include at least one BMMF read, which is the highest percentage of the three RNA cohorts analyzed. After applying the three-hits threshold only 2.49% of the 5429 GTEx samples analyzed remain to be consider as positive, which is approximately at the level of the PCAWG RNA data set (fig. 3.6). The TCGA cohorts exhibit the lowest percentage of BMMF positive samples of all three RNA data sets included in this analysis.

Looking at the WGS data sets, the three-hits threshold reduces the percentage of samples defined as positive from 19.89% to 8.79% in case of the PCAWG WGS tumor data sets and from 12.79 % to 4.58 % in case of the PCAWG WGS normal tissue/blood data (fig. 3.6). In case of the DepMap data, the three-hits threshold eliminated the single BMMF read identified in this data set. Consequently, all samples of the DepMap cell line data are considered as BMMF

negative after applying the three-hits threshold (fig. 3.6). The DepMap data set was analyzed with the intention to use it as negative control, since BMMFs are not expected in cancer cell lines because of their indirect role in carcinogenesis. The elimination of BMMF hits in this data set by the three-hits threshold supports this assumption. Consequently, the three-hits threshold is applied to remove samples in which the low BMMF detection could also be attributed to contaminations instead of an actual BMMF signal. On the other hand, this filtering step radically reduces the number of files considered as positive for all data sets, which might lead to the exclusion of reads mapping to the BMMF library with a high sequence identity from further analysis. While the three-hits threshold might cause some false negative samples, it also ensures, that only samples with a robust BMMF signal are counted as positive and included in subsequent analysis steps and conclusions drawn based on them.

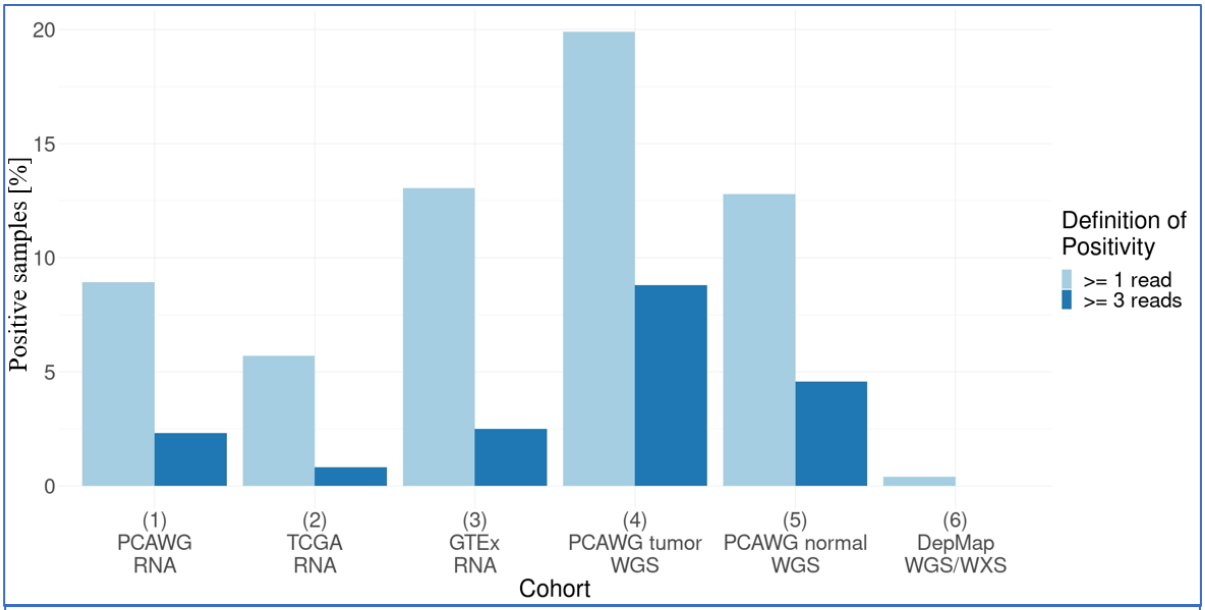


Fig. 3.6: BMMF positive samples in RNA sequencing and WGS data sets: BMMF positive samples [%] for all five data sets with two different definitions of positivity: Light blue: Each sample with at least one BMMF read is counted as positive. Dark blue: samples with at least three reads detected for the same BMMF groups are considered positive. The BMMF reads were detected with D-ViSioN, the plot was generated using R.

3.1.3 Analysis and impact of sequencing depth

The sequencing depth of a sample might have an impact on detection levels. Since there are only low detection levels expected for BMMFs due to the perceived indirect link to cancer, the sequencing depth might have an impact on the likelihood of a sample being classified as positive or negative. To assess the effect of the sequencing depth on BMMF detection, I generated histograms contrasting the sequencing depth of positive and negative samples for all

five data sets (fig. 3.7). The continuous lines intercepting the histograms highlight the mean numbers of total mapped reads per sample.

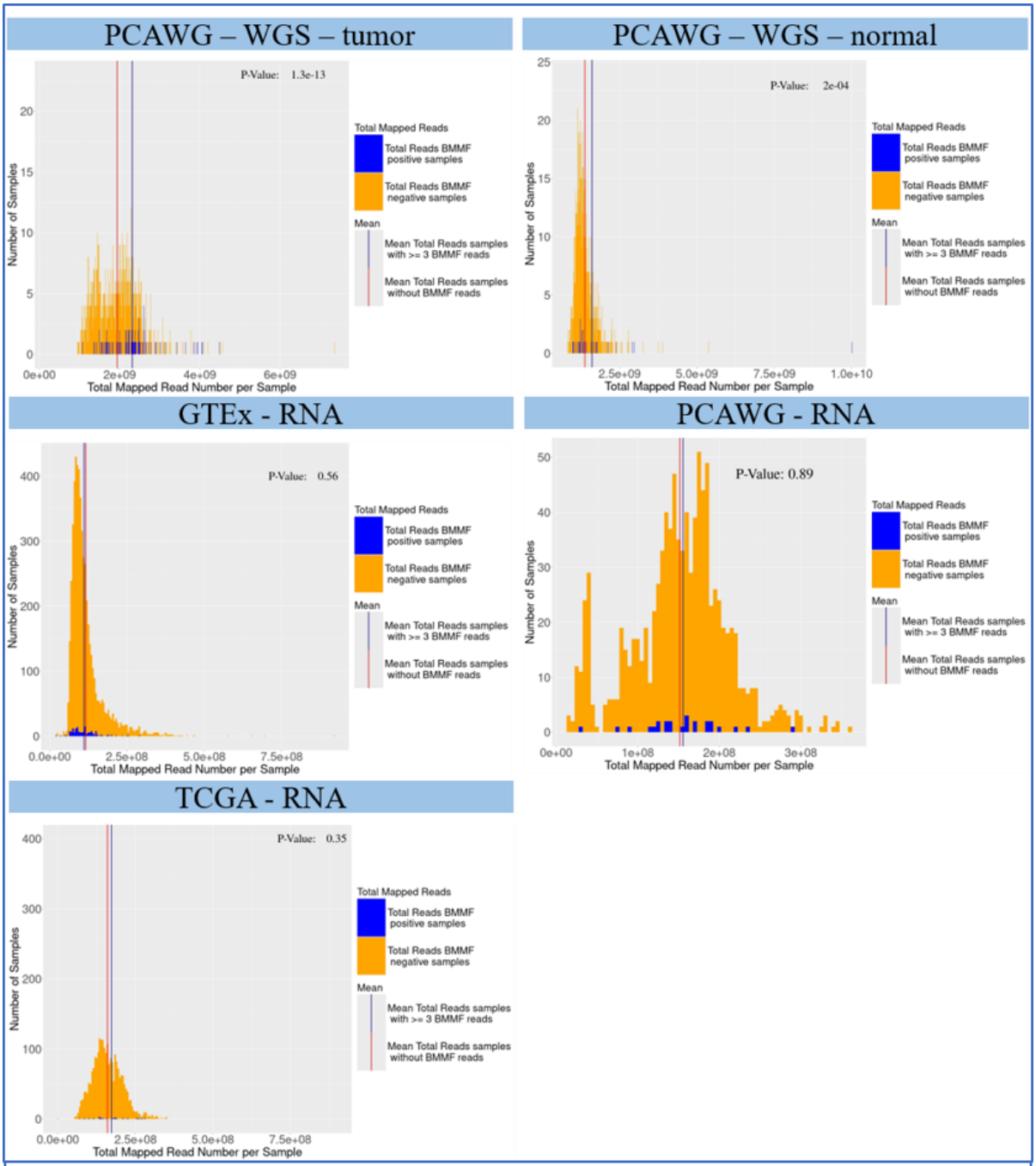


Fig. 3.7: Histograms of sequencing depth of BMMF positive and negative samples for RNA-seq and WGS data data sets: The histograms for the sequencing depth were generated for the PCAWG, TCGA and GTEx RNA seq data and for the PCAWG tumor and normal tissue/blood WGS data to verify, if there are differences between the sequencing depth in BMMF positive and negative samples. The blue and red lines indicate the respective mean sequencing depth, the p-value was calculated using the Mann-Whitney-U test in R. The sequencing depths were determined using samtools, the histograms were generated using R.

The mean sequencing depth of BMMF positive PCAWG WGS samples is higher than the mean

sequencing depth of BMMF negative PCAWG WGS samples for both tumor and normal tissue data (fig 3.7). Since Q-Q-plots indicated, that the sequencing depth is not normally distributed for all data sets (fig. S7), I used the Mann-Whitney-U test in R to assess if the optically observed differences between the mean sequencing depths for BMMF positive and negative samples are statistically significant. The p-values show a statistically significant difference between the sequencing depth of BMMF positive and negative samples for both tumor and normal tissue PCAWG WGS data (fig. 3.7). The RNA data sets however show no significant differences between the mean sequencing depth of positive and negative samples. Consequently, I conclude that the sequencing depth has an impact on BMMF detection for the WGS data analyzed, but not in case of the RNA sequencing data. To account for this, I will use the numbers of total mapped reads per sample to normalize the detected BMMF hits for sequencing depth as described in chapter 2.3.4.

3.2. Detection of BMMF reads assigned to four BMMF groups

The BMMF library contains BMMF sequences assigned to the four different BMMF groups. BMMFs within the same groups are often highly conserved, however BMMFs of different groups are only distantly related. For this reason, I analyzed the detection of the four main BMMF groups in the five sequencing data sets to assess if there are differences between the detection levels of these BMMF groups. Furthermore, I examined if the four main BMMF groups are associated with certain cancer or tissue sites at either RNA or DNA level. Samples are only counted as positive for one of these four BMMF groups, if at least three BMMF reads of the respective group are detected.

Tab. 3.1: BMMF reads detected in RNA sequencing and WGS data sets: Number of BMMF reads of all 6 data bases listed for the four main BMMF groups separately and in total. BMMF reads detected using D-ViSioN were only included in this table, if at least 3 reads of the respective BMMF group were detected in one sample. The data analyzed includes the RNA sequencing data from TCGA, PCAWG and GTEx as well as the WGS data from PCAWG and DepMap.

BMMF Group	TCGA RNA	PCAWG RNA	GTEx RNA	PCAWG WGS - Tumor	PCAWG WGS - Normal	DepMap	Total
BMMF1	165	729	1103	5763	2090	0	**9851**
BMMF2	18	149	267	913	227	0	**1569**
BMMF3	0	0	11	3	0	0	**14**
BMMF4	0	0	21	223	18	0	**262**
Total	183	878	1402	6902	2335	0	**11696**

Table 3.1 shows the read numbers detected for the four BMMF groups for each data set after the application of the three-hits threshold. As already depicted in figure 3.6. there are no cell line samples within the DepMap data analyzed, that meet the requirements of the three-hits threshold. There were however sufficient BMMF hits detected to exceed this threshold in all other data sets analyzed. The tumor PCAWG WGS data contains the highest amount of BMMF hits with in total 6902 BMMF reads detected, which are about three times as many reads as the 2335 BMMF reads found in the normal tissue/blood PCAWG WGS data (tab. 3.1). The three RNA data sets exhibit lower hit numbers than the WGS data set. The highest number of BMMF reads were detected in the GTEx data with 1402 total reads, followed by the PCAWG RNA data with 878 reads and the TCGA data with 183 BMMF reads (tab. 3.1). BMMF group 1 is covered by the highest number of reads in all five data sets, while BMMF group 2 is also detected in all data sets albeit at lower levels (tab. 3.1). BMMF1 reads account for 84.2% of the total BMMF reads detected across all cohorts, whereas 13.4% of the reads detected in all cohorts are BMMF group 2 reads (supplementary table S8). Looking at the five data sets individually, the percentage of BMMF1 hits ranges from 78.7% in GTEx to 83% and 83.5% in PCAWG RNA and PCAWG WGS tumor data up to 89.5% and 90.2% in the PCAWG WGS normal and TCGA RNA data sets (S8). The GTEx cohort shows the highest ratio of BMMF2 hits, which comprise 19% of the total hits detected, whereas the TCGA RNA and PCAWG WGS normal cohorts exhibit the lowest percentages of BMMF2 hits, since only 9.8% and 9.7% of the detected reads are assigned to BMMF group 2 (S8).

BMMF group 4 hits are detected in three of the five data sets. The majority of BMMF group 4 hits detected across all data sets were found in the PCAWG WGS tumor data, where about 3 % of the detected hits are assigned to BMMF group 4 (tab. 3.1, S8). Additionally, there were BMMF4 hits detected in the GTEx RNA data and in the normal tissue/blood PCAWG WGS data, which account for 1.5 % and 0.77 % of the total reads detected in these two cohorts (tab. 3.1, S8). Table 3.1 shows only sporadic detection of BMMF group 3. However, this is also caused by the removal of hits covering the BMMF group 3 template HCBI8.215, which always mapped to the same short segment of about 20 bp of the HCBI8.215 genome (fig. 3.8). This suggests that the HCBI8.215 reads detected by D-ViSioN are most likely artifacts and not caused by the actual presence of the BMMF3 isolate. Consequently, the BMMF reads assigned to this genome were excluded from further analysis. After the removal of HCBI8.215 reads, there are only very few samples with BMMF3 hits left that meet the three-hits threshold. The remaining BMMF 3 hits were found in the GTEx RNA and the PCAWG WGS tumor data (tab. 3.1, S8).

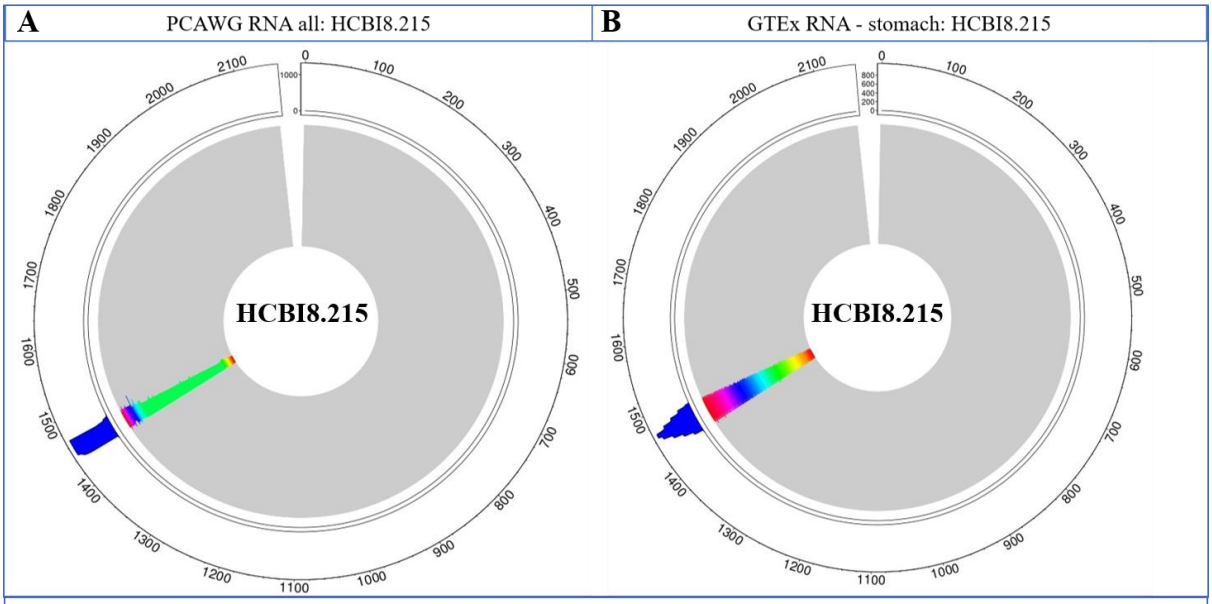


Fig. 3.8: Detected reads for BMMF group 3 isolate HCBI8.215: The BMMF3 isolate HCBI8.215 was detected in several different samples in the PCAWG RNA data (A) and GTEx RNA stomach (B) data sets. The outer circle shows in blue the number of detected reads per base pair of the HCBI8.215 genome. The inner circle maps the individual reads detected for the HCBI8.215 genome. Reads detected in the same patient are depicted in the same color.

Despite an apparent increase of BMMF detection in WGS data compared to the RNA data, this data sets are not directly comparable due to the different experimental protocols applied to the samples. Additionally, the five data sets analyzed also differ regarding their samples size and sequencing depth, which was shown to significantly affect BMMF detection at least in case of WGS data (chapter 3.1.3, fig. 3.7). Consequently, the detected BMMF reads were normalized for sequencing depth and cohort size in the next step. For this, I used the sequencing depth of the samples as well as the sample size of the respective data sets to determine the number of BMMF reads detected per billion mapped reads per 100 samples for all four BMMF groups as well as for the total number of BMMF reads detected per data set (tab. 3.2). While the raw numbers of BMMF hits indicated higher BMMF read numbers for the GTEx normal tissue data than for the two RNA data sets derived from cancer patients, the PCAWG RNA data set stands out with the highest number of BMMF reads detected per billion mapped reads per 100 samples across all three RNA sequencing data sets (tab. 3.2). Both the normalized BMMF group 1 and BMMF group 2 numbers of the PCAWG RNA data are more than 10 folds higher than in the other two RNA data sets. The TCGA RNA data sets shows slightly higher numbers of normalized BMMF1 and total BMMF reads detected compared to the GTEx data. However, the GTEx data sets exhibits marginally higher normalized read numbers for the other three BMMF groups.

Tab. 3.2 BMMF reads in RNA sequencing and WGS data sets normalized for sequencing depth and cohort size: BMMF reads per billion mapped reads per 100 samples listed for the four main BMMF groups separately. The means were calculated including only BMMF positive samples (at least 3 reads detected per BMMF group). The data analyzed includes the RNA sequencing data from TCGA, PCAWG and GTEx as well as the PCAWG WGS data.

BMMF Group	TCGA RNA	PCAWG RNA	GTEx RNA	PCAWG WGS - Tumor	PCAWG WGS - Normal
BMMF1	1.6934	24.6855	1.1587	0.9363	0.8326
BMMF2	0.1169	4.8156	0.2618	0.1353	0.0899
BMMF3	0.0000	0.0000	0.0118	0.0008	0.0000
BMMF4	0.0000	0.0000	0.0224	0.0277	0.0060
Total	1.8103	29.5010	1.4547	1.1000	0.9284

Looking at the two PCAWG WGS data sets, the normalization also relativizes the initial observations. While the PCAWG WGS tumor data showed both a higher percentage of BMMF positive samples and higher absolute read numbers compared to the PCAWG WGS normal data, the difference between these data sets becomes much smaller after normalization. The tumor data shows still higher normalized BMMF read levels in total as well as for all four BMMF groups, however the number are only slightly increased compared to the normal tissue/blood data. While the sample size only marginally differs between both data sets, the average sequencing depth is higher in case of the tumor cohort.

3.2.1 Identification of cancer and tissue sites with BMMF positive samples

To identify cancer and tissues sites with potentially increased BMMF detection, I focus on twelve different tissues-of-interest. This includes both tissues affecting inflammation-linked cancer types considered as relevant for BMMF research and tissues which stood out with increased BMMF detection during the D-ViSioN analysis. Since the five data sets analyzed are comprised of different compositions of cancer and tissue cohorts, there are not all tissue types represented in all data sets. The PCAWG RNA and the TCGA RNA data cover only ten respectively five of the ten tissues of interest (fig. 3.9). On the other hand, I also screened PCAWG RNA and WGS data of other cancer sites besides of these twelve selected tissues-of-interest, which is not included in this comparison. In case of some tissues or cancer sites, there is more than one cohort available covering the same tissue, but from either different cancer types or different geographic origins. In this step of analysis, the D-ViSioN results of cohorts covering the same tissue are combined.

Comparing the twelve tissues across the different data sets analyzed, the TCGA cohort shows less than 1% positive samples in all five cohorts analyzed (fig. 3.9). In case of the PCAWG

RNA data set, the positivity of the most frequently detected BMMF group 1 ranges from 1% to 3% for most cancer sites analyzed. BMMF group 2 reads are detected in 1.3% of the samples derived from blood cancer patients, which makes blood the only cancer site with BMMF2 positivity after applying the three-hits threshold. The liver and prostate cancer RNA data cohorts of the PCAWG project are even completely negative. However, there are three cancer sites of the PCAWG RNA data set that attract further attention: the cancer samples derived from the uterus, stomach and ovaries are exhibiting positivity levels for BMMF1 genomes of between 10% and 20% (fig. 3.9). The GTEx RNA data set shows higher levels of BMMF positivity compared to the TCGA RNA data, however no tissue cohort shows positivity rate higher than 5% for any of the four BMMF groups. In contrast to the two cancer sequencing data cohorts, the GTEx tissue cohorts exhibit BMMF2 positivity in several different tissues albeit at low positivity rates. The uterus and breast cohorts show the highest BMMF1 positivity within the GTEx data set with about 4.2% BMMF 1 positive samples each (fig. 3.9).

The PCAWG WGS tumor and normal tissue/blood data sets exhibit on average higher BMMF positivity rates compared to the RNA data sets. For most cancer sites, the BMMF positivity is higher in the tumor cohorts than in the normal tissue/blood cohorts (fig. 3.9). This includes the BMMF1 positivity of the breast, esophagus, kidney, lung, pancreas, prostate and stomach cancer cohorts. Additionally, the livery and ovary cohorts show slightly elevated BMMF1 positivity rates in the tumor data compared to the normal tissue/blood data. In case of blood cancer, the BMMF1 positivity is nearly identical between the tumor and the control data, whereas the colon and uterus cohorts exhibit higher percentages of BMMF1 positivity in the normal tissue/blood data than in the tumor data. The BMMF2 positivity is higher in all tumor data cohorts with BMMF2 reads reported compared to the respective normal tissue/blood samples with the exception of blood cancer. In general, the BMMF1 positivity exceeds the BMMF2 positivity for all cancer sites in both tumor and normal tissue data with exception of the control samples of blood cancer patients (fig.3.9). The cancer sites with BMMF1 positivity rates exceeding 5% in the PCAWG WGS tumor data set include breast, kidney, lung, pancreas, prostate and stomach cancer. 15.56% of the lung cancer patients, 20% of the prostate cancer patients and 23.07% of the stomach cancer patients included in the PCAWG WGS data set are positive for BMMF group 1. These three cancer sites also show the highest BMMF2 positivity rates of the PCAWG WGS tumor data set. The normal tissue/blood samples of breast, lung, prostate and stomach cancer patients also exhibit BMMF1 positivity in at least 5% of all samples. 15.38% of the normal tissue/blood-derived samples of stomach cancer patients are positive for BMMF1. Interestingly, the cancer sites with the highest levels of BMMF1 positivity

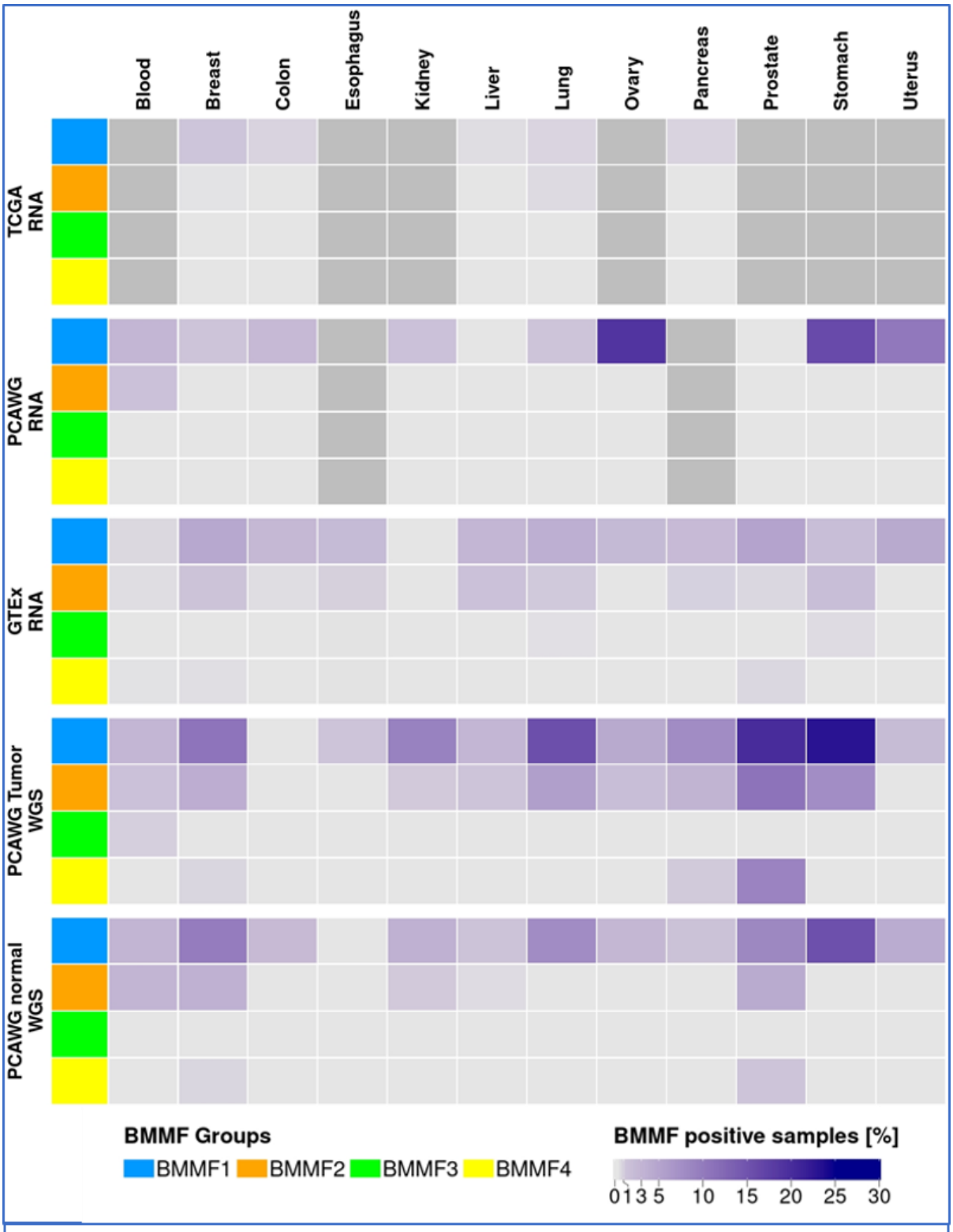


Fig. 3.9: Positivity for 4 BMMF groups in 12 tissues-of-interest: Overview of positivity of BMMF groups 1-4 under three-hits threshold in RNA and WGS data of 12 tissues-of-interest. The cancer sequencing data analysed is from the TCGA (RNA) and the PCAWG (RNA+WGS) project, the normal tissue data from the GTEx (RNA) and the PCAWG (WGS) project. Dark grey fields indicate, that no data is available for this tissue type in this sequencing project.

overlap between the PCAWG WGS tumor and normal tissue/blood data. The stomach cancer

data stands out with the highest BMMF1 positivity for both tumor and control samples, whereas the highest percentage of BMMF2 positive samples was detected in the tumor and normal tissue/blood samples of prostate cancer patients.

3.2.2 Identification of cancer and tissue sites with increased BMMF signal

Besides of the positivity rates in percent, the strength of the signal in positive samples is another important benchmark to evaluate the positivity of the different data sets analyzed for the presence of BMMFs. For this, the number of BMMF hits detected in the respective cohorts is again normalized for the sequencing depth of the positive samples as well as for the cohort size by calculating the number of BMMF reads detected per million mapped reads in a cohort with 100 samples (fig. 3.10). Since the sequencing depth in RNA sequencing samples is lower than in WGS samples, this normalization step reinforces the reads detected in RNA sequencing data sets. Despite the low number and percentage of positive samples in the TCGA RNA data, the BMMF signal is enhanced in the five TCGA cohorts upon normalization, especially in case of the colorectal and pancreatic cancer TCGA cohorts (fig. 3.10). In addition, the other two RNA sequencing data cohorts exhibit more prominent BMMF signals, especially the PCAWG RNA data set, which contains the cohorts with the highest numbers of BMMF hits per billion mapped reads per 100 samples of all data sets analyzed. Here, the PCAWG RNA blood cancer samples stand out with the by far strongest BMMF signal both for BMMF group 1 and BMMF group 2, followed by the BMMF1 detection in the ovarian and uterine cancer cohorts of the PCAWG RNA data set (fig. 3.10).

The BMMF signal in the GTEx data is also elevated by the normalization of the BMMF reads detected. However, when comparing the data of different tissues within the GTEx data set, the picture resembles the pattern previously shown by the visualization of the percentage of BMMF positive samples (fig. 3.9, fig. 3.10). While BMMF reads are found in all GTEx tissue cohorts with exception of the kidney data, the BMMF1 positivity seems to be at a similar level across most of the different tissues analyzed without any cohorts clearly standing out from the others. While the BMMF1 positivity is the highest in uterus-derived samples, the BMMF1 signal is nearly as strong in the breast, liver, lung, ovary, pancreas and prostate samples. The stomach, colon, esophagus and blood samples show less prominent BMMF1 signals (fig. 3.10). The BMMF2 signal shows more variation within the GTEx data set, since the breast, liver, pancreas and stomach cohorts show higher normalized detection compared to the other GTEx tissue cohorts with little to no BMMF2 positivity. In contrast to the two data sets with sequencing data of cancer patients, the GTEx data also contains three, respectively four tissue cohorts

positive for BMMF group 3 and 4, however the signal is weak in all of these cohorts (fig. 3.10).

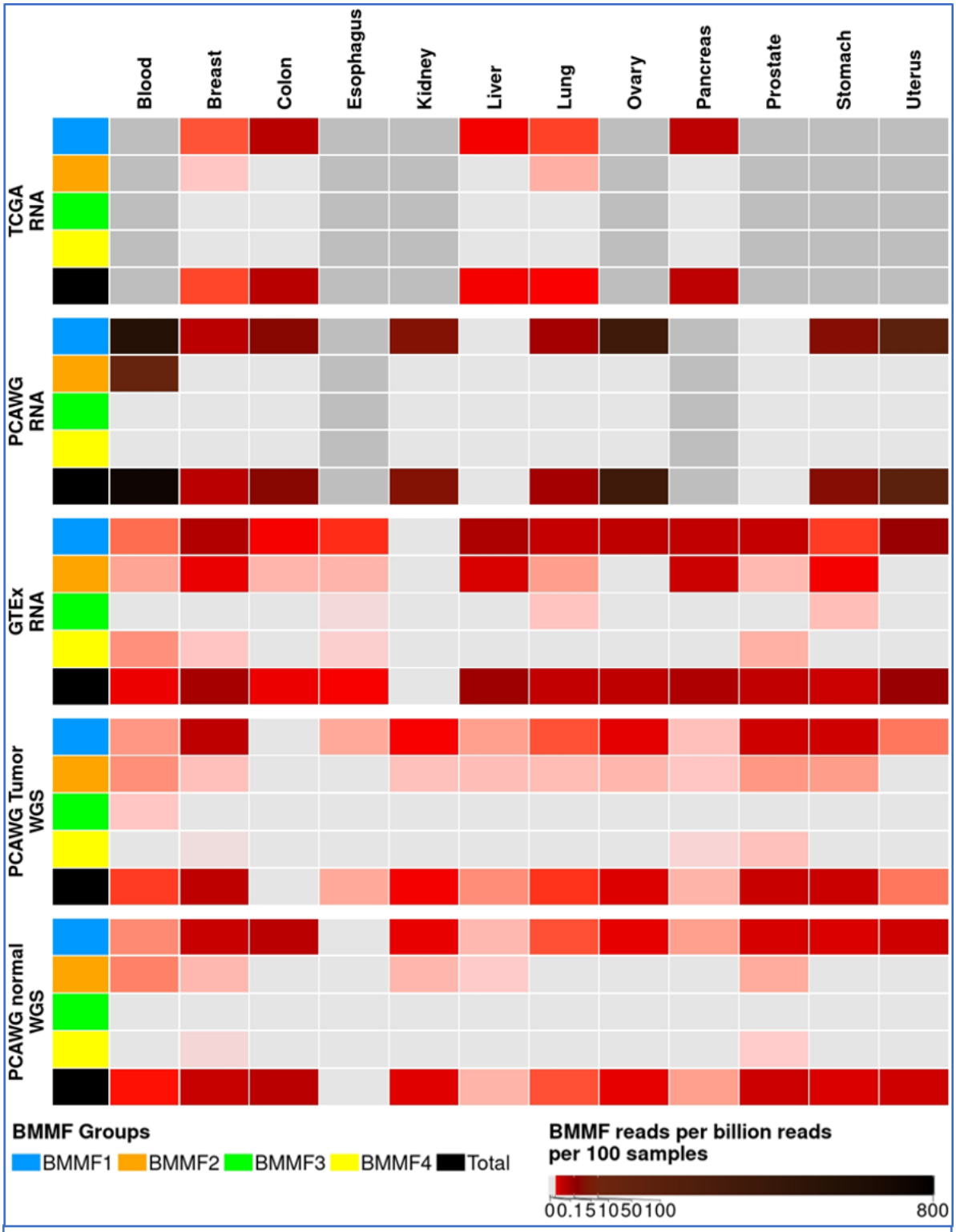


Fig. 3.10: Mean BMMF reads per billion reads per 100 samples of positive samples for 4 BMMF groups in 12 tissues-of-interest: Overview of BMMF reads per billion mapped reads per 100 samples of BMMF groups 1-4 of samples considered positive using the three-hits threshold in RNA and WGS data of 12 tissues-of-interest. The cancer sequencing data analyzed is from the TCGA (RNA) and the PCAWG (RNA+WGS) project, the normal tissue data from the GTEx (RNA) and the PCAWG (WGS) project. Dark grey fields indicate, that no data is available for this tissue type in this sequencing project.

When comparing the normalized BMMF signal of the PCAWG WGS tumor and normal tissue/blood data, the patterns of BMMF positivity between the different cancer sites show remarkable similarities especially with regards to the detection of BMMF group 1. Samples derived from breast, ovary, prostate and stomach cancer patients show a stronger BMMF1 DNA signal compared to the other cancer sites no matter if the data is derived from the tumor itself or from a control normal tissue or blood sample (fig. 3.10). Importantly, the mostly blood-derived control samples of the colon and uterus cancer patients in the PCAWG WGS normal data set show an equally strong BMMF1 signal which is comparable to the levels of detection in tumor samples of breast, ovary, prostate and stomach cancer patients. However, for uterus cancer the signal is less strong in the uterus tumor data than in the correlated normal tissue/blood data. Interestingly, for colorectal cancer, the tumor data is negative for all four BMMF groups, while BMMFs were detected in the non-tumor samples of colorectal cancer patients. The kidney and lung samples show a moderate BMMF1 signal in both PCAWG WGS tumor and normal data, whereas the blood, liver and pancreas cohorts show lower levels of BMMF1 detection respectively (fig 3.10).

Nine of twelve cancer sites show BMMF2 positivity in the respective tumor samples, whereas only five cohorts of the PCAWG WGS normal data show BMMF2 signal. Besides of this, there is a weak BMMF4 signal detected in the tumor and normal tissue/blood data of breast and prostate cancer patients as well as in the tumor samples of pancreatic cancer patients. BMMF group 4 was also detected in the breast and prostate normal tissue data of the GTEx project (fig. 3.10). In total, the BMMF1 signal overshadows the weaker signals of the three other BMMF groups in all five data sets with exception of the PCAWG RNA blood cancer data.

3.2.3 Comparison of BMMF detection in tumor and non-tumor samples on RNA and DNA level

To obtain a better overview of the BMMF detection in tumor and normal tissue data, I compared the differences between the BMMF detection in WGS tumor and normal tissue cohorts with the differences between the BMMF detection in RNA tumor and normal tissue data (fig. 3.11). The visualization of the differences regarding the BMMF reads detected per billion mapped reads per 100 samples in tumor and normal tissue data, highlighted the differences between the D-ViSioN results found in RNA and WGS data. The stomach, ovarian and lung cancer cohorts of the PCAWG tumor data show higher levels of BMMF detection in both WGS and RNA tumor data compared to the normal tissue PCAWG WGS and GTEx RNA. However, in case of the

PCAWG WGS lung and ovarian cancer data, the BMMF detection was only marginally increased compared to the normal tissue/blood data (fig. 3.11). On the other hand, there are several cohorts, where the BMMF detection is higher in tumor than in normal tissue data either only in the RNA sequencing data or in the WGS data. For example, the blood cancer RNA data shows the highest increase in BMMF detection compared to the GTEx normal blood data of all cohort differences shown in figure 3.11, whereas the BMMF detection in WGS data is slightly lower in tumor than in the non-tumor samples of blood cancer patients. Breast cancer on the other hand is the cancer type, for which the WGS data shows the largest positive difference between BMMF detection in tumor and normal tissue data. The breast RNA data of both the TCGA (triangle) and the PCAWG (dot) project however shows lower levels of BMMF detection compared to the normal tissue data of the GTEx project (fig. 3.11). However, the colorectal cancer data shows higher BMMF detection in the non-tumor WGS data. The colon RNA data on the other hand shows higher detection in the tumor data.

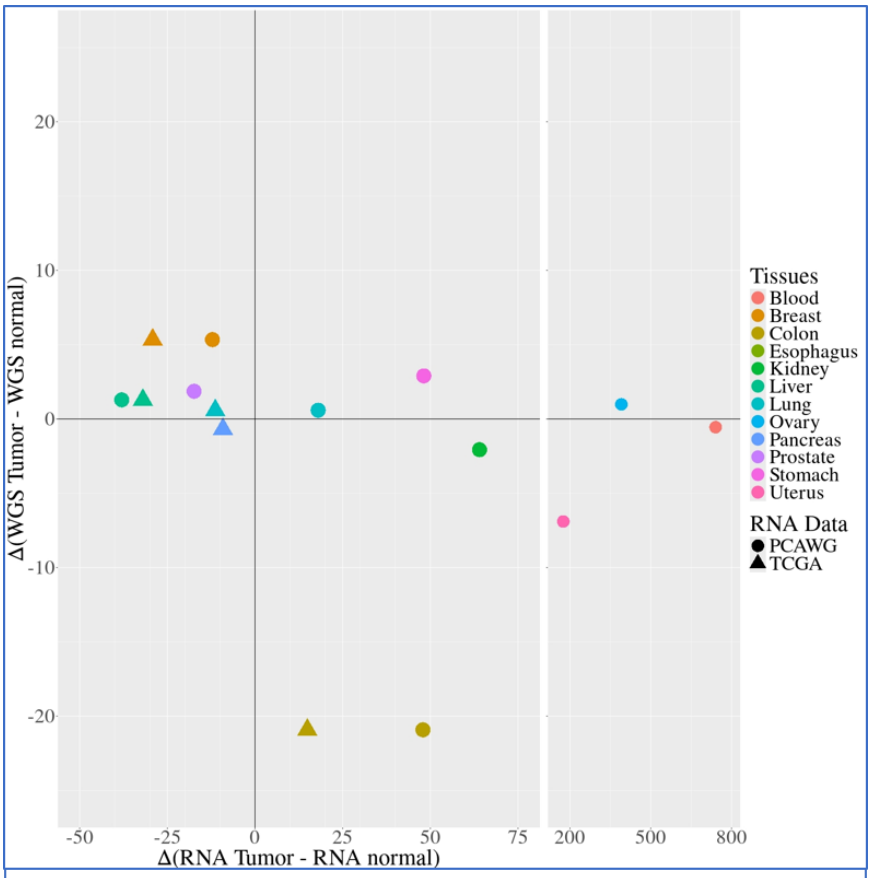


Fig. 3.11: Difference between mean BMMF reads per billion reads per 100 samples in tumor and normal tissue samples: Comparison of the difference between BMMF reads per billion mapped reads per 100 samples in tumor and normal tissue data. The x-axis shows the difference between the PCAWG RNA and the GTEx RNA tissue cohorts (dots) as well as the difference between the TCGA RNA cohorts and the GTEx RNA tissue cohorts (triangles). The y-axis shows the difference between the PCAWG WGS tumor data and the PCAWG WGS normal tissue/blood samples of the cancer patients.

Splitting up the total BMMF detection into BMMF1 and BMMF2 detection, the differences in BMMF1 detection exhibit a similar pattern (fig. 3.12 A). In fact, the removal of the non-BMMF 1 groups, leads to even lower differences between the BMMF detection in tumor and non-tumor WGS data of prostate, lung and ovarian cancer patients. However, the BMMF2 detection provides a strikingly different view. Nearly all cohorts are plotted at or around the origin of the coordinate system, indicating little to no difference between BMMF detection in

tumor and normal tissue data in either WGS or RNA sequencing data. The only exception to this is caused by the blood cancer RNA samples, which stands out with much higher BMMF2 positivity than the healthy blood RNA samples provided by the GTEx project (fig. 3.12 B).

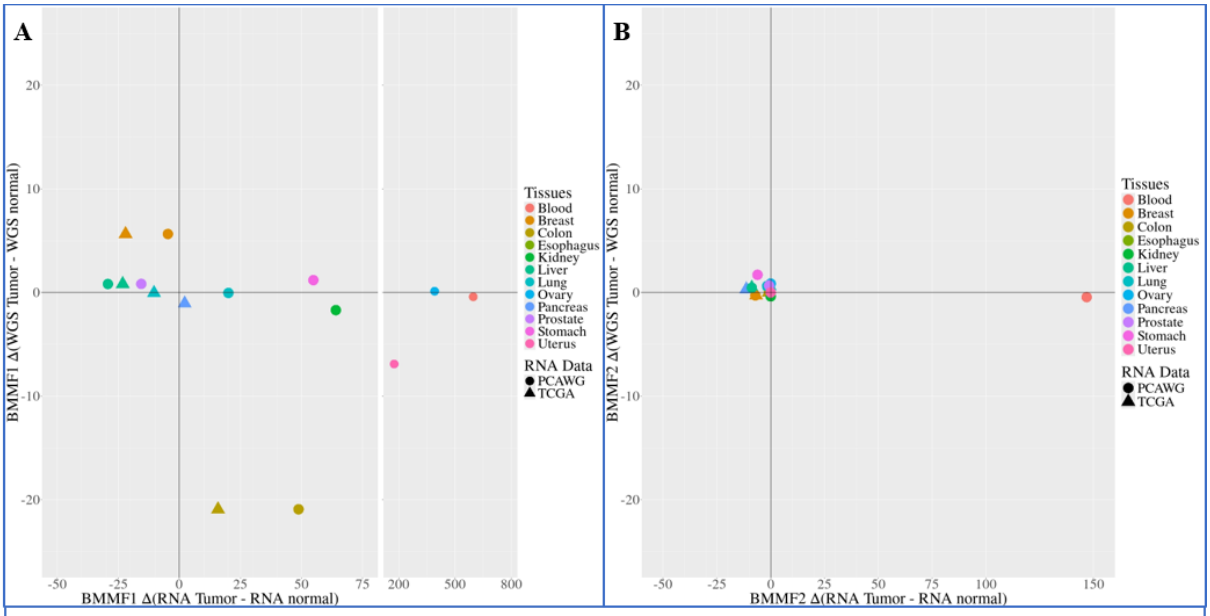


Fig. 3.12: Difference between mean BMMF1 and BMMF2 reads per billion reads per 100 samples in tumor and normal tissue samples: Comparison of the difference between BMMF1 (A) and BMMF2 (B) reads per billion mapped reads per 100 samples in tumor and normal tissue data. The x-axis shows the difference between the PCAWG RNA and the GTEx RNA tissue cohorts (dots) as well as the difference between the TCGA RNA cohorts and the GTEx RNA tissue cohorts (triangles). The y-axis shows the difference between the PCAWG WGS tumor data and the PCAWG WGS normal tissue/blood samples of the cancer patients.

3.3 Analysis of detected BMMF reads at subgroup level

3.3.1 Division of the four main BMMF groups into subgroups

Previous BMMF screening of the detection levels of the four main BMMF groups revealed significant differences between the detection of the four BMMF groups. BMMF1 detection dominates, BMMF2 reads are detected frequently, but at lower levels, whereas BMMF group 3 and 4 are only sparsely detected both on RNA and DNA level. While BMMF groups 3 and 4 comprise three respectively one isolate, the reads reported for BMMF1 and BMMF2 summarizes the detected of more than 50 respectively of more than 100 different genomes. Consequently, there is a more detailed analysis required to identify, which genomes contribute to the BMMF1 and BMMF2 signals detected. The high sequence identity between many BMMF isolates within BMMF group 1 or 2 often complicates the unambiguous assignment of BMMF reads to one single BMMF genome. Especially the BMMF1 isolates derived from colorectal tissue samples often differ only by few mutations in the entire sequence. This results

in reads covering highly conserved BMMF regions being assigned to multiple isolates. To address this issue, I decided to split the two large BMMF groups 1 and 2 into different subgroups based on phylogenetic timetrees to analyze BMMF clusters with reasonable differences on the nucleotide sequence level.

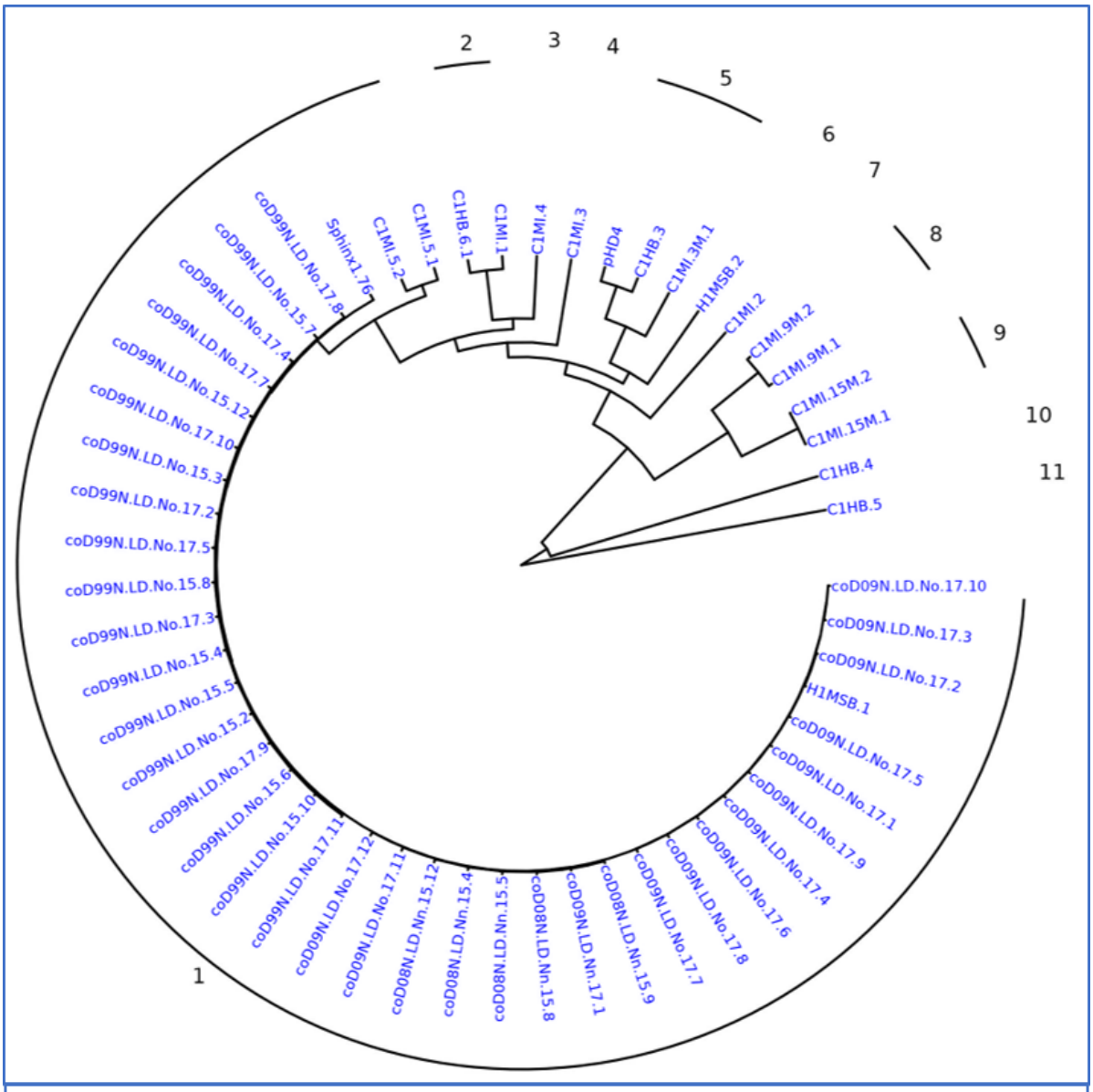


Fig. 3.13: BMMF1 phylogeny: Timetree of BMMF group 1 genomes divided in 11 different subgroups. The small BMMF1 isolates C1MIs.3M.1 and C1HB.6.2 were excluded from the phylogenetic analysis due to the different genome size and manually assigned to the subgroups defined upon the phylogenetic analysis. Based on MUSCLE and BLAST analysis C1HB.6.2 was added to subgroup 2 and C1MIs.3M.1 to subgroup 8. The UPGMA tree was generated based on a muscle alignment using MEGA 11.

The 57 isolates of BMMF group 1 were split into eleven different BMMF1 subgroups. Subgroup 1 is by far the largest BMMF1 subgroup, since it not only contains all colorectal tissue isolates (all isolate names starting with “co…”), but also two cow milk isolates

(C1MI.5.1, C1MI.5.2), the human multiple sclerosis brain isolate 1 (H1MSB.1) and the Sphinx1.76 isolate (fig. 3.11). Despite the large size of this subgroup, the isolates included in subgroup 1 share sequence identity levels of nearly 100% with each other and form a large cluster on a percent identity matrix of BMMF group 1 (S9). The remaining ten BMMF1 subgroups comprise between one and three BMMF genomes (fig. 3.11). The percent identity matrix calculated for BMMF group 1 confirms these subgroups and shows sequence identities larger than 95% within the respective subgroups (S9). However, the isolates of the adjacent subgroups 1 and 2 as well as 8 and 9 also share sequence identities of between 90% and 95% with each other and even more distant subgroups still exhibit sequence identities of 80% to 90% to the other BMMF subgroups (S9). Consequently, the subgroups can resolve many, but not all reads, that cannot be clearly assigned to one BMMF genome, since there are regions on the BMMF genome with matching sequences between different subgroups. The BMMF1 isolates C1MIs.3M.1 and C1HB.6.2 are not included in figure 3.11. These two BMMF1 genomes were excluded from the initial phylogenetic analysis, since their much smaller genome sizes disrupted their assignment to a subgroup. Based on manual MUSCLE and BLAST analysis, C1HB.6.2 was added to BMMF1 subgroup 2, whereas C1MIs.3M.1 was placed in subgroup 8.

The 112 BMMF group 2 genomes were also split into 13 different subgroups, which I named subgroup 12 to 24 to continue the subgroup names of the BMMF1 subgroups (fig. 3.12). BMMF group 2 comprises almost twice as many isolates as BMMF

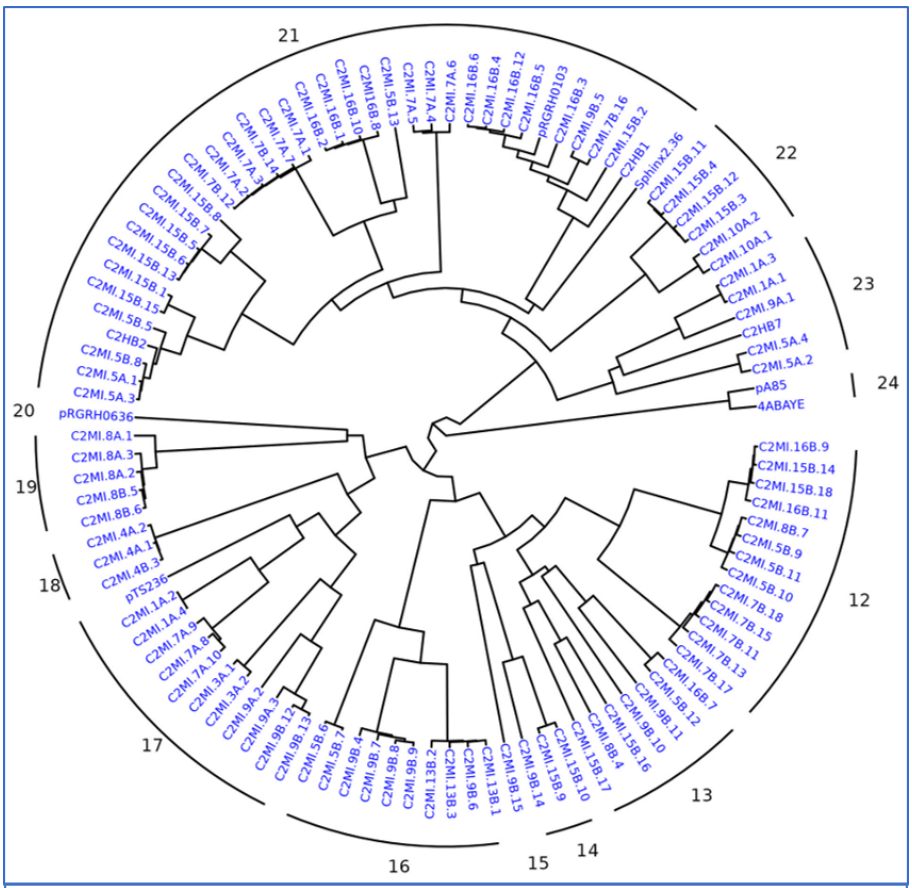


Fig. 3.14: BMMF2 phylogeny: Timetree of BMMF group 2 genomes divided in 13 different subgroups. The small BMMF1 isolates C2MI.10As.1, C2MI.5As.1, C2MI.7As.1, C2MI.9As.2, C2MI.9Bs.4 and C2MI.9Bs.6 were excluded from the phylogenetic analysis due to the different genome size and manually assigned to their subgroup. The UPGMA tree was generated based on a muscle alignment using MEGA 11.

group 1. For BMMF group 2 several large clusters of BMMF2 genomes have been defined as subgroups, however there were also few very small BMMF2 subgroups determined with only up to three isolates (fig. 3.12). The percent identity matrix generated for the genomes of BMMF group 2 confirms the separation into 13 different subgroups, but also shows lower sequence identities between genomes assigned to the same group than observed for the subgroups of BMMF group 1 (S10). Most sequences within the same subgroups share sequence identities larger than 90 %, however the largest subgroup 21 also contain sequences that are more distantly related to their other group members (S10). On the other hand, sequences assigned to different groups are not as closely related as in BMMF group 1, which might facilitate the unambiguous assignment of BMMF group 2 reads to their respective subgroups. Just as BMMF group 1, BMMF group 2 also contains six small genomes, that were manually assigned to the generated subgroups and that are thus not represented in figure 3.12.

3.3.2 Detection of BMMF subgroups on RNA and DNA level

The subgroups defined for BMMF1 and BMMF2 genomes were applied to the D-ViSioN results of all five data sets to examine which subgroups are most frequently detected regarding absolute read numbers as well as which subgroups show the highest normalized BMMF signal.

3.3.2.1 Detection of BMMF subgroups in PCAWG RNA data

In total, 878 BMMF reads were found in the PCAWG RNA data set, 729 BMMF1 reads and 149 BMMF2 reads. 5.1% of the reads detected could not be unambiguously assigned to one specific BMMF cluster. Upon splitting these numbers between BMMF1 and BMMF2 reads, 5.9% of the BMMF1 reads detected are so-called unclear BMMF1 reads (fig.3.15: “?/1”), whereas only 1.3% of the BMMF2 reads are classified as unclear (fig. 3.15: “?/2”). Looking at the number of BMMF hits detected for the different cohorts, the US acute myeloid leukemia cohort (LAML-US) of the PCAWG RNA data set contained both the highest number of hits detected per subgroup and the broadest range of different subgroups detected (fig. 3.15). BMMF1 subgroup 10 is covered by the the highest number of reads by far. Since BMMF1 subgroup 10 contains only one BMMF genome – C1HB.4 – this sequence is the most frequently detected BMMF genome in the LAML-US data. Additionally, BMMF2 subgroup 21 is covered by the second highest number of reads detected in this cohort (fig. 3.15). Since subgroup 21 is the largest BMMF2 subgroup, these reads can be caused by a number of different closely related BMMF genomes (fig. 3.14). Besides of these two signatures, the unclear BMMF1 column and BMMF1 subgroup 6 also show higher numbers of BMMF hits compared to the other BMMF

subgroups within the LAML-US cohort. Only BMMF1 subgroups 1 and 4 have not been found in this cohort, whereas the BMMF2 reads only cover three subgroups and a few additional unclear hits (fig. 3.15). The remaining PCAWG RNA cohorts show much lower BMMF read numbers compared to LAML-US. The BMMF reads detected in ovarian (OV-US), stomach (STAD-US) and uterine (UCEC-US) cancer also cover between four and seven different BMMF1 subgroups – BMMF1 subgroups 5 and 6 are the only subgroups detected in all three of these cohorts.

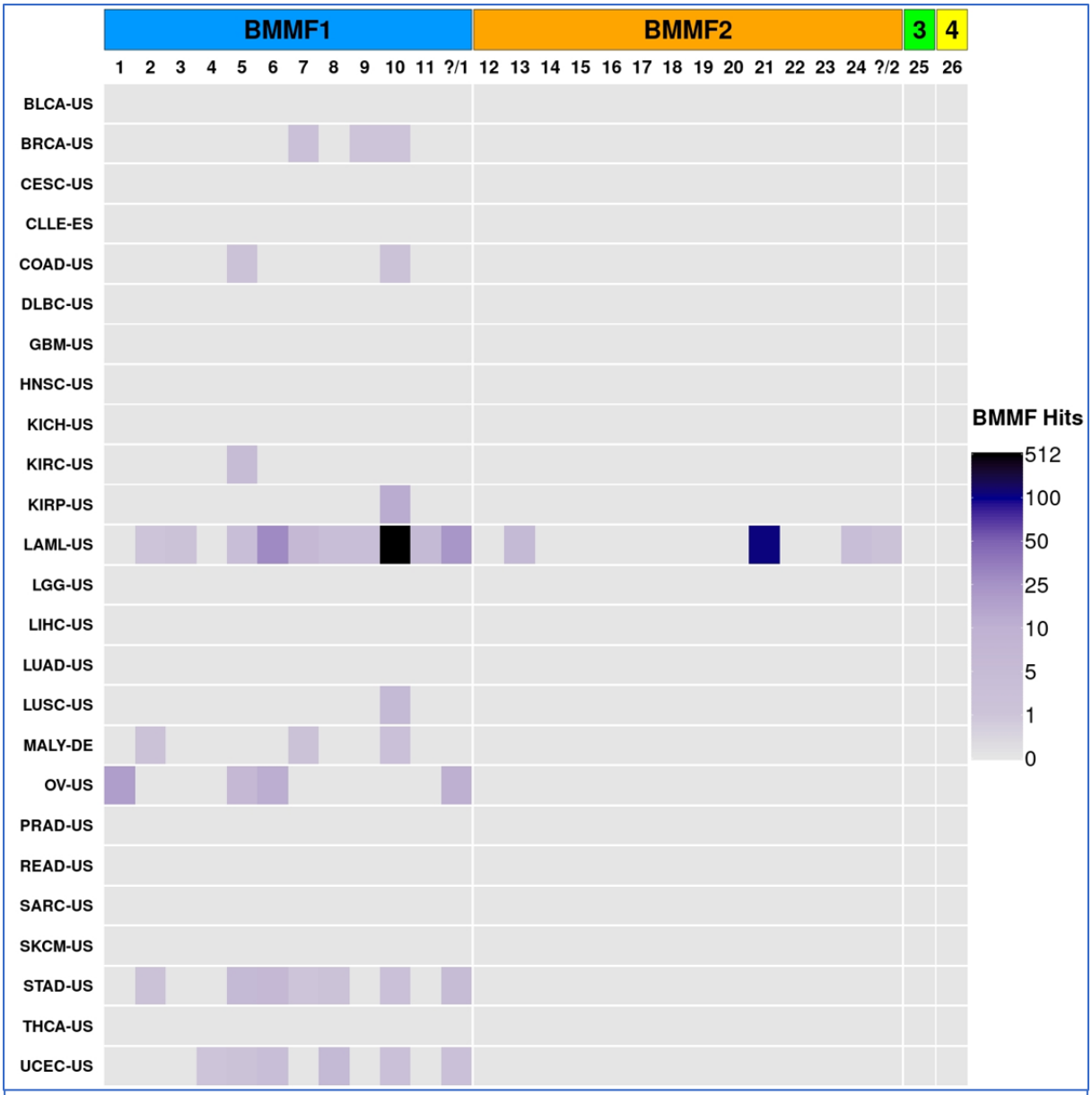


Fig. 3.15: BMMF subgroups detected in all cancer cohorts of PCAWG RNA data: Overview of BMMF reads detected in PCAWG RNA data cohorts after application of the three-hits threshold. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

In breast cancer (BRCA-US) and malignant lymphoma (MALY-DE) three different BMMF1

subgroups were found. BMMF2 reads have only been detected in the LAML-US cohort. Interestingly, there have been four blood cancer cohorts combined for the "blood"- PCAWG-RNA data discussed in the previous chapter – LAML-US, malignant lymphoma (MALY-DE), chronic lymphocytic leukemia (CLLE-ES) and lymphoid neoplasm diffuse large B-cell lymphoma (DLBC-US). The strong BMMF signal observed for blood cancers is mainly caused by the LAML-US data. In the MALY-DE cohort, there were a few BMMF1 reads detected including reads assigned to subgroup 10, which was the most frequently detected subgroup in the LAML-US cohort. The two remaining blood cancer cohorts were BMMF negative.

3.3.

Looking at the normalized BMMF read numbers, the LAML-US cohort again stands out as the cohort with the highest number of BMMF reads per billion mapped reads per 100 samples detected, followed by the OV-US and UCEC-US cohorts (fig. 3.16). BMMF subgroup 10 representing the BMMF genome C1HB.4 is not only the most frequently detected subgroup across all cohorts of the PCAWG RNA data set, but also in several cohorts the subgroup that is covered by the strongest BMMF signal. There are only two cohorts with BMMF positive samples – OV-US and the kidney cancer cohort KIRP-US – in which subgroup 10 was not detected. Besides of BMMF1 subgroup 10, the BMMF1 subgroup 5 was the second most frequently detected subgroup, which was found in 6 different PCAWG RNA cohorts.

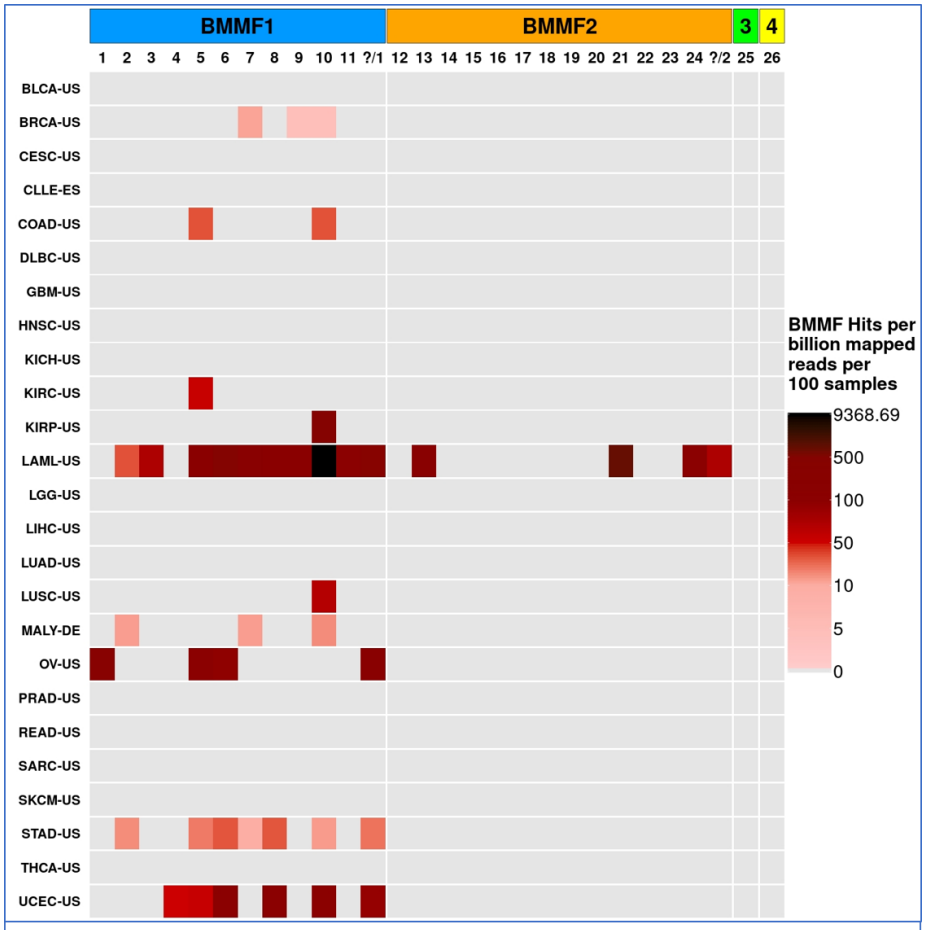


Fig. 3.16: BMMF reads detected for BMMF subgroups in PCAWG RNA data normalized for sequencing depth and a cohort size of 100 samples: Overview of BMMF reads detected in PCAWG RNA data cohorts after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

3.3.2.2 Detection of BMMF subgroups in TCGA RNA data

In the TCGA RNA data set, BMMF1 subgroup 10 is the only subgroup detected in all five cancer types (fig. 3.17). Consequently, this subgroup stands out in both RNA sequencing data sets of cancer patients in a range of different cancer types. Additionally, this subgroup also shows the strongest BMMF signal in both the TCGA colorectal and pancreatic cancer cohorts compared to all other subgroups in all five cancer types (fig. 3.17). The highest number of raw BMMF reads without normalization can be observed in the TCGA breast cancer cohort, in which eight of eleven BMMF1 subgroups were identified (fig. 3.17, S11). Upon normalization, the colorectal, lung and pancreatic cancer cohorts shows higher levels of BMMF reads detected per billion reads per 100 samples compared to the breast cancer cohort. However, the normalized BMMF signal in the TCGA RNA data set is much weaker than in the PCAWG RNA data (fig. 3.16/3.17). Just as in case of the PCAWG RNA data, the majority of the detected reads could be unambiguously assigned to BMMF subgroups. 7.1% of the total amount of reads detected are classified as unclear reads, that cover BMMF group 1, but cannot be clearly assigned to one of the subgroups or one specific BMMF genome (S11).

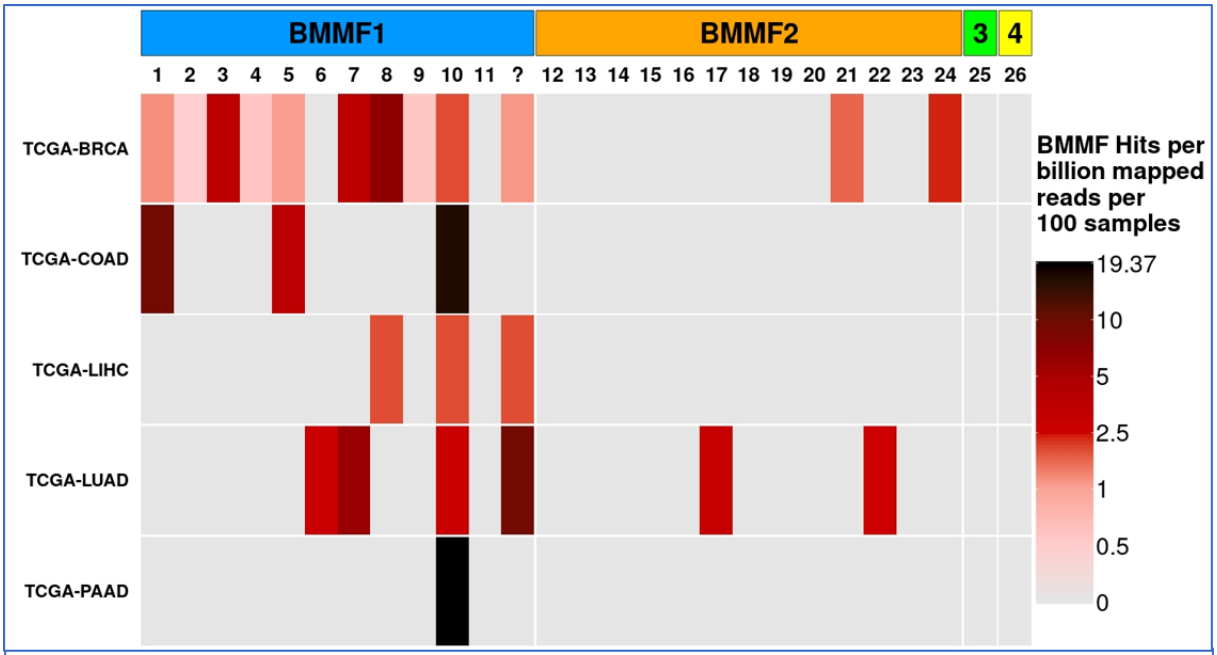


Fig. 3.17: BMMF reads detected for BMMF subgroups in TCGA RNA data normalized for sequencing depth and a cohort size of 100 samples: Overview of BMMF reads detected in TCGA RNA data cohorts after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

3.3.2.3 Detection of BMMF subgroups in GTEx RNA data

The GTEx RNA data shows a broader coverage of different BMMF subgroups for both BMMF group 1 and BMMF group 2 compared to the BMMF reads detected in the two RNA sequencing data sets of cancer patients (fig. 3.18). The number of reads detected per subgroup per cohort is at a similar level compared to the PCAWG RNA data, but higher than in the TCGA RNA cohorts. Compared to the two other RNA sequencing data sets, there was a higher fraction of unclear reads reported for the GTEx data set. 13.6% of the total number of reads detected could not be assigned to a subgroup. This affects mostly the classification of BMMF1 reads, since 15.2% of all BMMF1 reads are unspecific, whereas this is only the case for 8.6% of the BMMF2 reads. The highest number of both BMMF1 and BMMF2 hits per cohort was found in the breast mammary tissue data. The esophagus mucosa and lung tissue cohorts also exhibit higher levels of BMMF hits detected. In contrast to the PCAWG RNA data, the blood GTEx RNA data contained the lowest number of BMMF hits of all tissue cohorts besides of the negative kidney cohort (fig. 3.18).

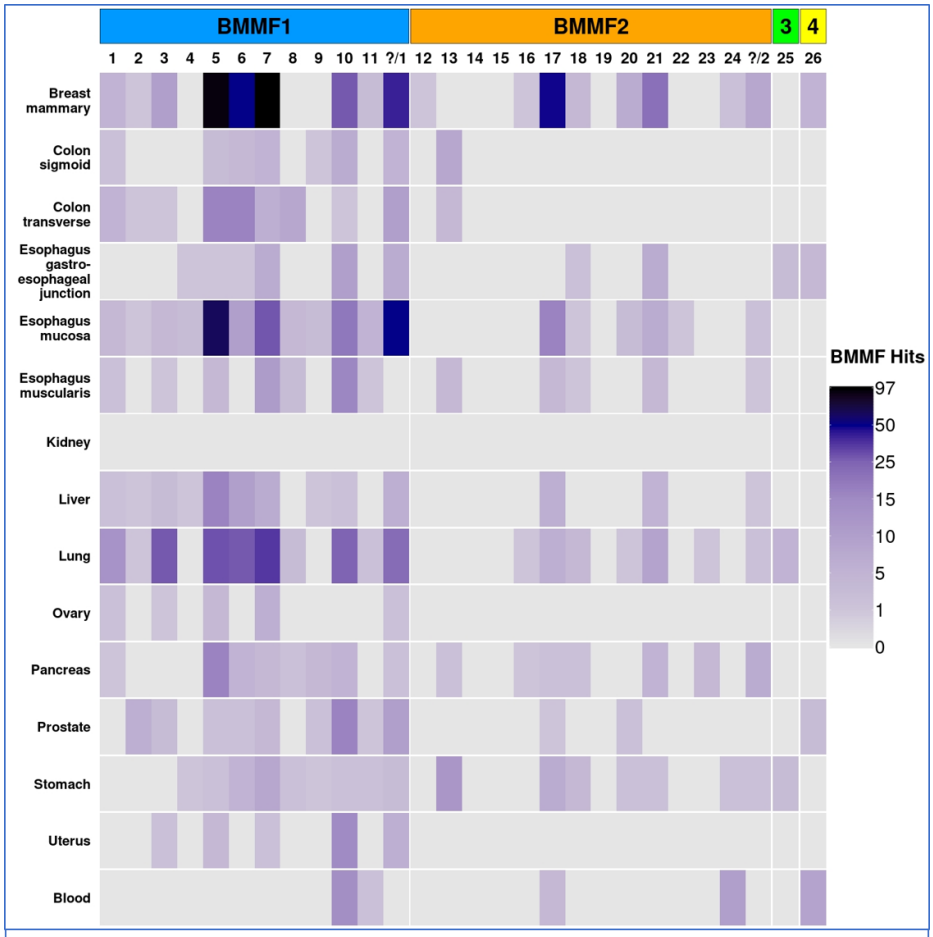


Fig. 3.18: BMMF subgroups detected in all cancer cohorts of GTEx RNA data: Overview of BMMF reads detected in GTEx RNA data cohorts after application of the three-hits threshold. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

The normalization of the detected BMMF reads for sequencing depth and cohort size confirms the low BMMF signal in the GTEx blood data (fig. 3.19). However, the normalization empowers the BMMF positivity data for the subgroups 5,6 and 7 in several tissues such as breast, liver and prostate. Additionally, group 5 and 7 were detected with a strong signal in the

uterus cohort, which does not contain any reads assigned to subgroup 6 (fig. 3.19). Besides of

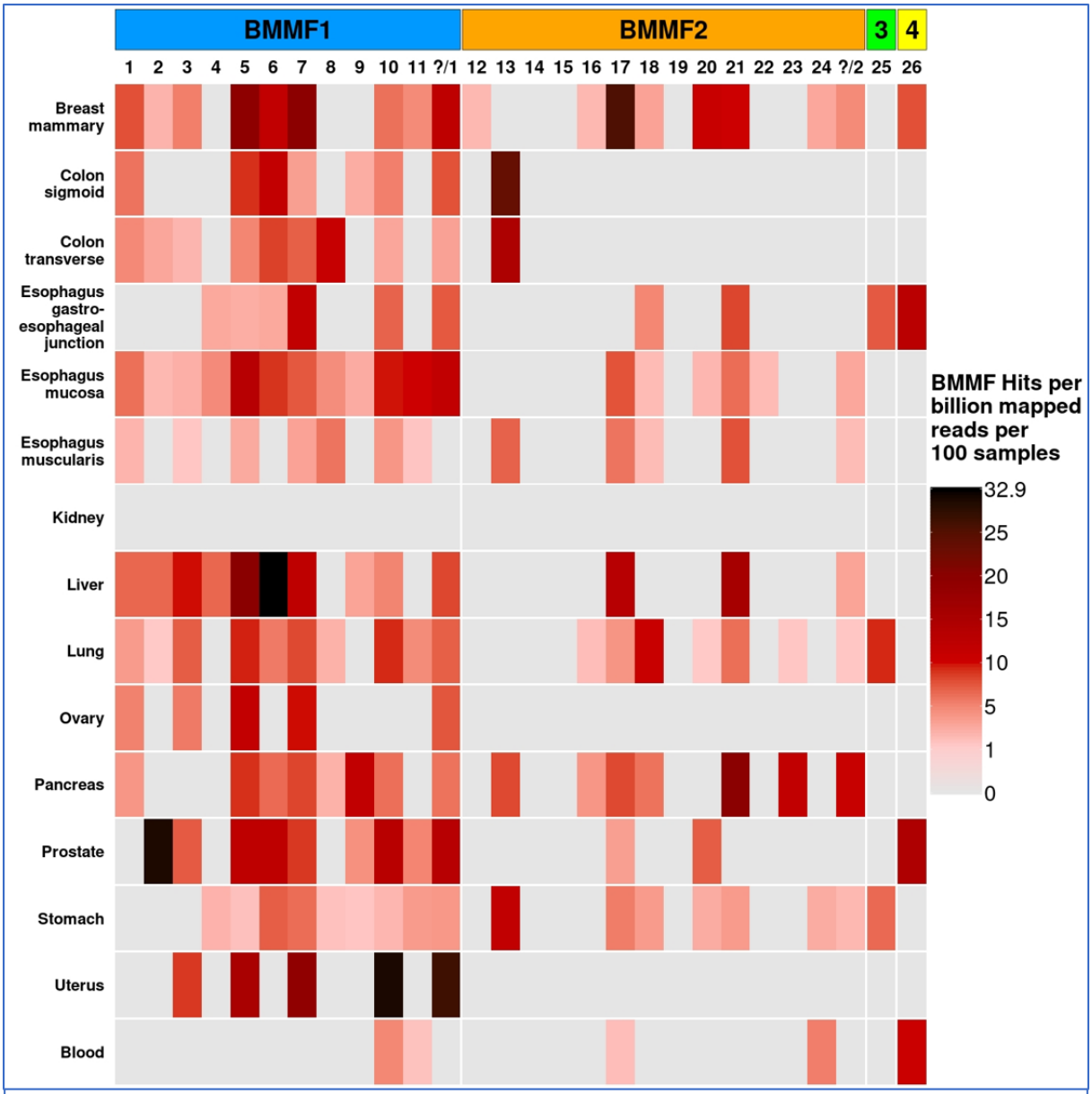


Fig. 3.19: BMMF reads detected for BMMF subgroups in GTEx RNA data normalized for sequencing depth and a cohort size of 100 samples: Overview of BMMF reads detected in GTEx RNA data cohorts after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

this, the uterus cohort stands out with the strongest signal for subgroup 10 as well as for unclear BMMF1 reads. Subgroup 2 is covered with a strong BMMF signal in the GTEx prostate data, whereas this subgroup is only weakly represented in the remaining GTEx cohorts as well as in the PCAWG RNA and TCGA data. The BMMF2 subgroups with the highest BMMF positivity include subgroup 13, 17 and 21. While subgroups 17 and 21 are detected in a broad range of different tissues, subgroup 13 is found in tissues linked to the digestive system – especially in the two colon cohorts and in the stomach data (fig. 3.19). Comparing the numbers of BMMF

reads detected per billion mapped reads per 100 samples reported for the GTEx cohorts, the BMMF signal is on average elevated in the positive cohorts of the PCAWG RNA data set when compared to GTEx, whereas the GTEx data shows positivity in a higher number of different tissue types and subgroups.

3.3.2.4 Detection of BMMF subgroups in PCAWG WGS tumor data

In the tumor whole genome sequencing data of the PCAWG project, 6902 BMMF reads have been detected. 14.8% of these reads could not be unambiguously assigned to a BMMF subgroup. While this includes 17.5% of the BMMF1 reads detected, only 1.4% of the BMMF2 reads are classified as unclear reads. The highest number of BMMF hits was found in the European breast cancer cohort (BRCA-EU) and in the Canadian prostate cancer cohort (PRAD-CA) (S12). While both of these cohorts exhibit high BMMF1 read numbers, the prostate cancer data also included the most BMMF2 and BMMF4 hits detected for any cancer cohort of the PCAWG WGS tumor data set. Besides of the Canadian prostate cancer cohort and the European breast cancer cohort, the PCAWG project also provides WGS tumor data of prostate and breast cancer patients of two other geographical locations. While the breast cancer tumor samples originating from the UK (BRCA-UK), contain hits of several different BMMF1 and BMMF2 subgroups, the US breast cancer samples (BRCA-US) are BMMF negative after application of the three-hits threshold. The three prostate cancer cohorts and the early onset prostate cancer cohort also show very different levels of BMMF detection: A high number of BMMF hits in the Canadian data, weaker BMMF detection in the prostate cancer data from the UK (PRAD-UK) and in the early onset prostate cancer data from Germany (EOPC-DE) and no detection events in the US prostate cancer cohort (PRAD-US) (S12).

Besides of these two cancer types, BMMF hits are also detected in several other cohorts, especially in the lung squamous cell carcinoma (LUSC-US), the stomach adenocarcinoma (STAD-US), the Australian ovarian cancer (OV-AU) and the Canadian pancreatic cancer (PACA-CA) cohorts. Additionally, I detected a broad range of BMMF1 subgroups in two of the three kidney cancer cohorts: in the kidney chromophobe cohort (KICH-US) and in the kidney renal clear cell carcinoma cohort (KIRC US) (S12). The third kidney cancer cohort derived from kidney renal papillary cell carcinoma samples (KIRP-US) is however BMMF negative, just as the second pancreatic cancer cohort originating from Australia (PACA-AU) and the lung adenocarcinoma cohort (LUAD-US). Additionally, the WGS data of the colorectal cancer tumor cohort (COAD-US) is BMMF negative in contrast to the PCAWG RNA colon cancer data (S12).

Looking at the normalized BMMF read numbers, both BMMF positive breast cancer cohorts stand out with a strong BMMF1 signal, whereas the BMMF2 signal is less prominent (fig. 3.20). Both breast cancer cohorts show peaks for the same BMMF1 subgroups: Subgroup 5 shows the strongest signal in both cohorts, followed by subgroups 6, 8 and 1. Additionally, a considerable fraction of BMMF1 hits is classified as "unclear" in both cohorts (fig. 3.20, S12).

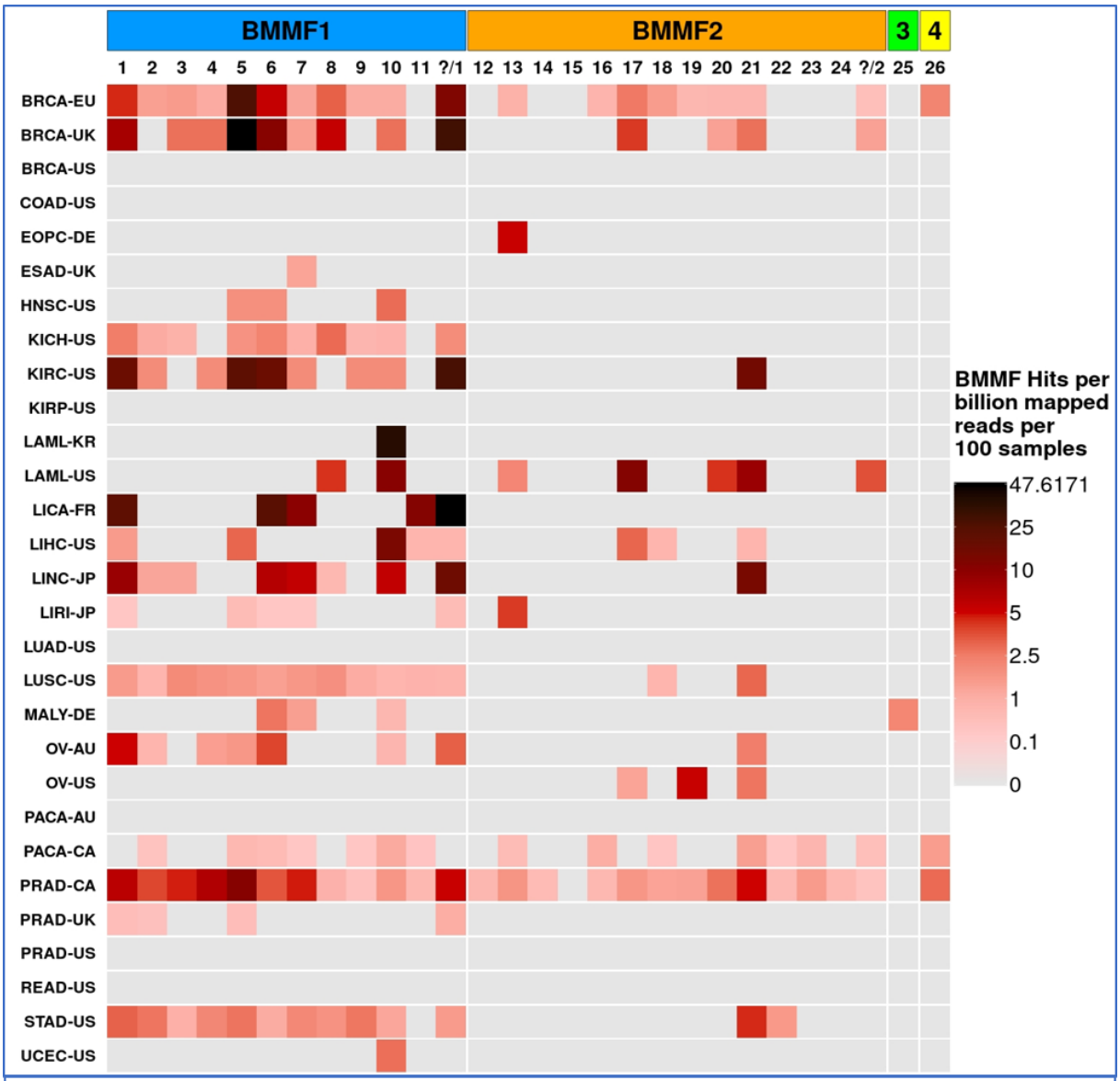


Fig. 3.20: BMMF reads detected for BMMF subgroups in PCAWG WGS tumor data normalized for sequencing depth and a cohort size of 100 samples: Overview of BMMF reads detected in PCAWG WGS tumor data cohorts after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

The kidney renal clear cell carcinoma cohort shows higher normalized hit numbers compared to the other BMMF positive kidney cancer cohort KICH-US, especially BMMF1 subgroups 1,

5 and 6 as well as BMMF2 subgroup 21 are prominently detected. The two acute myeloid leukemia cohorts were less eye-catching looking at the absolute hit numbers, however considering the normalized read numbers, BMMF1 subgroup 10 stands out in both the South Korean and the US LAML cohort (fig. 3.20). Additionally, the LAML-US cohort shows elevated positivity in BMMF subgroup 8 and BMF2 subgroups 17 and 21. The latter BMMF2 subgroup as well as BMMF1 subgroup 10 were also standing out in the PCAWG RNA LAML-US data. The four different liver cancer cohorts originating from France, the US and Japan show different shades of positivity as well as different patterns of BMMF subgroups detected. The small French liver cancer cohort exhibits the highest amount of normalized BMMF reads detected, with a strong BMMF1 signal found in subgroups 1, 6, 7, 11 and for unassigned BMMF1 hits. Especially the detection of subgroup 11 is noteworthy for this cohort, since this subgroup was only sparsely detected in all other PCAWG WGS tumor cohorts. On the other hand, the LICA-FR cohort is the only liver cancer cohort with no BMMF2 signal detected. The first Japanese liver cancer cohort LINC-JP shows the second strongest BMMF signal of the four liver cancer cohorts. Just as in the LICA-FR data, BMMF1 subgroups 1, 6, 7 and the BMMF1 unclear hits column are highlighted, whereas BMMF subgroup 11 is missing in this cohort. However, BMMF1 subgroup 10 and BMMF2 subgroup 21 appear in the LINC-JP data with a dark band in contrast to the LICA-FR cohort. The US liver hepatocellular carcinoma cohort (LIHC-US) as well as the second Japanese cohort (LIRI-JP) show lower levels of BMMF signal detected compared to the other two liver cancer cohorts. In the LIHC-US data BMMF1 subgroup 10 is most prominently detected, whereas the LIRI-JP cohort shows the highest signature for BMMF2 subgroup 13.

Just as the different liver cancer cohorts, the ovarian cancer cohorts from the US and Australia also exhibit different patterns of BMMF detection. Whereas mostly BMMF1 subgroups are detected in the OV-AU data with emphasis on subgroups 1 and 6, I found only BMMF2 hits in the OV-US cohort. The US lung squamous cell carcinoma cohort shows a broad detection of all BMMF1 subgroups as well as for BMMF2 subgroups 18 and 21, however without any subgroup standing out with an increased signal. A similar pattern can be observed for the US stomach adenocarcinoma and the Australian pancreatic cancer, which both cover a broad range of BMMF1 subgroups however without clear standouts. The only exception to this, is the BMMF2 subgroup 21 signal in the STAD-US data.

While the Canadian prostate cancer data set impressed with the highest number of absolute reads across all PCAWG WGS tumor cohorts, the normalized BMMF read counts per billion

mapped reads per 100 samples are not among the highest numbers identified in the entire data set. However, the BMMF signal detected in the PRAD-CA cohort is distributed among all BMMF1 subgroups as well as all but one BMMF2 subgroups. No other PCAWG WGS tumor cohort covers such as broad range of different BMMF subgroups. While some subgroups are only represented with a weaker signal in the PRAD-CA data, BMMF1 subgroups 1-7 and BMMF2 subgroup 21 are highlighted with stronger signals -especially in case of subgroup 5.

Across the entire PCAWG WGS tumor data set, BMMF1 subgroups 1, 5, 6, 7 and 10 are frequently standing out with a strong BMMF signal as well as BMMF2 subgroups 13, 17 and 21 regarding the BMMF2 reads detected.

3.3.2.5 Detection of BMMF subgroups in PCAWG WGS normal data

In total, there were 2335 BMMF reads identified in the PCAWG WGS normal tissue/blood data, which is approximately one third of the total read number detected in the PCAWG WGS tumor data. 14.8% of the PCAWG WGS normal tissue/blood reads could not be unambiguously assigned to a BMMF subgroup, which is the same percentage of unclassified reads as found in the PCAWG WGS tumor data. 16.4% of the BMMF1 reads are characterized as unclear, but only 1.3% of the BMMF2 reads. Just as in the PCAWG WGS tumor data, the BRCA-EU and PRAD-CA cohorts contain the highest number of BMMF reads detected across all cohorts in the PCAWG WGS normal tissue/blood data (S13). The LUSC-US, OV-AU, PACA-CA and STAD-US cohorts also stand out with hits reported for a broad range of different BMMF1 subgroups like in the PCAWG WGS tumor data set (S12, S13). However, in contrast to the tumor data, the normal samples of the early onset prostate cancer cohort (EOPC-DE) as well as of the South Korean LAML cohort show higher hit number, that cover several different subgroups. Additionally, the US breast cancer cohort, the US colon cancer cohort, the US lung adenocarcinoma cohort and the US prostate cancer cohort show BMMF positive samples, whereas the cohorts were BMMF negative in case of the respective tumor samples (S13).

In the normalized reads data, the South Korean LAML cohort, the Canadian prostate cancer cohort, the French liver cancer cohort as well as the European and UK breast cancer cohorts exhibit the strongest BMMF signals detected, which cover a broad range of different BMMF1 subgroups as well as several BMMF2 subgroups (fig. 3.21). The intensity of the detected BMMF signal is at a comparable level to the PCAWG WGS tumor data, which is why I used the same color code in the respective heatmaps (fig. 3.20, fig. 3.21). The peak values of the maximum number of normalized BMMF reads detected in the tumor and normal tissue/blood

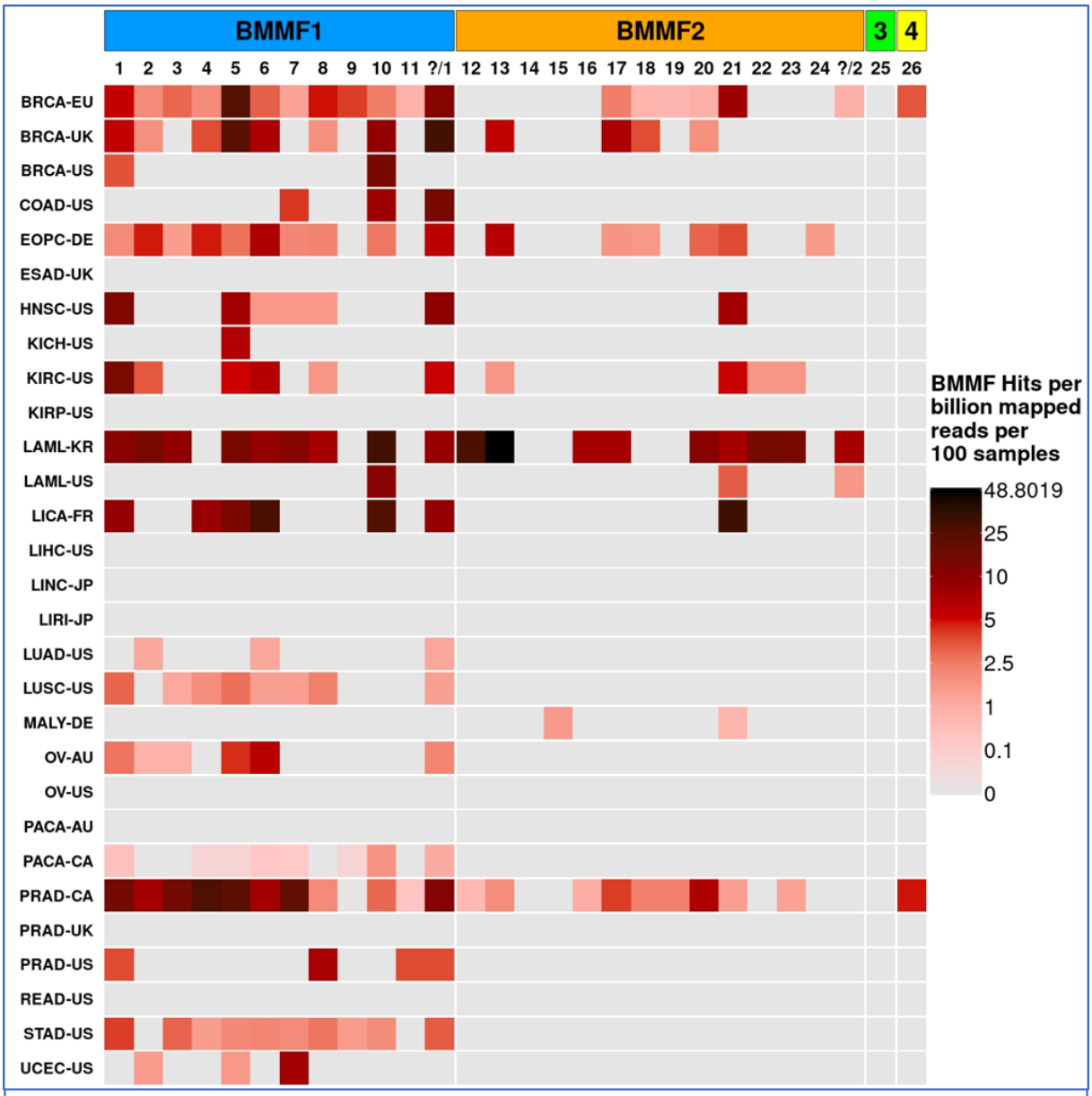


Fig. 3.21: BMMF reads detected for BMMF subgroups in PCAWG WGS normal tissue/blood data normalized for sequencing depth and a cohort size of 100 samples: Overview of BMMF reads detected in PCAWG WGS normal tissue/blood data cohorts after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

data is also nearly identical between the two data sets. Just as in the PCAWG WGS tumor data, BMMF1 subgroup 5 is the subgroup which was most prominently covered in the European and UK breast cancer cohorts, followed by subgroups 1 and 6. The UK breast cancer cohort additionally shows a strong signal for subgroup 10, which also stands out in the US breast cancer cohort. While the US breast cancer cohort did not contain BMMF2 hits, the other two breast cancer cohort exhibit BMMF2 signal albeit for different subgroups. In the BRCA-EU

cohort BMMF2 subgroup 21 dominate the BMMF2 detection, whereas subgroups 13 and 17 showed the strongest BMMF2 signal in the BRCA-UK data.

The normal tissue/blood colorectal cancer data showed positivity for BMMF1 subgroups 7 and 10. Subgroup 10 was also detected in the colorectal cancer data of the PCAWG RNA and TCGA RNA cohorts (fig. 3.16, fig. 3.17). The kidney cancer cohort KIRC-US showed the strongest BMMF signal in subgroups 1, 5, 6 and 21 just as in case of the tumor data of these cancer patients. The KICH-US cohort only showed positivity for subgroup 5, whereas the KIRP-US kidney cancer cohort is BMMF negative in both the tumor and the normal tissue/blood cohorts (fig. 3.21). The South Korean LAML cohort stands out with the detection of a strong BMMF signal in nine BMMF1 and BMMF2 subgroups each. No other PCAWG WGS normal data cohort covers such a broad range of BMMF subgroups with such a strong intensity. The tumor data of this cohort only showed positivity for BMMF subgroup 10, which is also the BMMF1 subgroup with the highest normalized read numbers detected. However, BMMF2 subgroups 12 and 13 show also a very strong signal, with subgroup 13 being the subgroup with the highest amount of BMMF reads detected per billion mapped reads per 100 samples across the entire PCAWG WGS normal tissue/blood data set. In the second acute myeloid leukemia data set – LAML-US – subgroups 10 and 21 are most prominently detected, which is consistent with the observations in the WGS tumor and RNA data of this cohort (fig. 3.20, fig. 3.21).

Three of the four liver cancer cohorts show no BMMF positivity in the normal tissue/blood samples, whereas the French liver cancer cohort stands out with a prominent signal in five BMMF1 subgroups and BMMF2 subgroup 21. Just as in the PCAWG WGS tumor data, the STAD-US, PACA-CA and LUSC-US cohorts show BMMF positivity for several BMMF1 subgroups, however without strong peaks visible. In contrast to the respective tumor data, there can be also a weak BMMF signal observed in the second lung cancer cohort (LUAD-US). Three of the four prostate cancer cohorts show higher normalized read numbers compared to the respective tumor cohorts. While the Canadian prostate cancer cohort belongs to the cohorts with the highest normalized read signal in both the tumor and the normal tissue/blood data sets, the BMMF positivity is on average higher in the subgroups detected in the control data. The early onset prostate cancer cohort from Germany only exhibits a very weak BMMF signal in the tumor data, however many different subgroups are detected in the normal tissue/blood data. The US prostate cancer tumor data is BMMF negative, whereas there is BMMF1 signal detected in the control data of this cohort. On the other hand, no BMMF detected was reported for the

UK prostate cancer cohort in the normal tissue data, whereas the tumor data exhibited weak BMMF positivity.

3.3.2.6 Differentiation between origins of PCAWG WGS non-tumor data

The data included in the PCAWG WGS normal cohort is derived from different origins. 72.5% of the samples are derived from blood samples, whereas most of the remaining normal tissue samples are either taken from solid tissue site distant from the primary tumor or from tissue adjacent to the primary tumor. Since BMMFs are generally expected to be detected at higher frequencies in the peritumor than in the primary tumor itself, samples taken from regions close to the tumor are most important for the analysis of BMMF detection, however only rarely available in sequencing projects. Normal tissue samples were obtained from tissue adjacent to the primary tumor for some European breast cancer samples, several Australian pancreatic cancer and Japanese liver cancer samples as well as for one UK breast cancer sample. All normal tissue samples of the French liver cancer cohort are derived from tissue close to the primary tumor. Five of eleven BRCA-EU normal tissue samples from tissue adjacent to the tumor show BMMF positivity, whereas four of six LICA-FR normal tissue samples contain BMMF reads (fig. 3.22). The samples from tissue adjacent to the tumors of the other cohorts were BMMF negative. Since the strong BMMF signal detected in the normal tissue data of the LICA-FR cohort is derived from samples close to the primary tumor, this might explain the strong BMMF signal in the normal tissue of the LICA-FR cohort in contrast to the negative normal data of the other three liver cancer cohorts, where the normal tissue samples are mostly derived from other origins. The normal tissue/blood BMMF signal reported for the BRCA-EU cohort is partly derived from tissue adjacent to the primary tumor and partly detected in blood samples (fig. 3.22, fig. 3.23). The normal tissue data

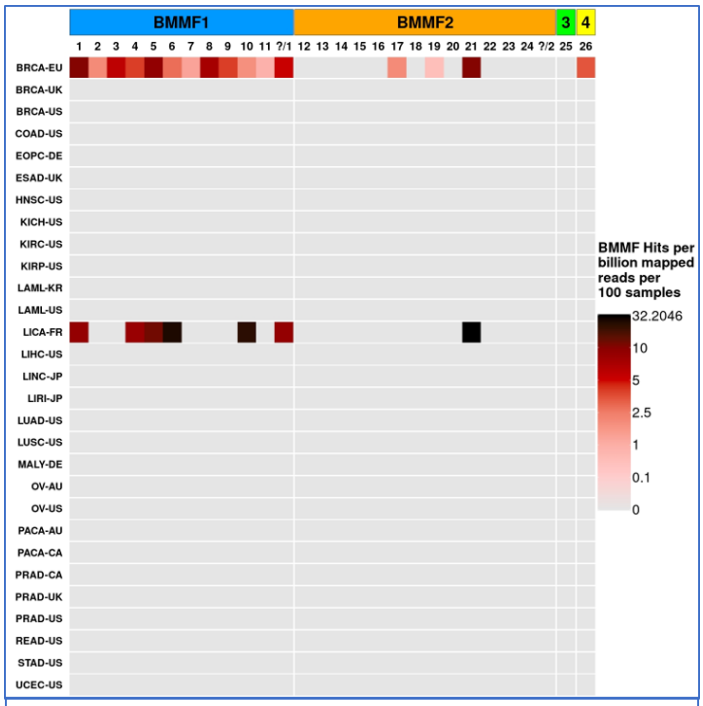


Fig. 3.22: Normalized BMMF reads detected for BMMF subgroups in PCAWG WGS normal tissue data from derived from tissue adjacent to the primary tumor: Overview of BMMF reads detected in PCAWG WGS normal tissue samples originating from tissue close to the tumor after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

obtained from samples close to the tumor is positive for all BMMF1 subgroups as well as for three BMMF2 subgroups. The strongest BMMF signal can be observed in BMMF1 subgroups 1, 5 and 8 and in BMMF2 subgroup 21 (fig. 3.22). The blood BRCA-EU data shows also BMMF positivity for a range of BMMF1 and BMMF2 subgroups, however BMMF1 subgroup 5 is clearly standing out with the highest normalized read numbers detected (fig. 3.23).

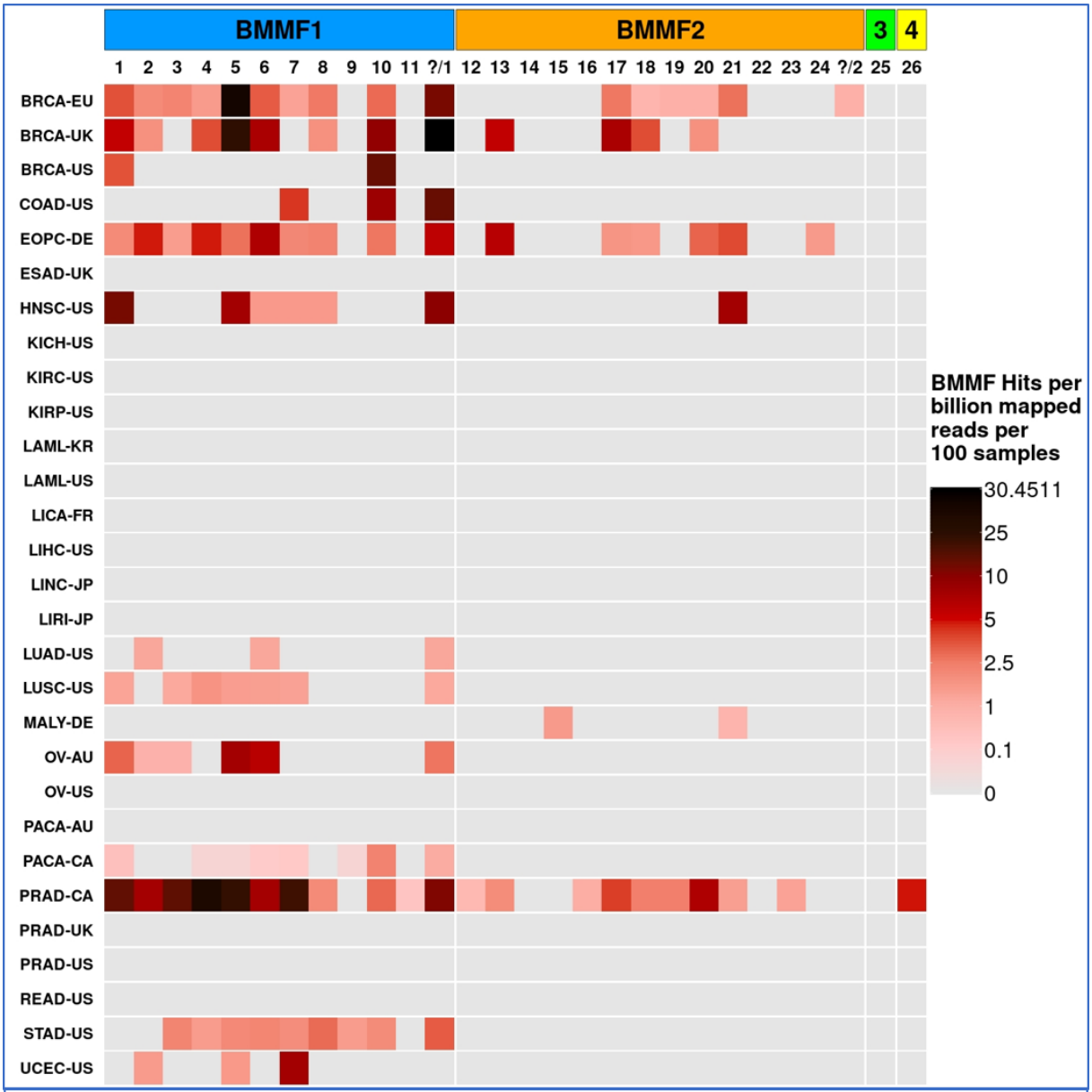


Fig. 3.23: Normalized BMMF reads detected for BMMF subgroups in PCAWG WGS normal blood data: Overview of BMMF reads detected in PCAWG WGS normal blood samples application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

Since 84.75 % of the total PCAWG WGS normal reads were discovered in the blood samples, the pattern of BMMF detection in the blood data resembles the picture of the entire PCAWG

WGS normal data set (fig. 23). However, some cohorts either include no blood samples or show no positivity in blood samples. Besides of the French liver cancer cohort, there are also several other cohorts negative when only blood samples are taken into account such as the KICH-US and KIRC-US kidney cancer cohorts, the US prostate cancer cohorts or both LAML cohorts (fig. 3.21/3.23).

Looking at the normal tissue data derived from solid tissue samples taken from locations distant to the primary tumor, it becomes apparent, that the normal tissue BMMF signal found in the KICH-US, KIRC-US, PRAD-US and LAML-US cohorts is originating from these samples. Additionally, the BMMF signal found in the LUSC-US, PACA-AU and STAD-US cohorts is constituted from reads found both in solid tissue and blood samples (S14).

Some of the cohorts include samples taken from other origins than the three sample sources discussed so far. The most interesting example for this, is the South Korean acute myeloid leukemia cohort. All PCAWG-WGS normal samples included in the LAML-KR data are obtained from saliva. Consequently, the BMMF signal detected in this cohort is not directly comparable to the other cohorts in the PCAWG WGS normal data set. Since BMMFs are expected to be ingested via nutrition, saliva is not a reliable negative control for BMMF research. However, this sample source might explain the remarkably high BMMF signal in the normal LAML-KR data compared to the tumor data.

3.3.3 Detection of BMMF subgroups at patient level in RNA and WGS data

The detection of BMMF subgroups at cohort level is useful to analyze, which subgroups might be linked to a certain cancer type or cohort. However, it is additionally necessary to investigate the results at patient level, to examine if observations made at cohort level are caused by single outliers or if they are supported by multiple detection events across a respective cohort.

3.3.3.1 BMMF subgroup detection in PCAWG RNA patients

Looking at the PCAWG-RNA data, it becomes apparent that the analysis at patient level is important for a better understanding of the results. The PCAWG-RNA data set contains 25 different cancer types, but just 987 samples in total. Consequently, most of the cohorts included in the PCAWG RNA data are quite small. Only six of the 25 cohorts comprise 50 or more samples. Due to this, the BMMF signal detected in several cohorts such as BRCA-US, COAD-US, KIRC-US, KIRP-US or LUSC-US is just caused by one single positive sample (fig. 3.24). In some of these cohorts, additional samples with BMMF reads detected were removed from the final representation because of not meeting the three hits threshold due to an unreliably low

number of BMMF reads per sample and per BMMF group. Thus, there would be a larger sample size needed to take more robust conclusions with regards to the BMMF positivity in these cancer types.

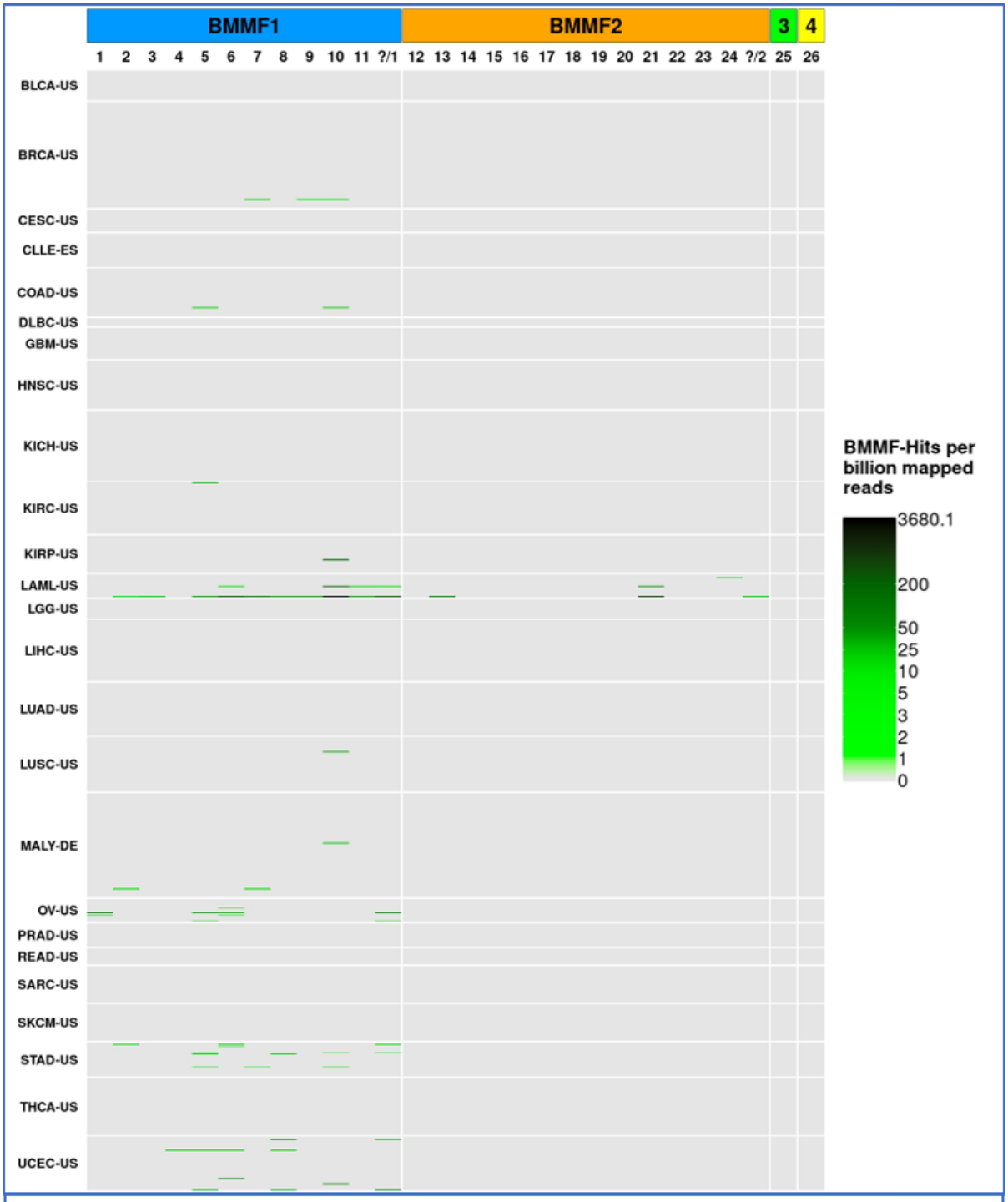


Fig. 3.24: Normalized BMMF reads detected per patient in PCAWG RNA data: Overview of BMMF reads detected in PCAWG RNA data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

However, the PCAWG-RNA data set also includes cancer types, for which three or more BMMF positive samples were reported after applying the three hits threshold. This includes the LAML-US, OV-US, STAD-US and UCEC-US cohorts, which are the same PCAWG RNA cohorts that stood out looking at the absolute and normalized read numbers (fig. 34, fig. 3.15, fig. 3.16). Interestingly, three of these four cohorts belong to the smaller cohorts included in the PCAWG RNA data: The LAML-US cohort consists of 20 samples, the OV-US cohort of 21 samples and the STAD-US cohort of 31 samples. Only the UCEC-US cohort with 49 samples is among the 10 PCAWG RNA cohorts with the highest sample numbers. Analyzing the PCAWG RNA data at patient level, the LAML-US peaks for BMMF1 subgroup 10 and BMMF2 subgroup 21 are both caused by the same two positive samples. However, subgroup 10 was also found samples of several other cancer cohorts. Besides of this, the LAML-US, OV-US, STAD-US and UCEC-US cohorts also contained several samples positive for BMMF1 subgroups 5, 6 and 8 (fig. 3.24).

3.3.3.2 BMMF subgroup detection in TCGA RNA patients

The TCGA RNA data set only includes 5 cancer cohorts, but with much larger sample sizes ranging from 182 in the pancreatic cancer cohort to 1222 samples included in the breast cancer cohort. In spite of this, the number of BMMF positive samples per cohort is low for four of the five cancer types (S15). The TCGA liver (TCGA-LIHC) and pancreatic cancer (TCGA-PAAD) cohorts contain only one BMMF positive sample. Additionally, there were only three, respectively five positive samples reported for the colorectal cancer (TCGA-COAD) and lung adenocarcinoma cohorts (TCGA-LUAD). Only the breast cancer cohort (TCGA-BRCA) contains a higher number of positive samples, however this cohort also includes more than twice as many samples compared to the other TCGA cohorts (S15). BMMF1 subgroup 10 is the only subgroup detected in all five TCGA cancer cohorts. It is also the subgroup that is found the in the highest number of different samples. Besides of subgroup 10, subgroup 7 is also detected in several different samples in the TCGA-BRCA cohort (S15).

3.3.3.3 BMMF subgroup detection in GTEx RNA patients

The GTEx cohort consists of more than 5500 samples. Due to this, the representation at samples level has to be split into three different heatmaps (fig. 3.25, S16, fig. 3.26). The largest GTEx cohort screened for BMMF reads, is the blood cohort, which shows only weak BMMF positivity in a low number of samples (fig. 3.25). The breast mammary tissue cohort showed both higher normalized read numbers per sample and the highest number of positive samples across the

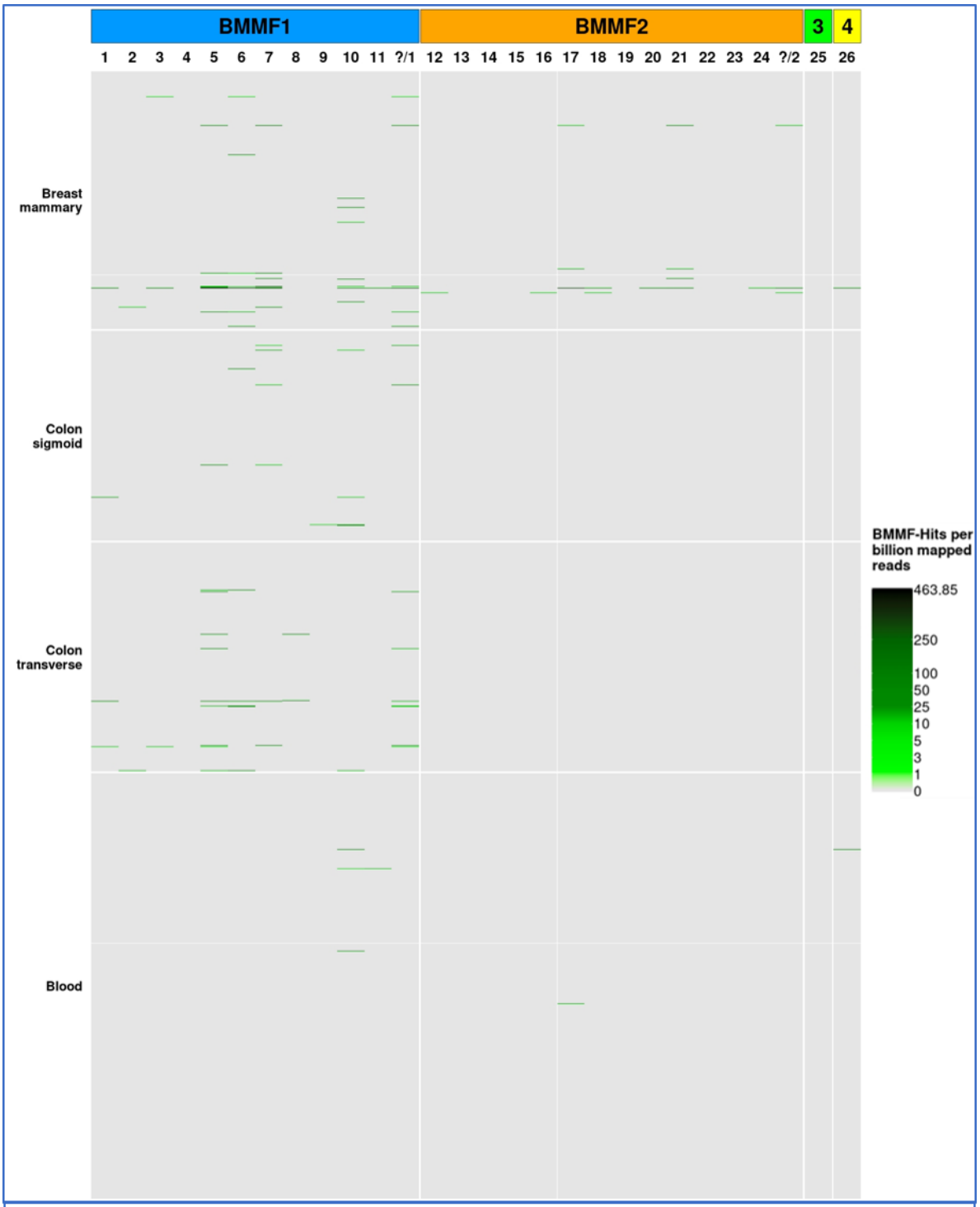


Fig. 3.25: Normalized BMMF reads detected per patient in GTEx RNA data (1): Overview of BMMF reads detected in GTEx RNA data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

entire GTEx data set with 22 positive samples (of 457). While some breast tissue samples only contain BMMF reads of one single subgroup, there are also several samples in this cohort with reads covering three or more subgroups. BMMF1 subgroups 5, 6, 7 and 10 are the only subgroups reported in more than 5 different tissue samples (fig. 3.25). The two colorectal tissue

cohorts as well as the three esophageal cohorts do not exhibit many overlapping patterns of BMMF detection. The colon sigmoid tissue data showed lower levels of BMMF positivity and a lower number of positive samples compared to the colon transverse cohort. While BMMF1 subgroup 5 is found in almost all BMMF positive samples in the colon transverse data, this subgroup is only detected in one colon sigmoid sample (fig. 3.25).

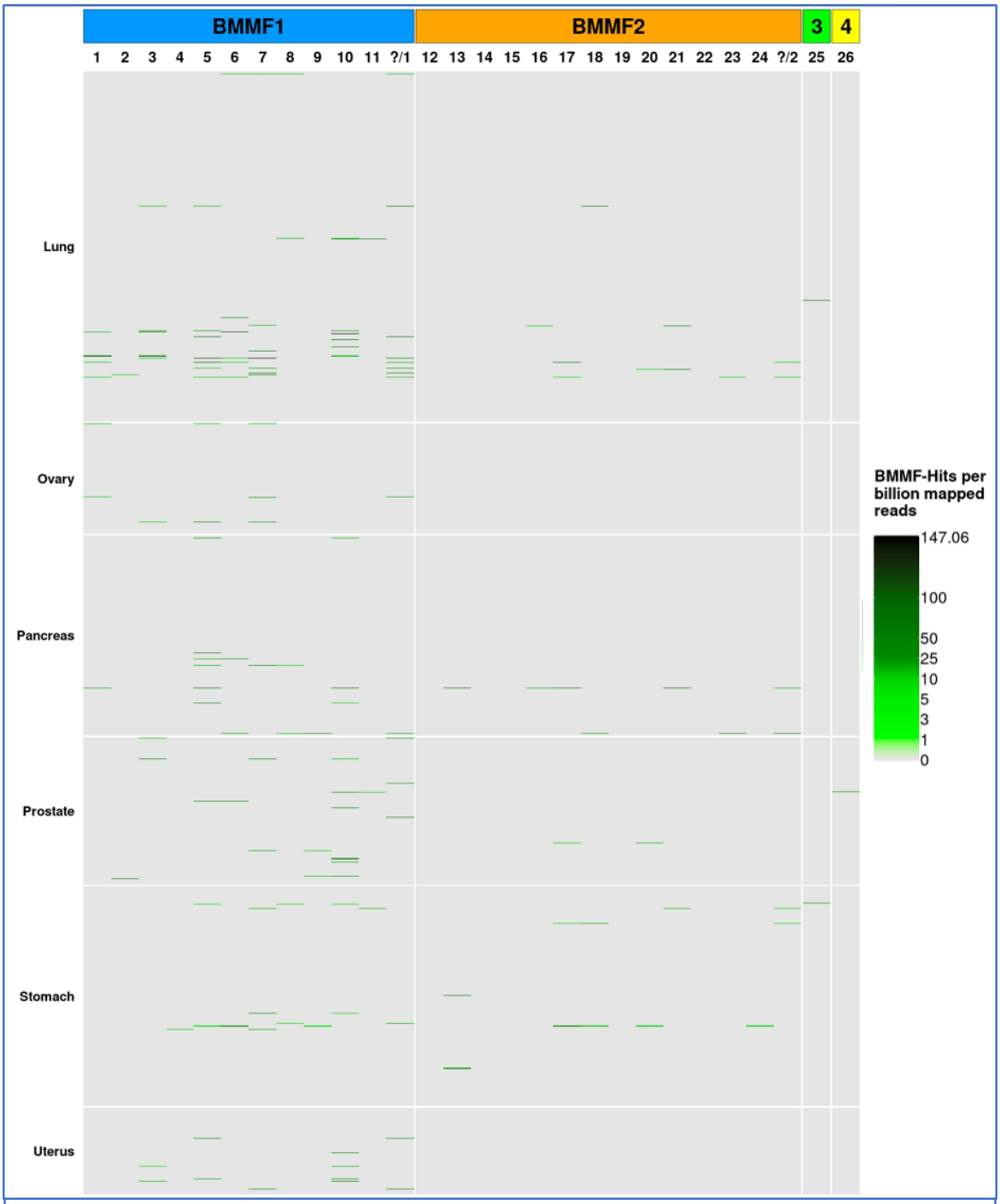


Fig. 3.26: Normalized BMMF reads detected per patient in GTEx RNA data (3): Overview of BMMF reads detected in GTEx RNA data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

In the three esophageal cohorts, BMMF1 subgroup 10 is consistently detected in several samples in all three cohorts. Besides of that, the esophagus gastroesophageal junction and esophagus muscularis cohorts contain low numbers of BMMF positive samples (S16). The esophagus mucosa cohort shows a stronger BMMF signal and 16 positive samples, mostly for BMMF1 subgroups 5, 7 and 10 (S16). The GTEx kidney cohort is BMMF negative, whereas the liver cohort contains a low number of positive samples (S16). The GTEx lung data set contains 22 positive samples, which is the highest number of positive samples together with the breast mammary cohort. In the lung data, BMMF1 subgroups 5, 6, 7 and 10 stand out with high normalized read numbers and with detection across several different samples (fig. 3.26). The prostate GTEx cohort contains 14 BMMF positive samples, which is the fourth highest number across the GTEx RNA data set. However the reads detected in these positive samples do not cover many different BMMF subgroups, only subgroup 10 in more than two different samples (fig. 3.26). Subgroup 10 is also found in four out of six BMMF positive samples in the uterus tissue cohort, whereas subgroup 5 is most often detected in pancreatic tissue samples. The BMMF reads detected in the ovary and stomach GTEx cohorts are distributed across several different subgroups without any preferences for certain subgroups visible (fig. 3.26).

3.3.3.4 BMMF subgroup detection in PCAWG WGS tumor samples

The PCAWG WGS tumor data set exhibits very different BMMF detection patterns compared to the three RNA data sets discussed previously. Not all PCAWG WGS cohorts are BMMF positive, but most PCAWG WGS cohorts with BMMF positive samples exceeding the three hits threshold contain a higher number of positive samples in relation to the sample size than observed in the RNA data (fig. 3.27). Additionally, BMMF positive samples often contain BMMF reads covering several different BMMF subgroups. Consequently, some samples of the BRCA-EU, BRCA-UK, LUSC-US, PRAD-CA and STAD-US cohorts basically represent a horizontal line crossing all or at least most columns of BMMF1 subgroups.

Looking at the BMMF2 subgroups, this can only be observed in the PRAD-CA cohort. In other cohorts, the detected reads favor only one or a low number of different subgroups. The Canadian pancreatic cancer cohort favors the detection of BMMF subgroup 5 and 10, whereas in the Australian ovarian cancer cohort mainly the subgroups 1, 5 and 6 are reported (fig. 3.27). Many cohorts for which high read numbers or a strong normalized BMMF signal have been observed, also exhibit positivity in many different samples such as in the BRCA-EU, PRAD-CA, LUSC-US, STAD-US or PACA-CA cohorts (fig. 3.20/fig. 3.27). However, for some cohorts only one or two samples are causing a strong normalized BMMF signal at cohort level. This can for

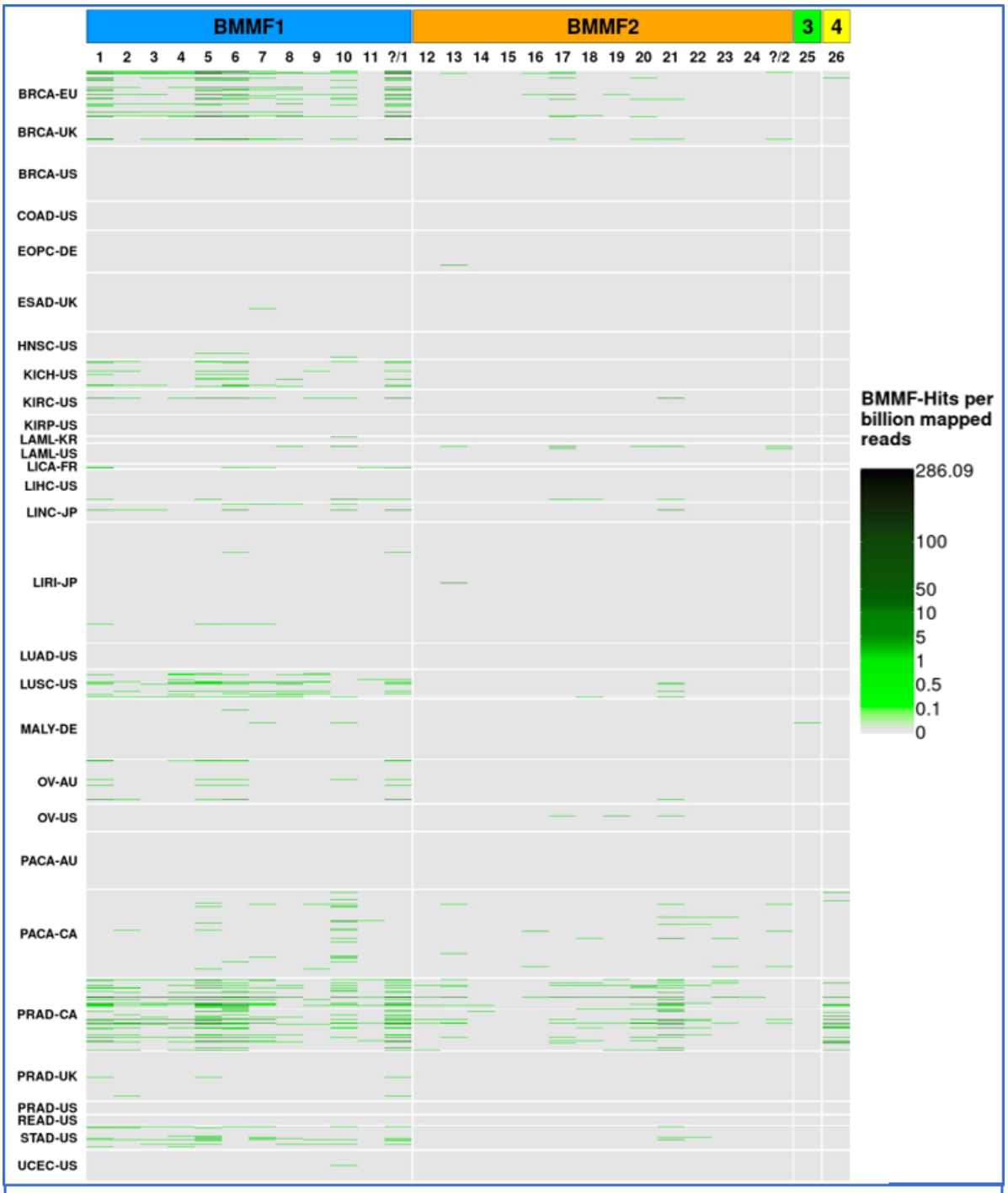


Fig. 3.27: Normalized BMMF reads detected for BMMF subgroups in PCAWG WGS tumor data: Overview of BMMF reads detected in PCAWG WGS tumor data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

example be observed in case of the BRCA-UK, KIRC-US and both LAML cohorts. Additionally, the strong signal in the LICA-FR cohort was also caused by just two samples, however the entire French liver cancer cohort only comprises six samples in total (fig. 3.27).

Across the entire PCAWG WGS data set, BMMF1 subgroups 1, 5, 6 and 10 are detected in the highest number of different samples as well as BMMF2 subgroup 21 (fig. 3.27).

3.3.3.5 BMMF subgroup detection in PCAWG WGS non-tumor samples

The PCAWG WGS normal tissue/blood data set shows lower numbers of positive samples detected in most positive cohorts as well as lower numbers of different BMMF subgroups

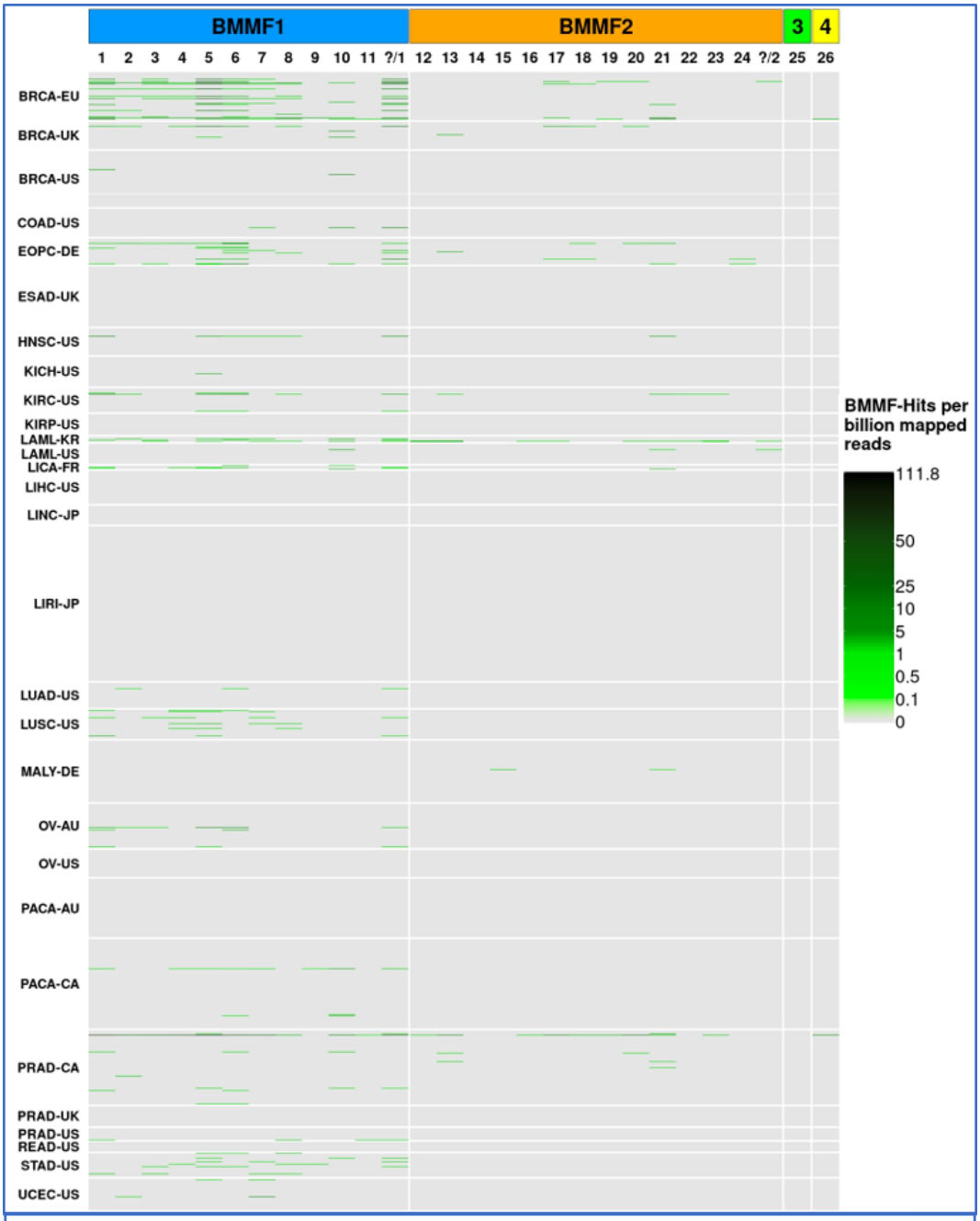


Fig. 3.28: Normalized BMMF reads detected per patient in PCAWG WGS normal tissue/blood data: Overview of BMMF reads detected in PCAWG WGS normal tissue/blood data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

covered by positive samples (fig. 3.28). The biggest outlier to this is caused by the BRCA-EU cohort, which shows both a high number of positive samples and a high coverage of a broad range of BMMF1 subgroups. While BMMF1 subgroups 1, 5 and 6 dominate with high read numbers and high numbers of positive patients, almost all BMMF1 groups are covered by reads obtained from several different non-tumor samples of European breast cancer patients (fig. 3.28). The LUSC-US and STAD-US also show several positive samples, however with lower numbers of positive samples and less BMMF1 subgroups covered than in the PCAWG WGS tumor data of the respective cohorts (fig. 3.27, fig. 3.28). On the other, there are higher numbers of positive patients observed for the German early onset prostate cancer data and the South Korean acute myeloid leukemia data. Whereas subgroups 5 and 6 stand out in case of the EOPC-DE cohort, almost all BMMF1 subgroups are covered by reads in different samples in the small LAML-KR cohort. The opposite can be observed for the Canadian pancreatic cancer cohort and the Canadian prostate cancer data. The BMMF signal of both cohorts is mainly caused by one strongly positive patient with reads covering a broad range of different BMMF subgroups. The tumor data of these cohorts showed much more positive samples. Whereas the tumor PRAD-CA data includes 48 samples with BMMF positivity, the negative control samples only contained ten positive samples with 122 files screened for both tumor and normal tissue/blood data. In case of the Canadian pancreatic cancer data 25 tumor samples are BMMF positive, but only three samples from the PCAWG WGS normal cohort.

3.4 Statistical comparison of BMMF subgroup detection in different data sets

For a more detailed comparison of the BMMF signal detected in tumor and non-tumor or healthy tissue data sets, I determined both the difference between the mean numbers of normalized reads in the tumor and normal tissue data and performed statistical testing for each BMMF subgroup and each cancer and tissue cohort. These analyses are supposed to give an overview, if there are on average more BMMF reads per billion mapped reads found in the tumor data or in the healthy tissue samples respectively in the blood/non-tumor tissue data and if the observed differences are statistically significant.

3.4.1 Comparison of BMMF subgroup detection in PCAWG WGS tumor and non-tumor data

The comparison of the mean normalized read numbers found in the PCAWG WGS tumor and the PCAWG WGS normal tissue/blood data, shows that there are more subgroups per cohort colored in shades of blue than in shades of magenta (fig. 3.29). This indicates, that the

normalized reads in the PCAWG WGS tumor data exceed the BMMF signal found in the PCAWG WGS normal data set for most cohorts, however cancer types and subgroups with opposing trends are observed as well (fig. 3.29). The Canadian prostate cancer cohort is highlighted in different shades of blue across almost all subgroups, which shows that for most subgroups there was a higher BMMF signal detected in the tumor than in the normal tissue/blood data. BMMF1 subgroup 5 is shown to have the biggest positive difference between the tumor and the normal data, but the subgroups 1,6 and 21 as well as the BMMF1 unclear

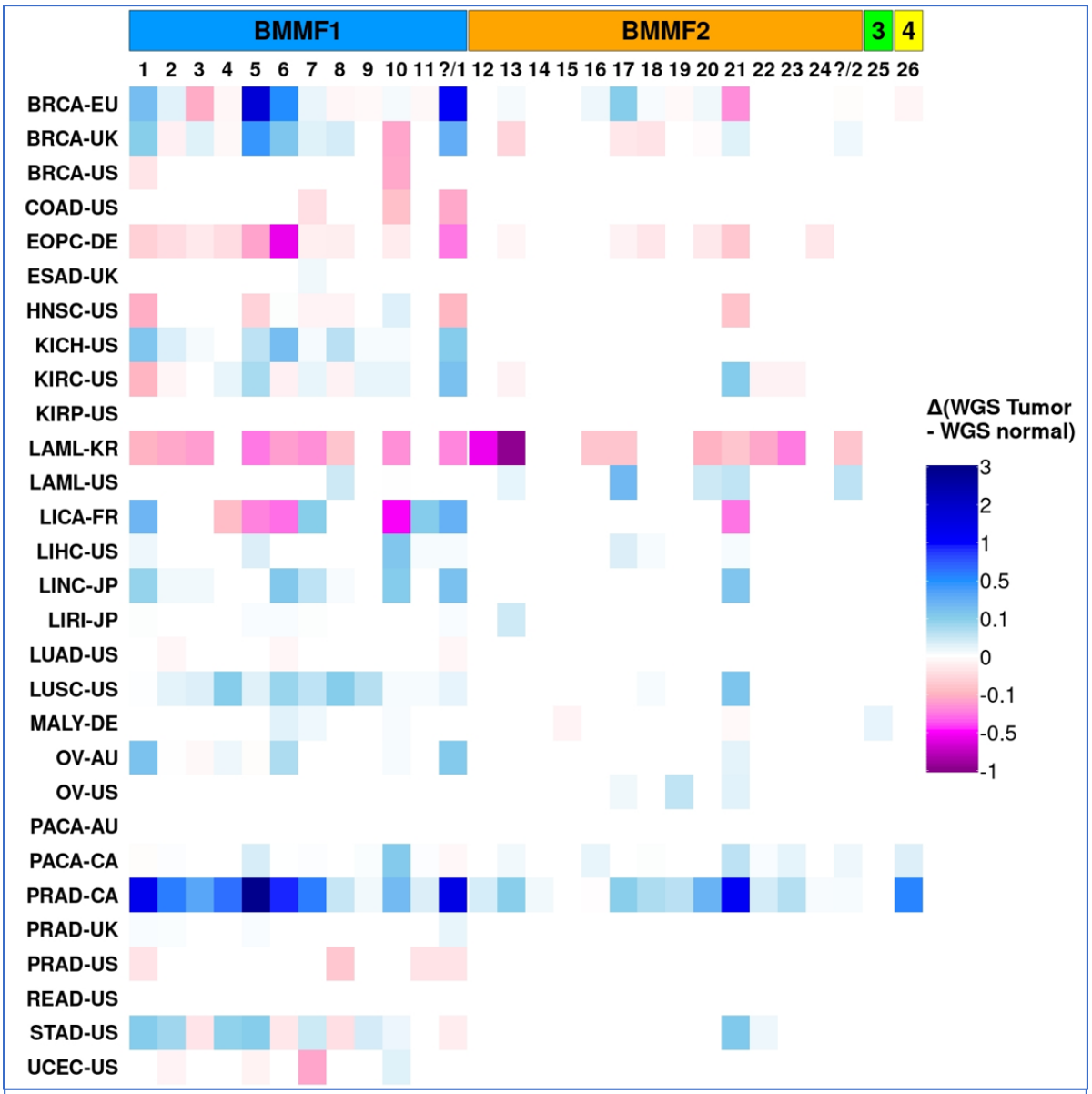


Fig. 3.29: Difference between normalized BMMF read numbers detected in PCAWG WGS tumor and normal tissue/blood data: The mean normalized BMMF reads per subgroup per cohort of the PCAWG WGS normal tissue/blood data set are subtracted from the mean normalized reads of the PCAWG WGS tumor data set. The differences between the normalized reads per subgroups per cohort of these two data sets are visualized in this heatmap. The color blue indicates higher mean normalized read numbers in the tumor data compared to the normal tissue/blood data. Magenta indicates a higher mean normalized BMMF signal in the normal tissue/blood data. White fields represent either no difference or no detection of this subgroup in the respective cohort in both tumor and normal data.

column also stand out with increased detection in the tumor data (fig. 3.29). Besides of the PRAD-CA cohorts, there are also several other cohorts with higher mean normalized read numbers detected in the tumor than in the normal tissue data such as the Canadian pancreatic cancer cohort, the US lung squamous cell carcinoma cohort, the Australian ovarian cancer cohort or the US kidney cancer cohort KICH-US (fig. 3.29). On the other hand, several cohorts such as the South Korean LAML cohort, the German early onset prostate cancer cohort and the US breast, colon and prostate cancer cohorts exhibit higher levels of BMMF signal in the normal tissue/blood data (fig. 3.29). Additionally, there are several cohorts in which some subgroups are found at higher frequencies in the tumor data, while other subgroups exhibit a higher BMMF signal in the normal tissue/blood data. This includes the French liver cancer cohort, the US stomach adenocarcinoma cohort and the European and UK breast cancer cohorts.

However, not all differences displayed in the difference matrix are actually statistically significant. For the statistical analysis I used the Zero-Inflated Rank Test provided by the ZIR library in R to account for the high number of samples with zero BMMF reads in the data sets analyzed using D-ViSioN. The visualization of the p-values determined by the ZIR-test shows, that the differences between the PCAWG WGS tumor and normal data is only significant for the detection of certain subgroups in specific cohorts such as the German early-onset prostate cancer cohort, the US kidney chromophobe cohort, the US lung squamous cell carcinoma cohort, the Canadian pancreatic cancer cohort and the Canadian prostate cancer cohort (fig. 3.30). While fields highlighted in green indicate a p-value lower than 0.05 for the comparison of the normalized tumor and normal tissue reads of the respective subgroups and cohort, this just indicates a significant difference between the cohorts compared. To identify if tumor or non-tumor samples show a statistically increased BMMF-signal, the fields with significant p-values need to be compared to the differences between the tumor and normal tissue/blood BMMF signal shown in figure 3.29.

In case of the German early-onset prostate cancer cohort, the tumor and normal data differ significantly regarding the detection of BMMF1 subgroups 1, 5 and 6 and of unassigned BMMF1 reads (fig. 3.30). The difference matrix shows a higher mean of normalized BMMF reads for all subgroups detected in the EOPC-DE cohort (fig. 3.29). Consequently, the BMMF signal in these subgroups in the EOPC-DE cohort is significantly higher in the normal data derived from blood samples than in the tumor data. However, all remaining cohorts and subgroups for which the p-value indicates a significant difference between the tumor and normal tissue/blood data, show a higher BMMF signal in the tumor than in the normal

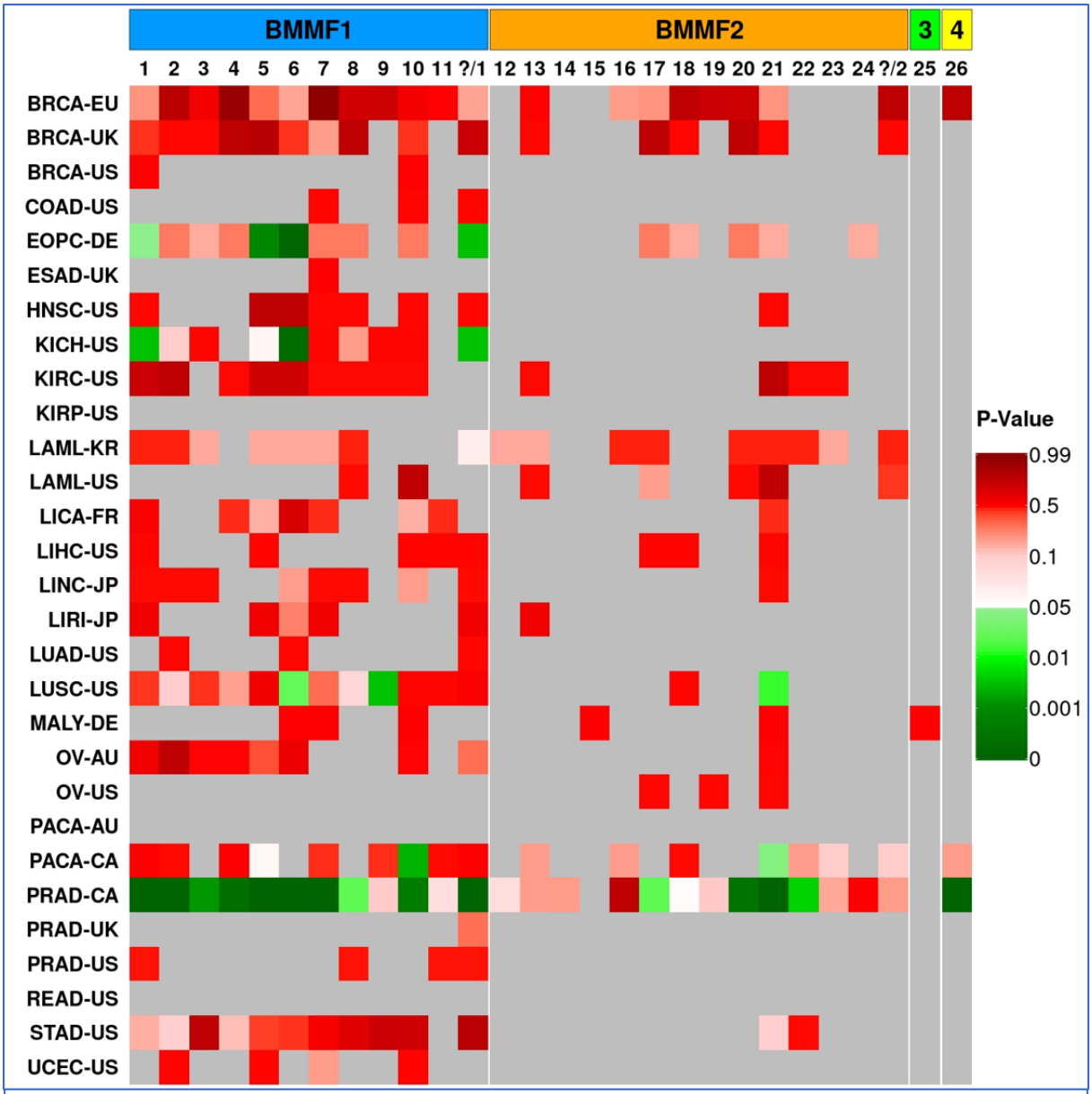


Fig. 3.30: P-values of statistical comparison of normalized BMMF reads detected in PCAWG WGS tumor and normal tissue/blood data: The normalized BMMF reads per patient of the PCAWG WGS tumor and the PCAWG WGS normal tissue/blood data set are compared using the Zero-Inflated Rank Test from the ZIR library in R. The reads normalized for sequencing depth of the patients of each cohort and each subgroup are compared between these two data sets and the p-value of the ZIR test is visualized in this heatmap. The color green indicates a significant difference regarding the detection of this subgroup in the respective cohort between the tumor and normal tissue/blood data. Red indicates no significant difference. Grey fields indicate no detection of this subgroup in the respective cohort in both tumor and normal data.

tissue/blood data (fig. 3.29/3.30). This includes BMMF1 subgroups 1, 6 and "BMMF1 unclear" for the KICH-US cohort as well as subgroup 6, 9 and 21 of the LUSC-US data. Additionally, the PACA-CA cohort shows a significantly increased BMMF signal in the tumor data for subgroups 10 and 21. The PRAD-CA cohort stands out most prominently in the PCAWG WGS data. Nine of eleven BMMF1 subgroups exhibit a significantly higher BMMF detection in the tumor data compared to the blood samples taken as control samples (fig. 3.30). Furthermore, four BMMF2 subgroups and BMMF group 4 shows a significantly increased BMMF signal.

3.4.2 Impact of origin of PCAWG WGS normal samples on comparison of BMMF subgroup detection in PCAWG WGS tumor and non-tumor data

In case of some PCAWG WGS cohorts the samples in the normal data set are derived from different origins. This includes the BRCA-EU and STAD-US cohorts. For this reason, I chose these two cohorts to calculate the differences and p-values seperately for the comparison of the tumor data with all normal samples as well as with only the normal samples derived from blood, tumor-adjacent tissue or solid tissue distant from the primary tumor. Comparing the tumor data with all normal samples, both cohorts exhibit higher BMMF detection in the tumor for some cohorts and higher BMMF detection in the normal samples in other cases (fig. 3.31A).

If the EU breast cancer samples are only compared with the blood samples taken from the cancer patients of this cohort, subgroups with a higher BMMF signal in tumor samples in the previous comparison still exhibit higher BMMF positivity in the tumor samples indicating less detection in blood. However, in case of subgroups 3 and 21 which showed higher BMMF detection in the normal data set, the difference between the respective cohorts becomes smaller (fig. 3.31A). When the tumor samples are compared to samples taken from tumor-adjacent tissue, the pattern changes even more. Only BMMF1 subgroups 5, 6 and unclear show a higher positivity in the tumor data. On the other hand, subgroups 1, 3, 4, 7, 8, 9, 10, 19, 21 and BMMF4 subgroup 26 all show higher levels of BMMF detection in the tumor-adjacent tissue. This indicates, that the origin of the normal data set samples is important when comparing the matching cohorts of the PCAWG WGS tumor and PCAWG WGS normal data sets (fig. 3.31A).

In case of the STAD-US cohort the differences between the different origins of the normal data are generally smaller. The comparison of the tumor samples with all normal samples results in an almost identical pattern in the difference matrix as the comparison with only the normal samples derived from blood samples. Taking into account only normal samples derived from solid tissue distant to the primary tumor changes the color of BMMF1 subgroups 1, 3, and 7. Whereas these subgroups are more frequently detected in tumor data than in the entire normal data or than in blood samples only, the normal tissue samples derived from solid tissue shows higher positivity for these subgroups than the STAD-US tumor samples (fig. 3.31A). The p-values calculated with the ZIR-test however indicate, that splitting the origins of the normal samples does not impact the significance of the differences observed between the tumor and the normal data (fig. 3.31B). There is one exception to this: The increased detection of BMMF1 subgroup 1 in the tumor data of the STAD-US cohort is not significant when comparing the tumor data with all normal samples. However, when the tumor data is compared only to normal

samples derived from blood, subgroup 1 shows a significantly increased detection in the tumor samples (fig. 3.31B).

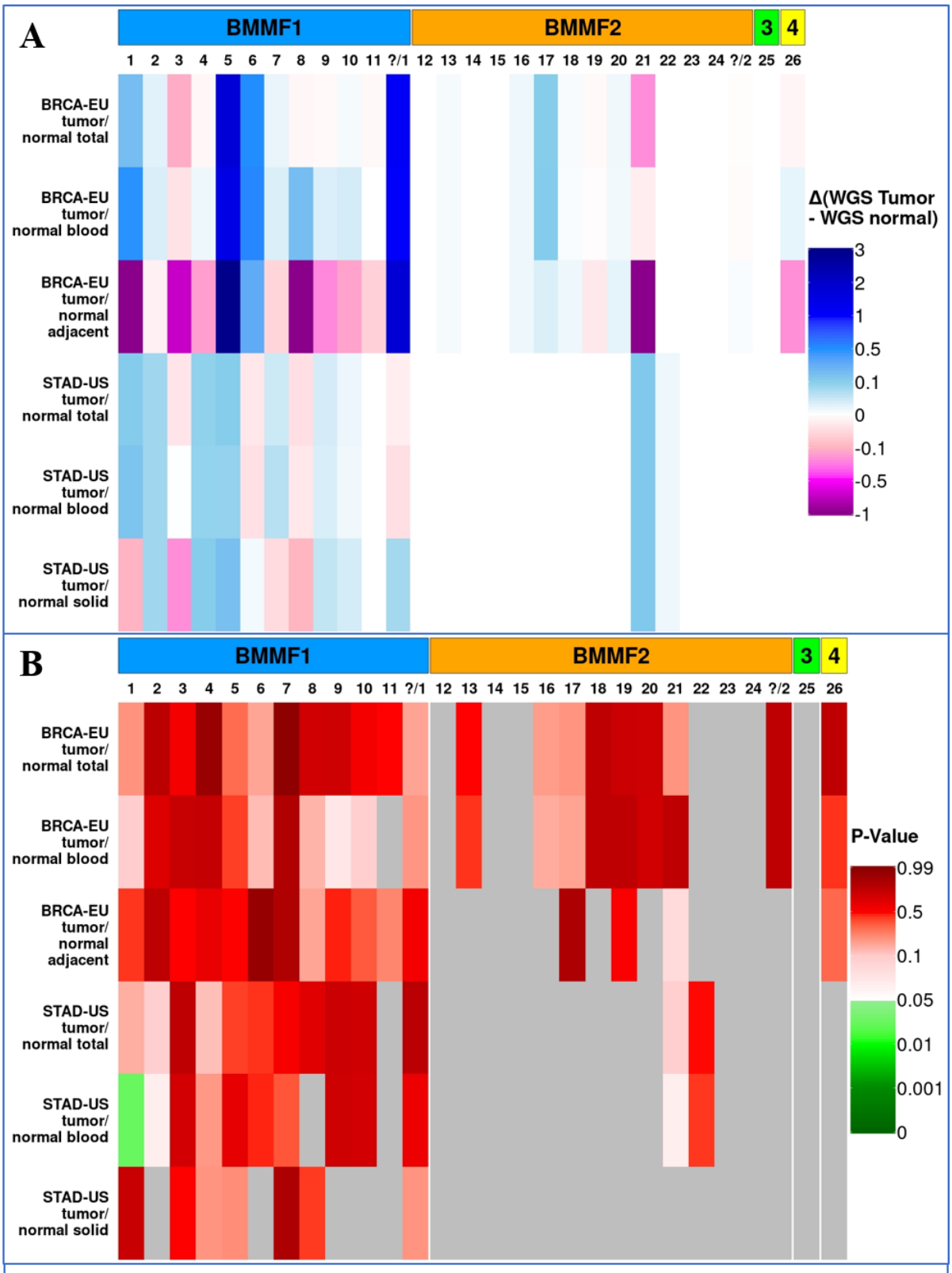


Fig. 3.31: Difference matrix (A) and P-values (B) of BRCA-EU and STAD-US PCAWG WGS data depending on origin of normal samples: The difference matrix (A) and p-values (B) determined using the ZIR-test for BRCA-EU and STAD-US samples split up in different origins of normal samples.

3.4.3 Comparison of subgroup detection in tumor and healthy tissue RNA data

Next, the TCGA and PCWAG RNA data sets are compared separately with the healthy tissue data of the GTEx data set. If there is more than one cancer cohort matching a normal tissue cohort of the GTEx data set, the GTEx cohort is compared seperately to both cancer cohorts. On the other hand, there are two different colon cohorts included in the GTEx RNA data set. These two healthy colon cohorts are also compared separately to the colon cancer data included in the TCGA and PCAWG RNA data sets.

The comparison of the differences regarding the mean BMMF reads per billion mapped reads found in the TCGA RNA data versus the GTEx RNA data shows, that for most subgroups the BMMF detection is lower in the cancer data than in the healthy tissue data (fig. 3.32). The clearest exception to this is the detection of BMMF1 subgroup 10 in the TCGA-COAD data, which is higher in the cancer data than in both the colon sigmoid and the colon transverse cohorts of the GTEx RNA data set. However, these observed differences turn out to be statistically insignificant looking at the p-values calculated using the ZIR-test (fig. 3.33).

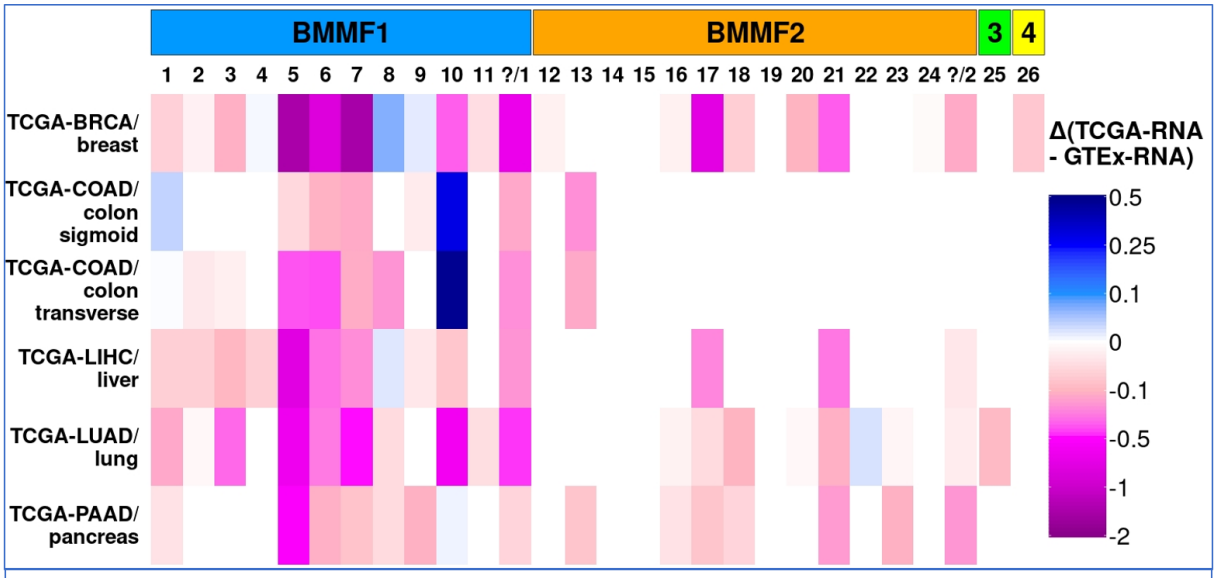


Fig. 3.32: Difference between normalized BMMF read numbers detected in TCGA RNA and GTEx RNA data: The mean normalized BMMF reads per subgroup per cohort of the GTEx RNA tissue data are subtracted from the mean normalized reads of the TCGA RNA data set. The differences between the normalized reads per subgroups per cohort of these two data sets are visualized in this heatmap. The color blue indicates higher mean normalized read numbers in the tumor data compared to the healthy tissue data. Magenta indicates a higher mean normalized BMMF signal in the healthy tissue data. White fields represent either no difference or no detection of this subgroup in the respective cohort in both tumor and normal data.

While a number of subgroups shows significant different levels of BMMF detection between the tumor and the normal tissue data, the BMMF signal is higher in the normal tissue data for all of these cases. The pancreatic cancer cohort is the only TCGA cohort for which no subgroup shows significantly higher levels of BMMF detection in the respective GTEx cohorts (fig. 3.33). Comparing the TCGA breast and lung cancer data sets, there are each five BMMF1 subgroups

with significantly higher signal in the GTEx tissue data. Additionally, there are significantly more normalized BMMF reads reported for three BMMF2 subgroups comparing the TCGA-BRCA and the GTEx breast tissue data (fig 3.33). The comparison between the TCGA colorectal and liver cancer data with the respective GTEx cohorts, only shows between one and three subgroups with significantly higher BMMF detection in the normal tissue data than in the cancer data (fig. 3.33).

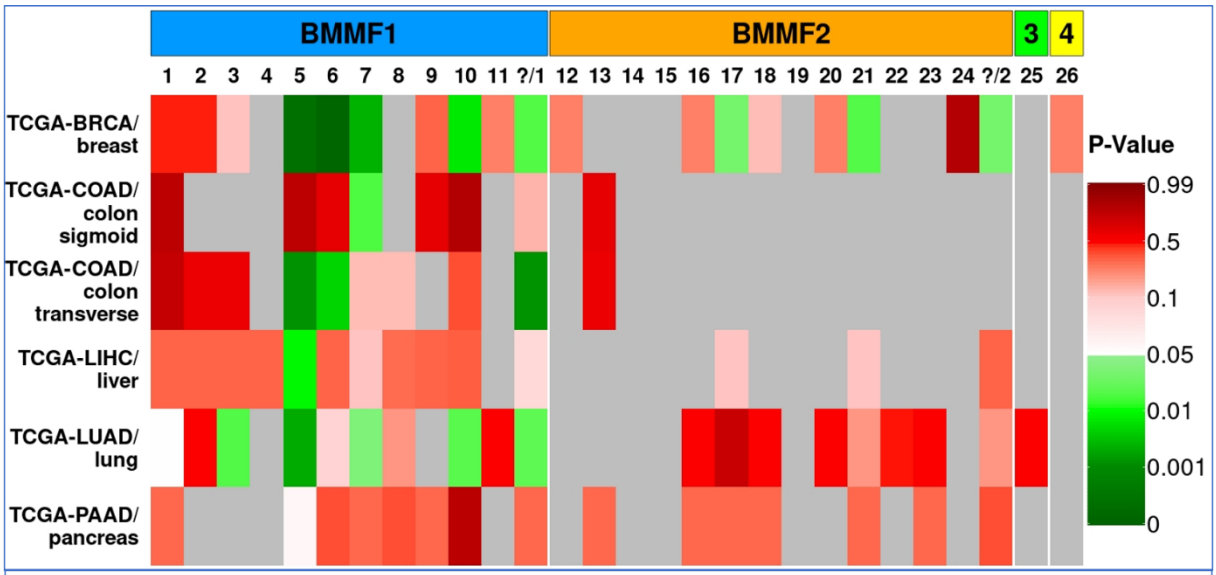


Fig. 3.33: P-values of statistical comparison of normalized BMMF reads detected in TCGA RNA and GTEx RNA data: The normalized BMMF reads per patient of the TCGA RNA data and the respective tissue cohorts of the GTEx data set are compared using the Zero-Inflated Rank Test from the ZIR library in R. The reads normalized for sequencing depth of the patients of each cohort and each subgroup are compared between these two data sets and the p-value of the ZIR test is visualized in this heatmap. The color green indicates a significant difference regarding the detection of this subgroup in the respective cohort between the tumor and normal tissue/blood data. Red indicates no significant difference. Grey fields indicate no detection of this subgroup in the respective cohort in both tumor and normal data.

When focusing on the comparison of the PCAWG RNA versus the GTEx RNA results, the difference matrix of the mean BMMF reads per billion mapped reads shows a mixed pattern. Since the PCAWG RNA CLLE-US, DLBC-US, LIHC-US, LUAD-US and PRAD-US cohorts were BMMF negative after application of the three-hits threshold, the level of BMMF reads detected is obviously higher in the respective tissue GTEx cohorts (fig. 3.34). The two kidney cancer cohorts with BMMF hits detected – KIRC-US and KIRP-US – on the other hand showed a higher BMMF signal compared with the kidney GTEx cohort, which was BMMF negative after application of the three-hits threshold (fig. 3.34).

The remaining cohorts have both BMMF positive samples reported in the PCAWG RNA data and in the respective normal tissue data. There are both subgroups with a higher signal in the tumor data and subgroups with a higher BMMF signal in the healthy tissue data reported for all of these cohorts. In case of the breast cancer, colorectal cancer and lung squamous cell carcinoma cohorts, most subgroups are more frequently detected in the healthy tissue data than

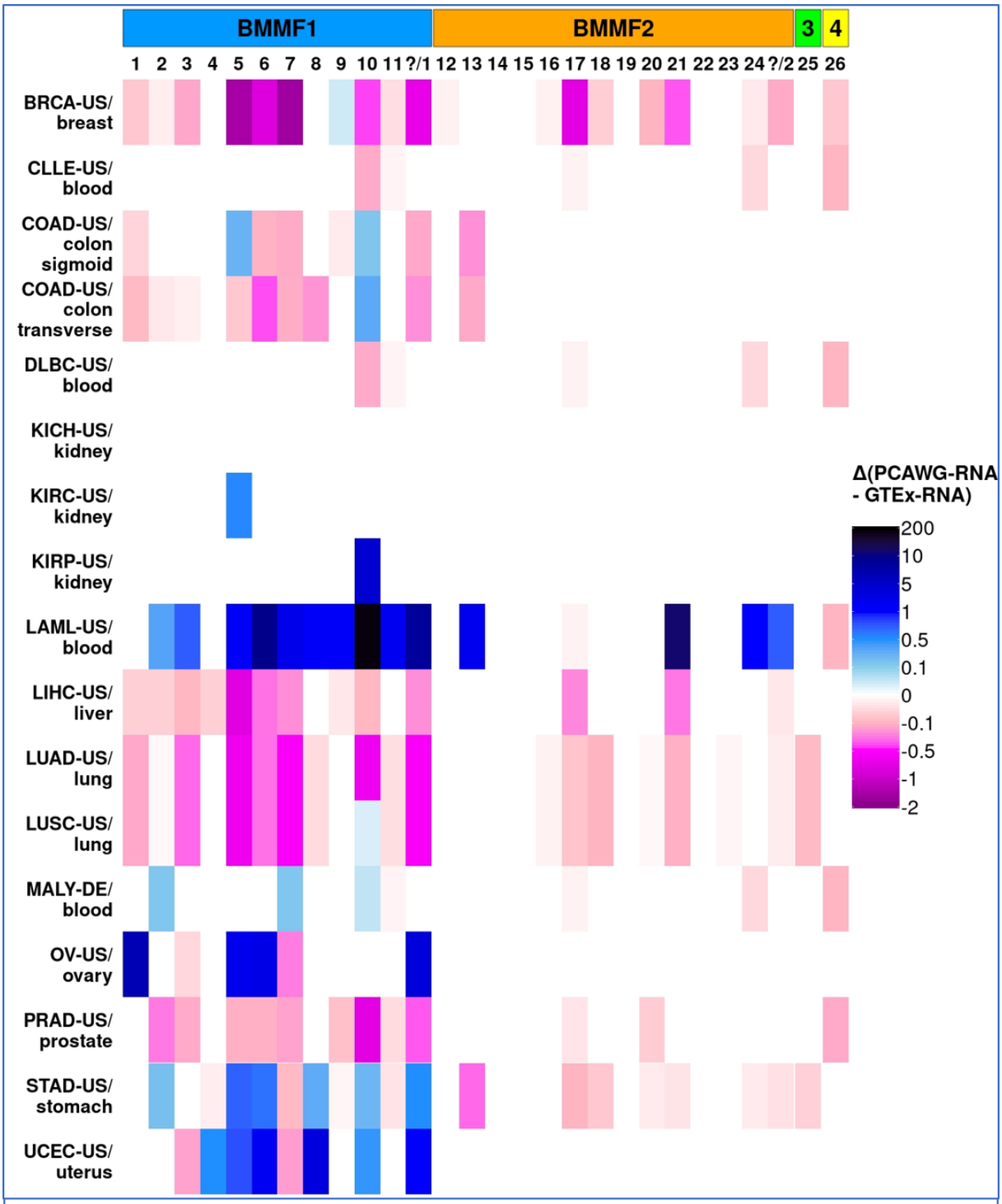


Fig. 3.34: Difference between normalized BMMF read numbers detected in PCAWG RNA and GTEx RNA data: The mean normalized BMMF reads per subgroup per cohort of the GTEx RNA tissue data are subtracted from the mean normalized reads of the PCAWG RNA data set. The differences between the normalized reads per subgroups per cohort of these two data sets are visualized in this heatmap. The color blue indicates higher mean normalized read numbers in the tumor data compared to the healthy tissue data. Magenta indicates a higher mean normalized BMMF signal in the healthy tissue data. White fields represent either no difference or no detection of this subgroup in the respective cohort in both tumor and normal data.

in the tumor data. On the other hand, there is a higher BMMF signal reported in the cancer data of the LAML-US, OV-US and UCEC-US cohorts compared to the healthy tissue controls. The

stomach cancer data contains a higher level of BMMF1 hits, whereas there were more BMMF2 hits detected in the healthy stomach data (fig. 3.34).

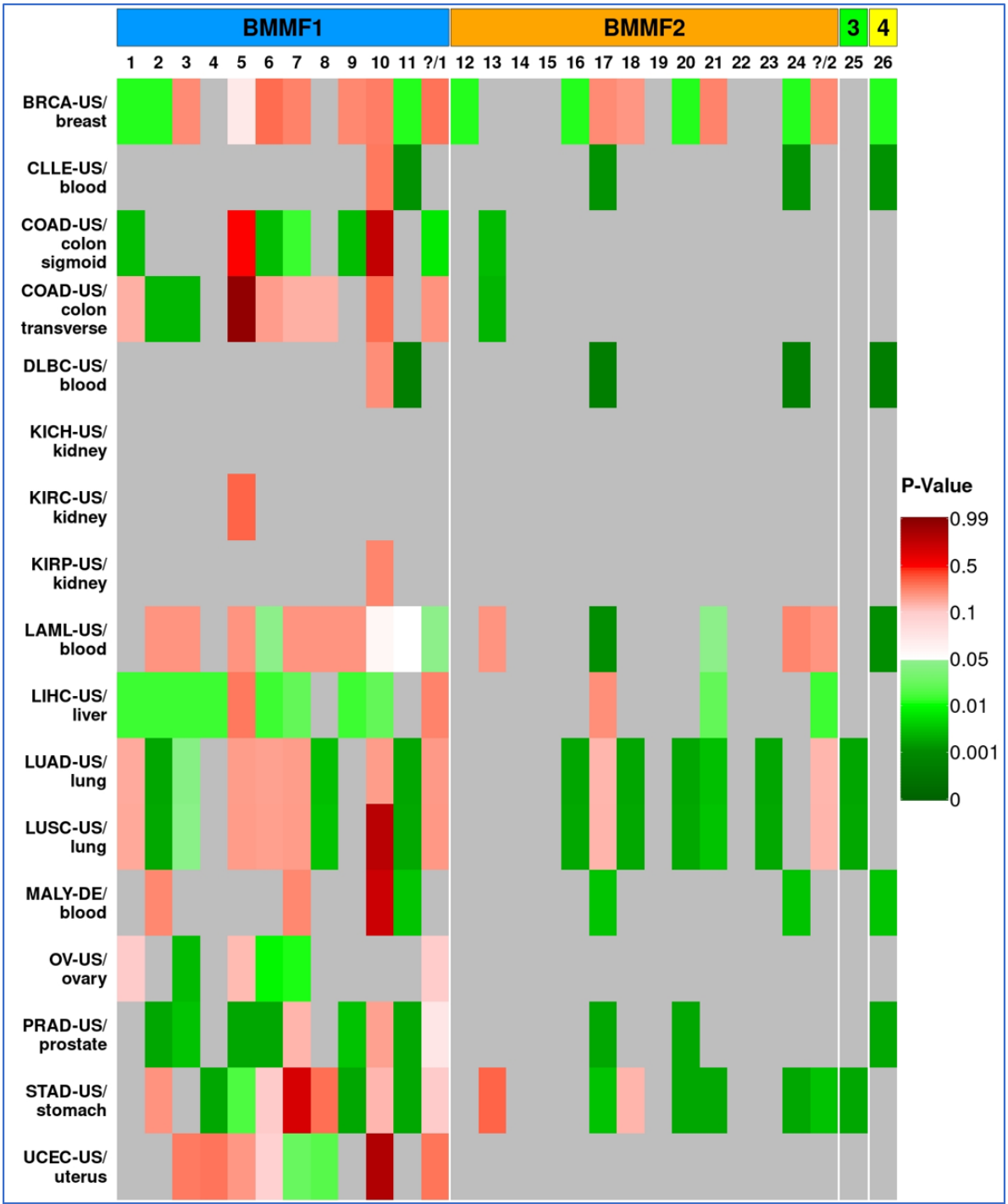


Fig. 3.35: P-values of statistical comparison of normalized BMMF reads detected in PCAWG RNA and GTEx RNA data: The normalized BMMF reads per patient of the PCAWG RNA data and the respective tissue cohorts of the GTEx data set are compared using the Zero-Inflated Rank Test from the ZIR library in R. The reads normalized for sequencing depth of the patients of each cohort and each subgroup are compared between these two data sets and the p-value of the ZIR test is visualized in this heatmap. The color green indicates a significant difference regarding the detection of this subgroup in the respective cohort between the tumor and normal tissue/blood data. Red indicates no significant difference. Grey fields indicate no detection of this subgroup in the respective cohort in both tumor and normal data.

Looking at the p-values calculated for the respective comparisons, significant differences are reported for all cohorts for which BMMF positive samples were reported both in the tumor and in the normal tissue data (fig. 3.35). Due to the much higher sample sizes of GTEx cohorts, the subgroups with higher BMMF signal reported in the healthy tissue data are more frequently classified as significantly increased, than the subgroups with higher numbers of mean normalized reads reported for the tumor data. The LAML-US cohort exemplifies this: While the mean read numbers per billion mapped reads are much higher in the LAML data than in the GTEx blood data for many different subgroups, only for subgroups have a p-value lower than 5 % (fig. 3.35). This includes on the one hand subgroups 6 and 21, which exhibited higher BMMF levels in the tumor data, but on the other hand also subgroup 17 and BMMF group 4, for which higher levels were detected in the GTEx blood data. While the LAML-US data contains only 20 samples, the GTEx blood cohort comprises 755 samples.

Besides of the LAML-US cohort, a significantly higher signal in the tumor data is only reported for BMMF1 subgroup 6 in the ovarian cancer data, for subgroup 5 in the stomach cancer data and for subgroup 8 in the uterine cancer cohort (fig. 3.35). The remaining subgroups reported to be significantly different between the two data sets show all higher levels of BMMF detection in the healthy tissue cohorts of the GTEx projects than in the cancer data of the PCAWG RNA data set.

Consequently, cancer RNA sequencing data sets contain – with few exceptions – lower levels of BMMF reads per billion mapped reads than the control data provided by the GTEx project. In case of the comparison of the PCAWG WGS tumor and normal data, there are no significant differences between most of the cohorts analyzed. However, in case of four of the five cohorts with significant differences reported, the tumor data showed higher levels of BMMF detection than the normal tissue/blood data. Only the German early-onset prostate cancer data exhibited higher BMMF positivity in the blood data of the PCAWG WGS normal data set than in the tumor data.

3.5 Quality assessment of detected BMMF reads

3.5.1 Read coverage of detected BMMF genomes

The BMMF reads detected so far have been characterized regarding the detecting of the four main groups and 26 subgroups in different tissues and cancer cohorts. However, it has not been discussed yet, which specific BMMF genomes are found. Additionally, it has not been analyzed in detail so far which regions of the detected genomes are covered by reads. I showed in figure

3.8 (chapter 3.2) the circular map of BMMF group 3 isolate. While there are high numbers of reads reported for this genome in all data sets analyzed so far, the detected reads always just cover the same very short section on the genome. Read fragments with a length of about 30 bp are mapped to this region of HCBI8.215 (fig. 3.8). This is a strong indication that the detected reads are caused by artifacts instead of the actual presence of the HCBI8.215 genome. In this chapter I will analyze the detection and read coverage of several BMMF genomes part of the most frequently detected BMMF subgroups described in chapter 3.4. On the one hand I will investigate if the reported reads are reliably caused by the actual presence of BMMFs and on the other hand I will examine if there are BMMF regions more frequently detected and if there is a correlation between read coverage and open reading frames of the BMMF genomes. The open reading frames are predicted using the Orffinder webtool. The predicted ORFs longer than 150 nucleotides are plotted into the circular graphics visualized the read covered to compare the location of the detected reads to the location of the predicted ORFs. Due to the high similarity between closely related BMMF genomes reads cannot always be clearly assigned to one BMMF genome or one BMMF subgroup, since they often map to a number of different genomes with a high sequence identity. Consequently, I differentiate between so-called “clear” und “unclear” hits. Clear hits can be unambiguously assigned to one subgroup and are thus reads specific for the genomes part of that subgroup. Unclear reads can be mapped equally well to the respective BMMF genome as well as to BMMF genomes of other subgroups, which is why it cannot be safely decided which specific BMMF genome was actually present in the sample and causes these reads.

BMMF1 subgroup 5 isolate C1MI.3M.1 was the most frequently detected BMMF genome in both the PCAWG WGS tumor and the PCAWG WGS normal tissue/blood data sets. C1MI.3M.1 was found in 95 samples in the PCAWG WGS primary tumor data set and in 46 samples in the PCAWG WGS normal data set. In case of 19 patients C1MI.3M.1 was detected both in their primary tumor sample and in their normal tissue/blood sample. Besides of the PCAWG WGS data sets, C1MI.3M.1 was also the second most frequently detected genome in the GTEx RNA data. While C1MI.3M.1 was also detected as well in the PCAWG RNA and TCGA RNA data it is not among the top hits reported for these data sets. The majority of the reads clearly assigned to C1MI.3M.1 are mapped to three different regions: the longest peak region ranges from 50 to 1050 bp, the second longest peak region ranges from 1250 to 1500 bp and the final peak region ranges from approximately 1625 to 1725 bp (fig. 3.36). These three peak regions can be observed in the PCAWG WGS tumor and normal data sets as well as in the GTEx RNA data (fig. 3.36). Most reads detected in the PCAWG RNA data also fall within

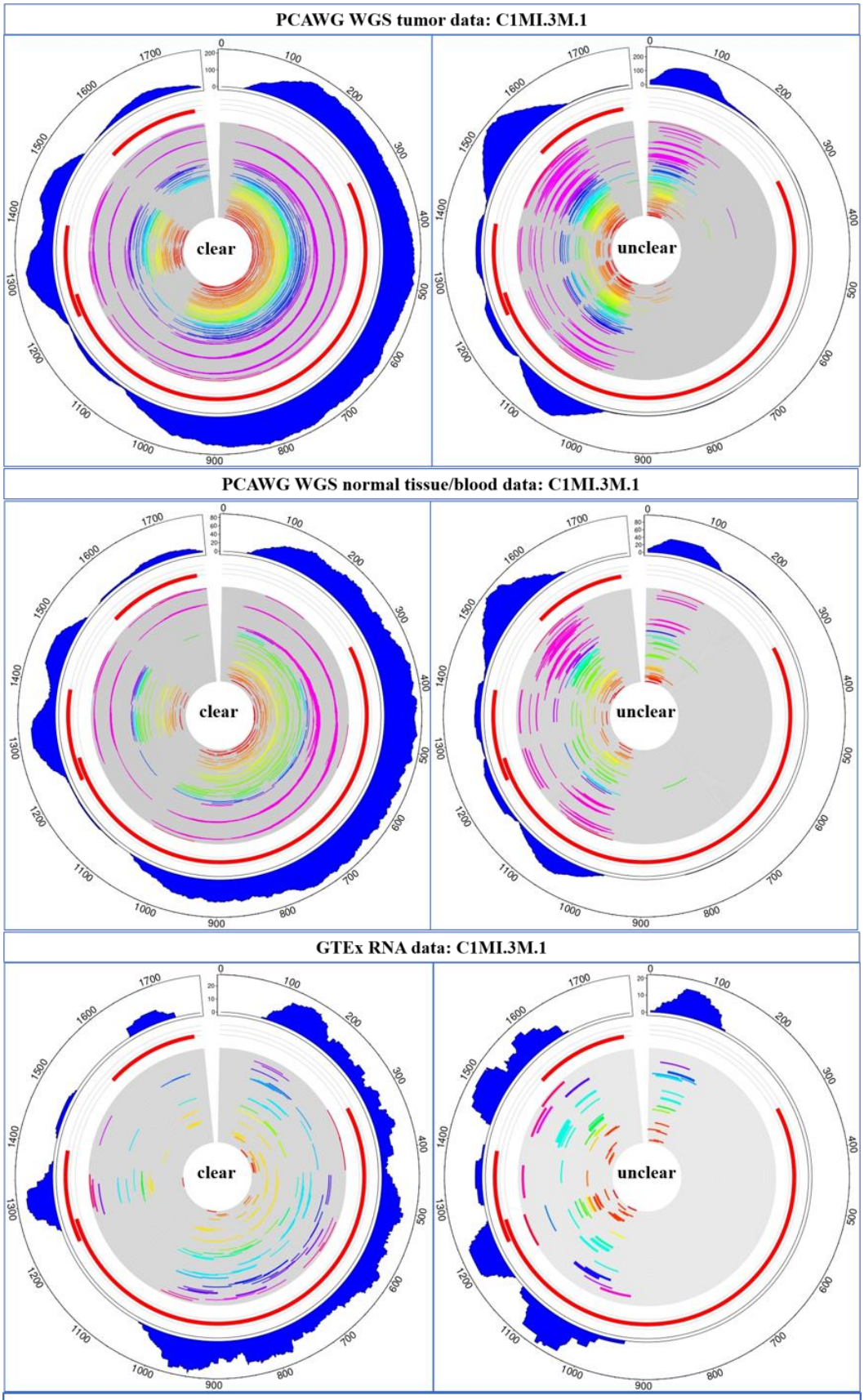


Fig. 3.36: Reads detected for BMMF group 1 isolate C1MI.3M.1 in PCAWG WGS and GTEx RNA data sets: Coverage of BMMF1 subgroup 5 genome C1MI.3M.1 by reads clearly assigned to subgroup 5 (left) and unclear reads (right). The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

these regions, but the read coverage within the PCAWG RNA data set is too low to constitute clearly visible peak regions (S17). The TCGA RNA data set only contained two C1MI.3M.1 reads. The highest number of C1MI.3M.1 reads was found in the PCAWG WGS tumor data set, where the first peak region is covered by up to 200 reads per position. The second peak is covered by up to 150 reads per position and the third short peak is covered by less than 100 reads per position in the PCAWG WGS tumor data. Peak 1 is also the most frequently detected region in the PCAWG WGS normal and GTEx RNA data sets, where this peak is covered by up to 80 respectively 20 hits per posititon (fig. 3.36). The regions outside of the peak regions are covered by lower read numbers. The first 40 bp of the genome as well as well as a short section at 1550 bp are even not covered by any clearly assigned reads at all in all three PCAWG data sets and the GTEx RNA data. Taking into account not only clearly assigned reads, but also unclear reads, it can be observed for all of these four data sets, that the clear and unclear reads fit together almost like puzzle pieces (fig. 3.36). The unclear reads show peaks at the first 100 bp of C1MI.3M.1 as well as between 1000 and 1100 bp and from 1450 to 1650 bp, which are regions covered by little to no clearly assigned C1MI.3M.1 reads (fig. 3.36/S17). Since the clear and unclear reads complement each other and show similar numbers of reads per position for all four data sets with substantial numbers of C1MI.3M.1 reads detected, it can be assumed, that the unclear reads can be really attributed to the detection of the C1MI.3M.1 genome in spite of their high sequence identity to BMMF1 genomes of other subgroups (fig. 3.36/S17). Putting the clear and unclear reads together, C1MI.3M.1 is quite evenly covered by reads without visible bias to the predicted open reading frames (fig. 3.36/S17).

When analyzing the C1MI.3M.1 coverage reported in the PCAWG WGS normal data, I was also interested in potential differences between the different origins of the samples in this data set. Most samples are derived from either blood, solid tissue distant to the tumor or solid tissue adjacent to the tumor. Consequently, I split the clearly assigned C1MI.3M.1 reads detected in the PCAWG WGS normal data into these three sample origins (fig. 3.37). The read coverage patterns observed in the total normal data set and in the blood samples are nearly identical, since more than 70 % of the total samples are blood-derived (fig. 3.37A+B). The C1MI.3M.1 reads detected in solid tissue samples taken from close or distant locations to the primary tumor all fall within the three peak regions observed for C1MI.3M.1 (fig. 3.37C+D). However, the peak regions are not as clearly visible as in the blood data due to the number of reads found in the samples from these origins (fig. 3.37).

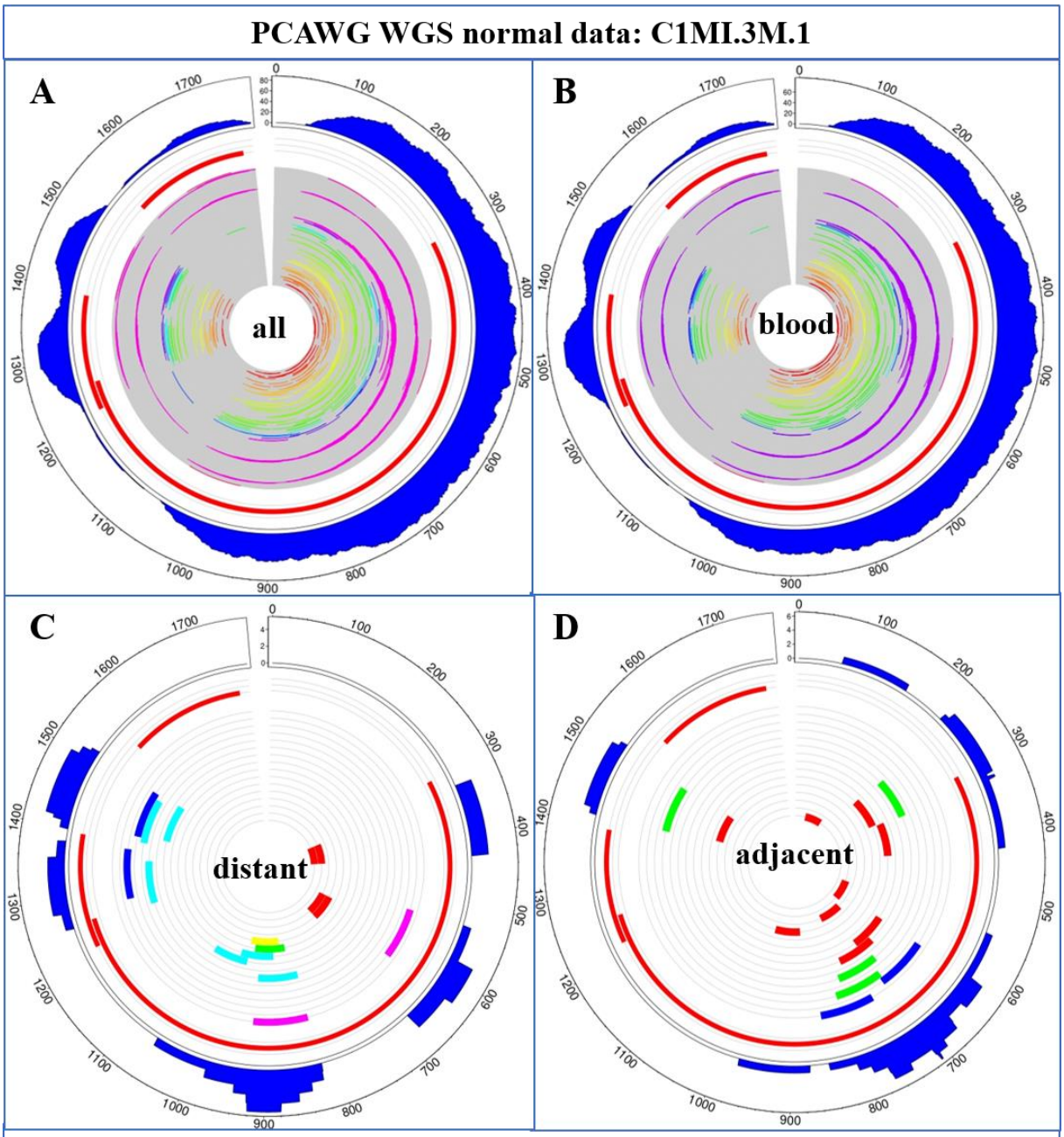


Fig. 3.37: Reads detected for BMMF group 1 isolate C1MI.3M.1 in PCAWG WGS normal data samples of different origins: Coverage of BMMF1 subgroup 5 genome C1MI.3M.1 by reads clearly assigned to subgroup 5 in all normal samples (A), blood samples (B), solid tissue samples distant to the tumor (C) and solid tissue samples adjacent to the tumor (D). The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

Sphinx2.36 is the most frequently detected BMMF2 genome in both the PCAWG WGS tumor and the PCAWG WGS normal tissue data (fig. 3.38). Since the number of unclear BMMF2 reads is much lower compared to the uncleared BMMF one reads, there were little to no unclear Sphinx2.36 reads reported. For this reason, only Sphinx2.36 reads clearly assigned to BMMF2 subgroup 21 are shown in figure 3.38. The Sphinx2.36 reads detected in both PCAWG WGS

data sets are covering the entire genome, however the number of reads mapping per position on the genome are much higher in the PCAWG WGS data set (fig. 3.38). In total there were 37 PCAWG WGS tumor samples with Sphinx2.36 reads detected, whereas only 11 samples of the PCAWG WGS normal data set contained Sphinx2.36 reads. One single patient was Sphinx2.36 positive in both the tumor and the normal tissue sample. The GTEx RNA data set contains lower numbers of Sphinx2.36 reads, which mainly cover the region ranging from 400 to 800 bp (S18). There were no Sphinx2.36 reads discovered in the PCAWG and TCGA RNA data sets. While there were very little BMMF2 reads reported for the TCGA RNA data set, there were higher numbers of BMMF2 reads found in the LAML-US cohort of the PCAWG RNA data set. The majority of these reads are found in one single LAML patient (S19). While the reads detected in this patient cover a large fraction of the BMMF2 subgroup 21 genome C2MI.5A.3, there is also a very short and suspicious peak region, that exceeds the read coverage of the remaining genome by far (S19).

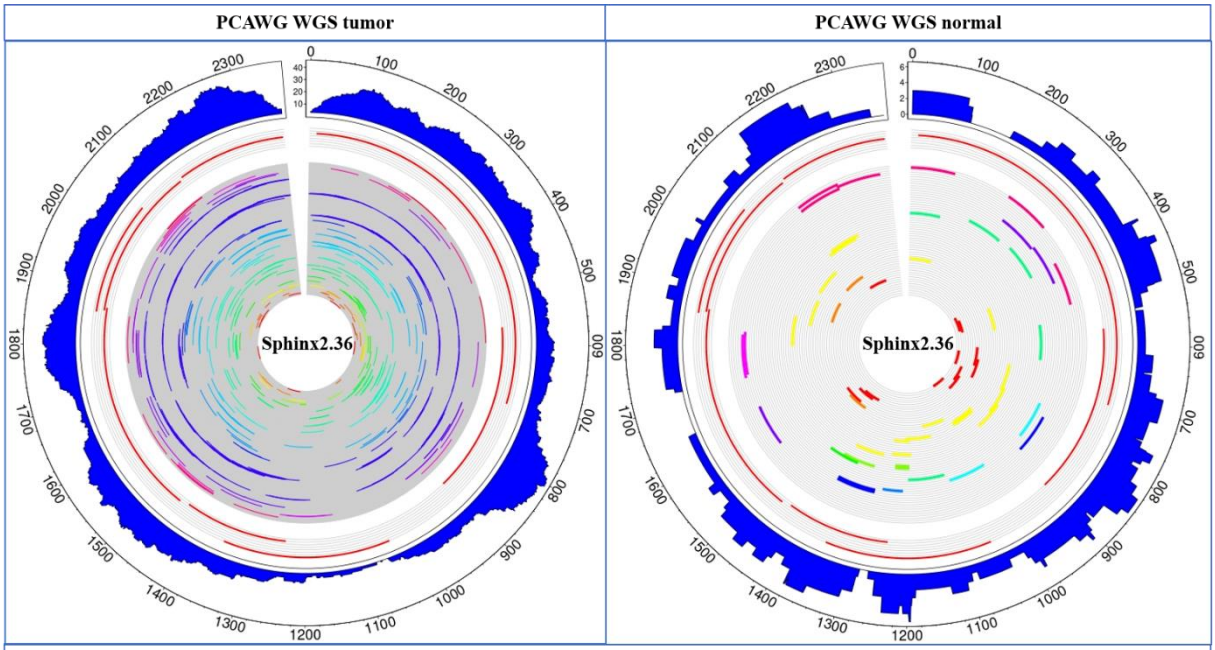


Fig. 3.38: Reads detected for BMMF group 2 isolate Sphinx2.36 in PCAWG WGS tumor and normal data: Coverage of BMMF2 subgroup 21 genome Sphinx2.36 by reads clearly assigned to subgroup 21. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

After the removal of the hits covering BMMF3 genome HCBI8.215 to exclude artifacts, BMMF group 3 was only sporadically detected in across all data sets. BMMF group 4 was detected most frequently in the PCAWG WGS tumor data with 23 MSSI1.162 positive patients. However, the detected reads are not evenly distributed across the only BMMF4 genome MSSI1.162. The majority of the reads either map to the beginning of the large predicted open

reading frame (100-400 bp) or to the region at the end of the sequence, which is not covered by any predicted open reading frame (1350-1600 bp) (fig. 3.39). In the PCAWG WGS normal tissue/blood data, MSSI1.162 was only found in two patients. The read distribution resembles the pattern observed in the PCAWG WGS tumor data, but the total BMMF4 read numbers are much lower (fig. 3.39). The GTEx RNA data also contained a few BMMF4 positive patients, whereas MSSI1.162 was not detected in the two cancer RNA sequencing data sets.

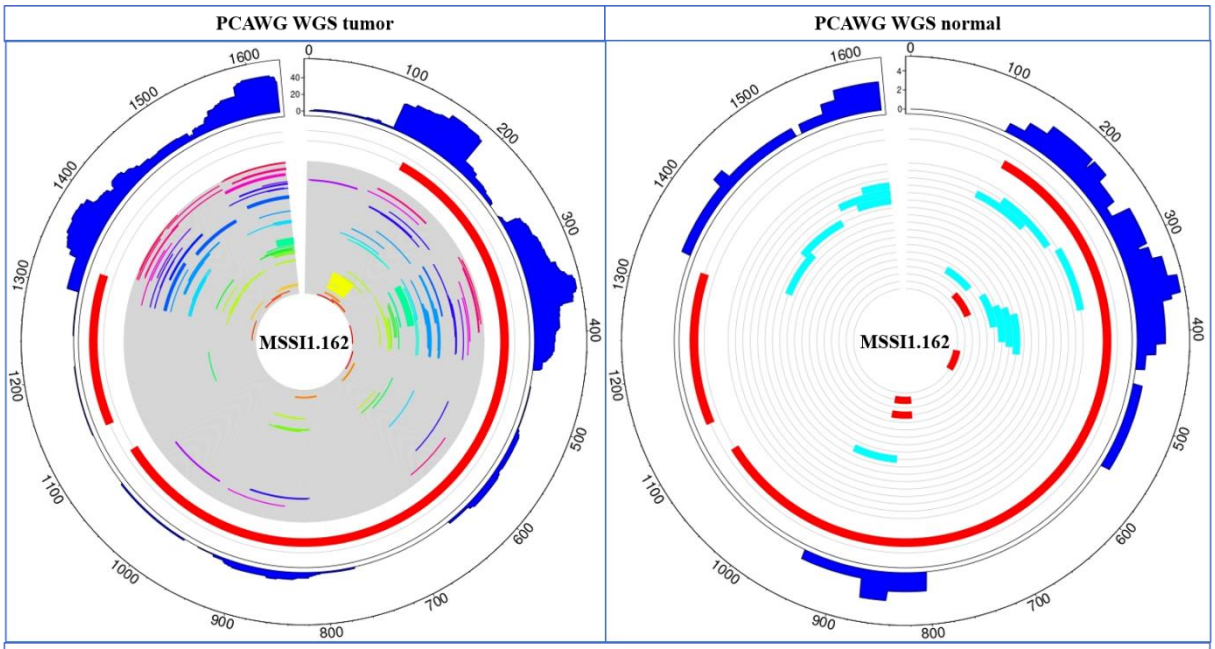


Fig. 3.39: Reads detected for BMMF group 4 isolate MSSI1.162 in PCAWG WGS tumor and normal data: Coverage of BMMF group 4 genome MSSI1.162. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

BMMF1 group 10 genome C1HB.4 was among the three most frequently detected BMMF isolates in all three RNA data sets, but it was also frequently detected in both PCAWG WGS data sets (fig. 3.40). Consequently, it is the most consistently detected BMMF genome across all five data sets analyzed using D-ViSioN. Figure 3.40 shows the reads clearly assigned to C1HB.4 for all five data sets. Since C1HB.4 is the only genome is subgroup 10 and an outlier in BMMF group 1, there are only very few ambiguous reads mapping to C1HB.4. The C1HB.4 reads detected in all five data sets are distributed across the genome, however there is a peak region between 100 and 200 bp, that shows up in all data sets except from the TCGA RNA data. Interestingly, this peak region is outside from any predicted open reading frame (fig 3.40). Additionally, there is an accumulation of reads in some of the data sets around position 1300, which is at the end of the first large open reading frame.

The PCAWG RNA data set contains the highest number of C1HB.4 reads. A high number of these reads is derived from one single LAML-US patient (fig. 3.40). This patient shows a strong

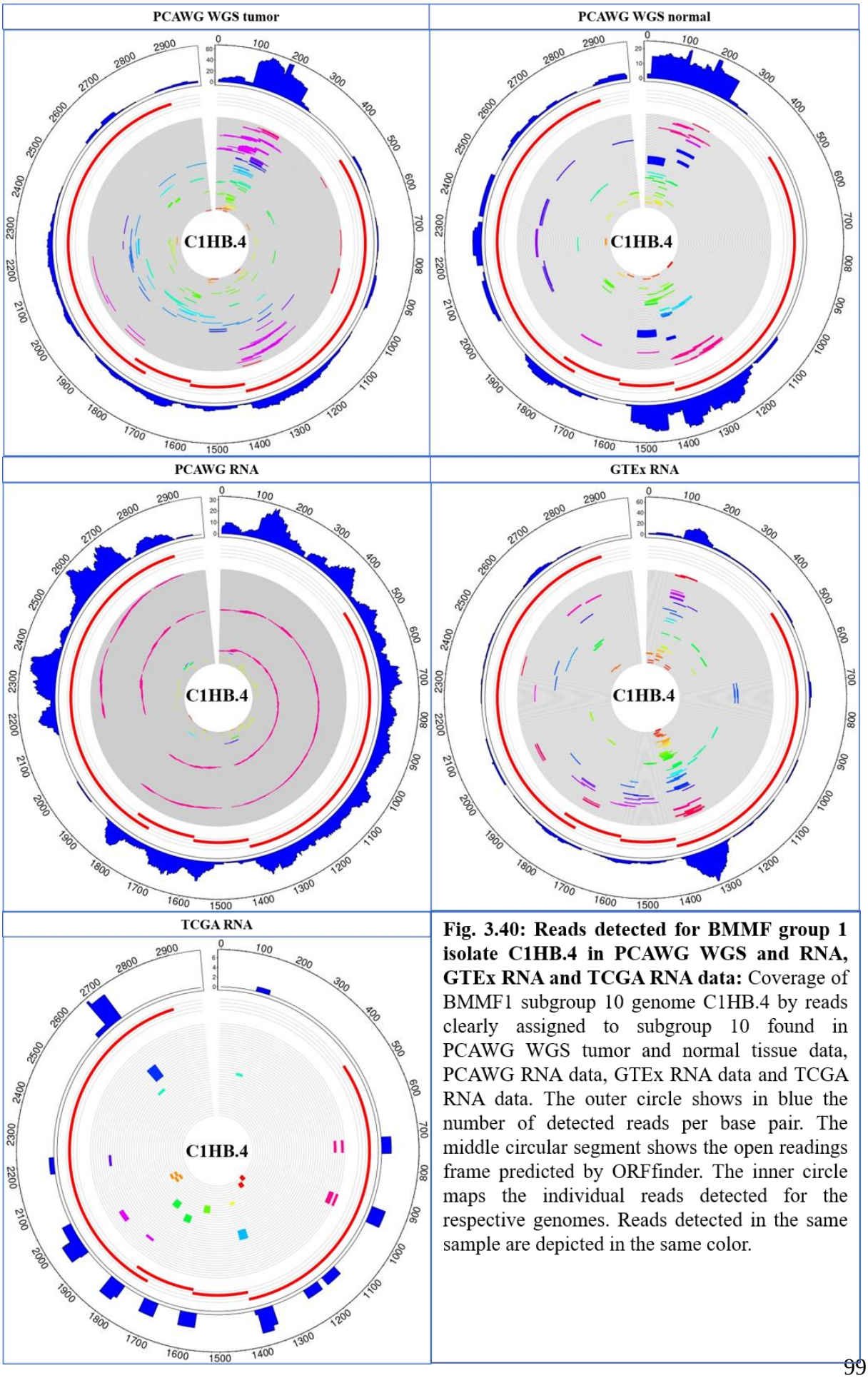


Fig. 3.40: Reads detected for BMMF group 1 isolate C1HB.4 in PCAWG WGS and RNA, GTEx RNA and TCGA RNA data: Coverage of BMMF1 subgroup 10 genome C1HB.4 by reads clearly assigned to subgroup 10 found in PCAWG WGS tumor and normal tissue data, PCAWG RNA data, GTEx RNA data and TCGA RNA data. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

read coverage of almost the entire sequence with one striking gap from 1900 to 2100 bp at the beginning of the second large predicted open reading frame. There is no comparable gap visible in the other data sets (fig. 3.40). This region is covered by reads derived from several different patients of the PCAWG WGS tumor data set, but there were also reads found for this region in the remaining three data sets. While the PCAWG WGS tumor data set contained 60 samples with C1HB.4 reads detected, only 21 samples in the PCAWG WGS normal data set exhibited C1HB.4. There were only few overlaps between the donors of the C1HB.4 positive samples in the tumor and normal data. 4 patients contained C1HB.4 reads in both of their samples.

The BMMF1 subgroup 7 template C1MI.2 is also detected in all five data albeit at different frequencies and with different coverage patterns. The C1MI.2 reads detected in both PCAWG WGS data sets mostly accumulate around nucleotide 1600, only few reads cover regions besides of this peak. Both the PCAWG WGS tumor and the PCAWG WGS normal data set contain one patient with a very large number of reads mapping at the peak region, but with no C1MI.2 reads outside of this region detected (fig. 3.41). Interestingly, this peak is caused by reads from the same Canadian prostate cancer patient, who contains reads of this peak region in both his tumor sample and in his blood sample. While there were 48 PCAWG WGS tumor samples with C1MI.2 reads detected, only 22 PCAWG WGS normal samples contained C1MI.2. 6 patients exhibit C1MI.2 positivity in both their normal tissue/blood samples.

The RNA data sets show a different picture compared to the PCAWG WGS data. Whereas there were only few C1MI.2 reads detected in the PCAWG RNA data set, C1MI.2 was the most frequently detected BMMF genome in the TCGA RNA and GTEx RNA data sets. Both of these data sets show a very similar pattern of the C1MI.2 reads detected. Whereas the reads cover most of the sequence, there seems to be a bias towards the second half of the sequence, which is covered by higher read numbers in both data sets (fig. 3.41). However, this pattern does not seem to be linked to the predicted open reading frames of C1MI.2. While a large fraction of the TCGA C1MI.2 reads is derived from the same patient, the GTEx data set contained several patients with high numbers of reads detected (fig. 3.41).

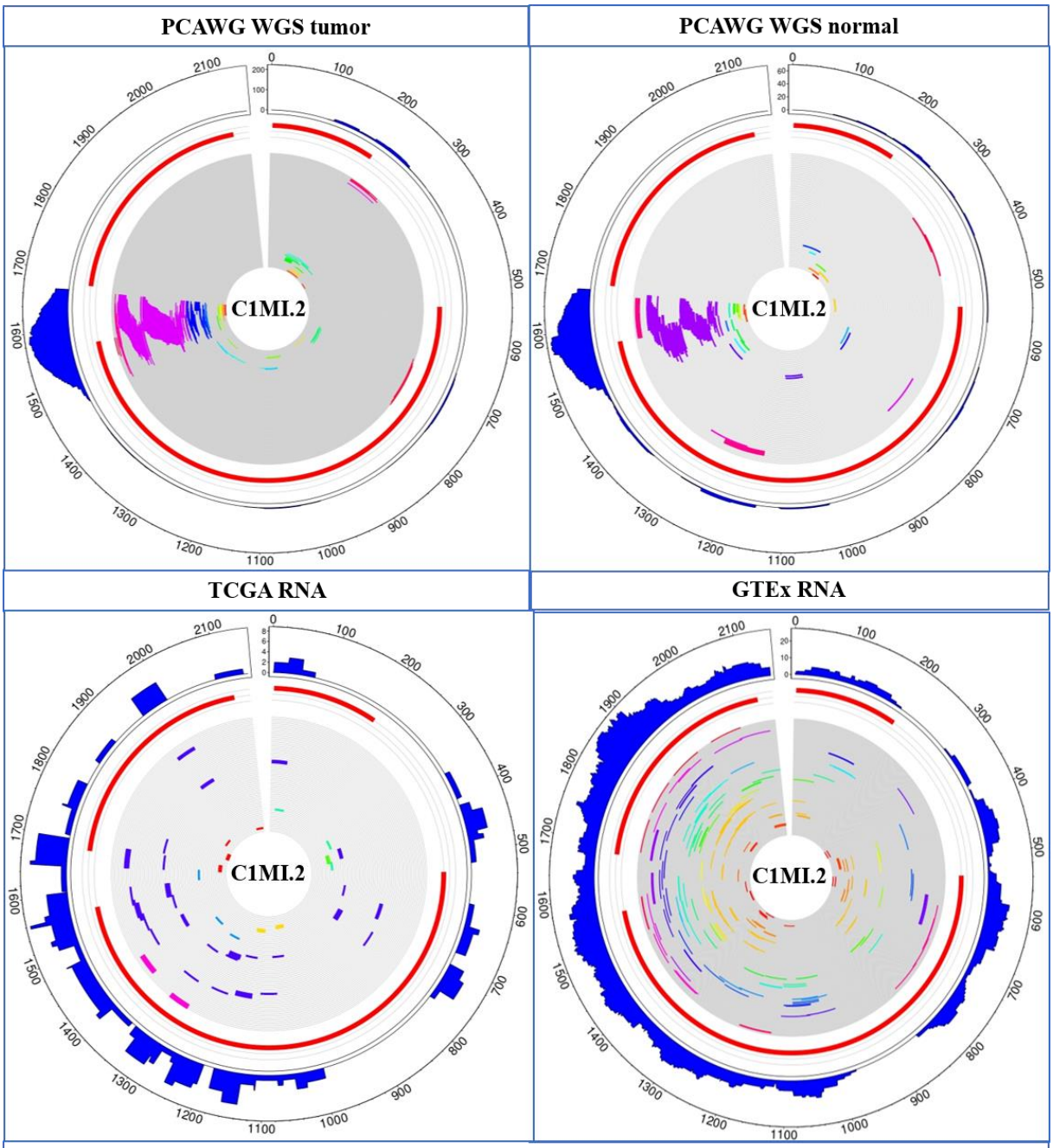


Fig. 3.41: Reads detected for BMMF group 1 isolate C1MI.2 in PCAWG WGS data and TCGA and GTEx RNA data: Coverage of BMMF1 subgroup 7 genome C1MI.2 by reads clearly assigned to subgroup 7. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

BMMF1 genome H1MSB.1 was also frequently detected especially in the PCAWG WGS tumor and normal data sets (fig. 3.42). Just as in case of the C1MI.3M.1 reads, there is a clear gap visible in the circular plot of the reads clearly assigned to H1MSB.2 in both the PCAWG WGS tumor and the PCAWG WGS normal data. Taking into account the unclear reads mapping to H1MSB.1, this gap is filled. While the clear reads show a peak ranging from 300 to 400 bp in both the tumor and the normal data, the unclear reads form two peak regions ranging from 0

to 300 bp and from 1600 to the end of the sequence (fig. 3.42). Interestingly, these peaks cover

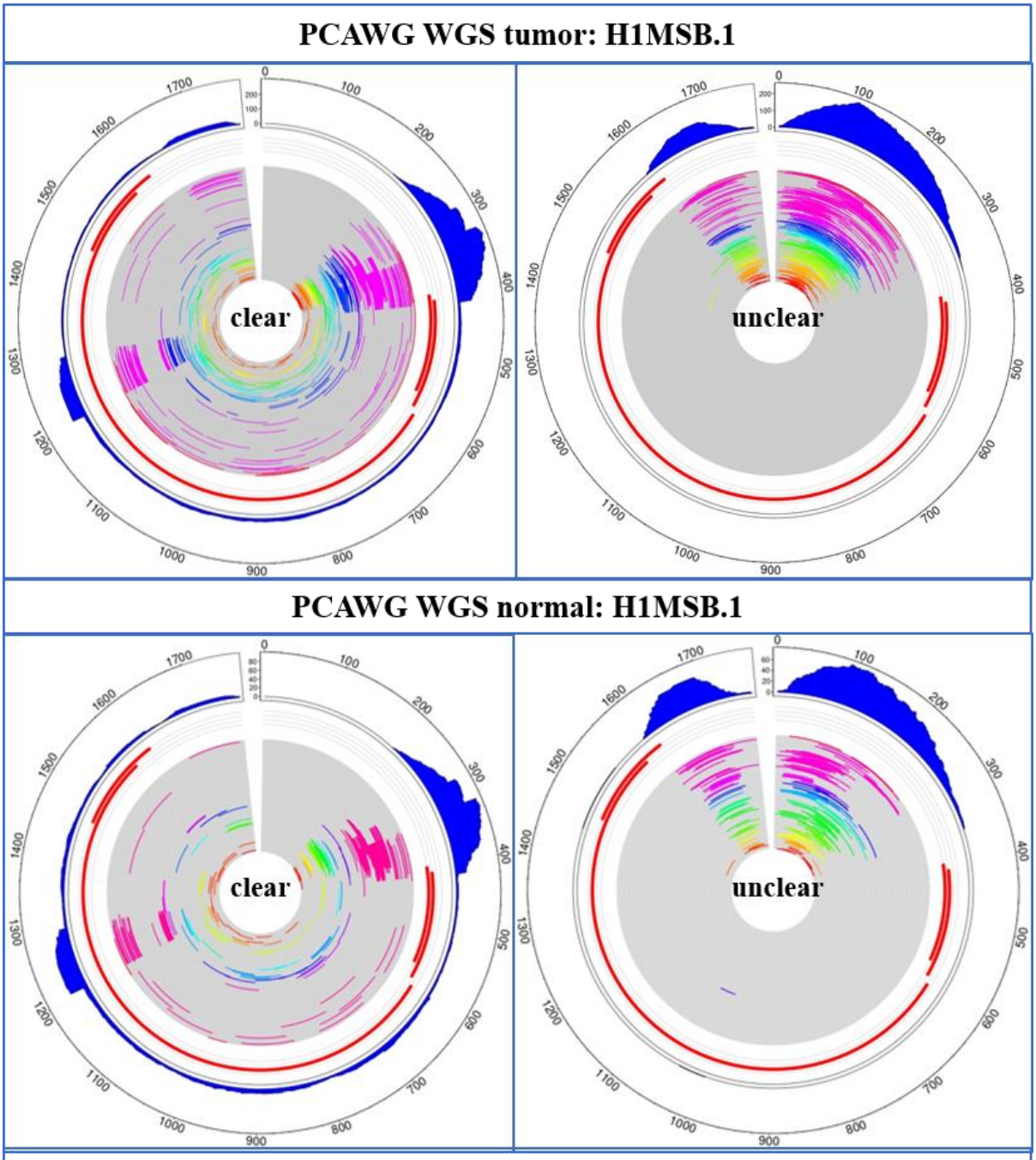


Fig. 3.42: Reads detected for BMMF group 1 isolate H1MSB.1 in PCAWG WGS tumor and normal data: Coverage of BMMF1 subgroup 1 genome H1MSB.1 by reads clearly assigned to subgroup 1 (left) and unclear reads (right). The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same sample are depicted in the same color.

all regions of H1MSB.1, for which no ORFs are predicted (fig. 3.42). Both the tumor and the normal data show one patient with the highest number of clear and unclear detected. This patient is the same Canadian prostate cancer patient, which also caused the C1MI.2 peak region found in the PCAWG WGS tumor and normal data. In total, there are 42 patients with H1MSB.1 reads

detected in the tumor data and 16 patients with H1MSB.1 reads found in the normal data. For 10 of these patients H1MSB.1 reads were found in both the tumor sample and their normal sample. 9 of these normal samples were derived from blood and one from solid tissue distant to the primary tumor. While the TCGA RNA data included only one single H1MSB.1 read, there were more H1MSB.1 reads detected in the PCAWG RNA and GTEx RNA data. While the read numbers were lower compared to the PCAWG WGS data sets, the peak regions described for the PCAWG WGS data sets were also visible in both RNA data sets, if the unclear reads are considered (S21).

3.5.2 Sequence identities of detected BMMF reads

The circular plots showed, that the majority of the detected BMMF reads are either distributed across the entire sequence of the respective BMMF genome or at least cover large parts of it. However, BMMF reads are not only closely related with each other, but also with several bacterial plasmids. For this reason, I analyzed the sequence identity of the detected BMMF reads to the BMMF genomes they are mapped to (fig. 3.43). The majority of the BMMF reads reported for all five data sets share a very high sequence identity with the BMMF genomes they are assigned to. The histogram of sequence identities visualizes that a small fraction of the BMMF reads shows sequence identities between 80 to 90 % when aligned to the respective BMMF genome. More than 90 % of the detected BMMF reads in each data set however exhibit sequence identities of higher than 90 % when they are aligned to the BMMF library sequences. In the TCGA RNA data even 100 % of the BMMF reads detected show sequence identities higher than 90 % (fig. 3.43 G+H). Between 72.7 and 75.3 % of the clear BMMF reads detected in the PCAWG WGS and GTEx RNA data sets exhibit a sequence identity higher than 95 % compared to the respective BMMF genomes. And between 43 and 44 % of the clear reads detected in these three data sets share 100 % sequence identity with the library sequences they are aligned to. The reads detected in the PCAWG RNA and TCGA RNA data are even of a higher quality. 88.5 % and 91.7 % of the clear reads found in the PCAWG and TCGA RNA data sets show a sequence identity of higher than 95 %, whereas 64.6 % and 60 % of the clearly assigned reads are 100 % identical to the respective BMMF library sequences. Interestingly, the reads that could not be unambiguously assigned to one BMMF subgroups exhibit even higher sequence identities when aligned to the BMMF library than the so-called clear BMMF reads (fig. 3.43 B, D, F, H, J). This indicates, that the unclear reads map generally very well to genomes of the BMMF library. However, these reads most likely cover highly conserved regions on the BMMF genomes, that are identical or nearly identical for many different

sequences, which is why reads mapping to these regions cannot be clearly assigned to a specific BMMF genome.

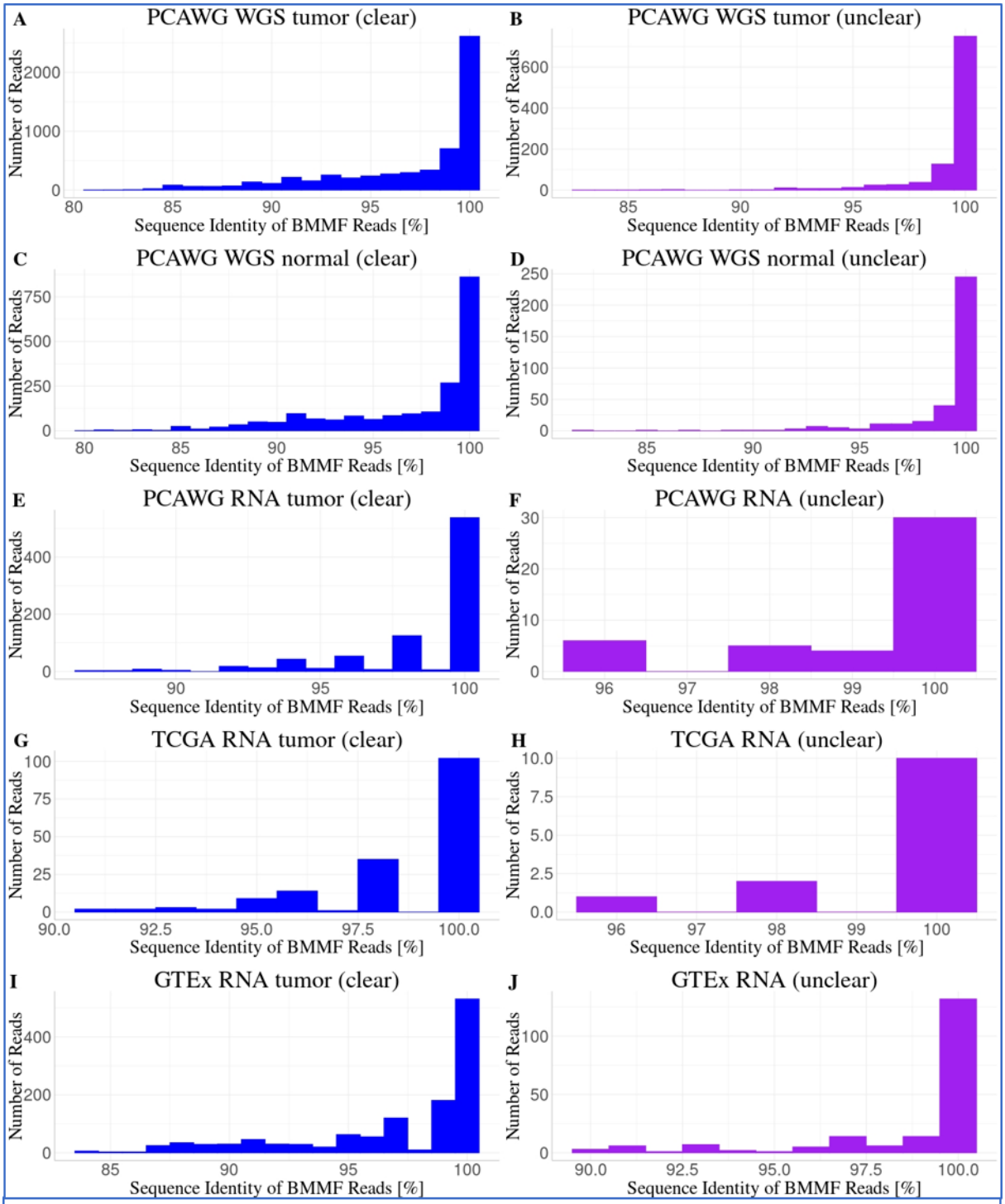


Fig. 3.43: Sequence identity [%] of clear and unclear BMMF reads detected in all five data sets: Overview of the sequence identity of clear (left, blue) and unclear (right, purple) reads detected in all five data sets. The x-axis shows the sequence identity [%], the y-axis the number of clear and unclear reads.

3.6 Analysis of single cell data of lung cancer tumor and metastasis biopsies

Besides of bulk sequencing data sets, I also used D-ViSioN to screen a single cell data set of 30 patients with metastatic lung cancer. The data includes 23420 cells derived from biopsies of both lung tumor and metastasis samples taken from adrenal, brain, liver, lymph node and pleura tissue. I detected 1827 BMMF reads in 131 different cells. Consequently, 0.56 % of the detected cells are BMMF positive. In contrast to the bulk sequencing samples, the three hits threshold is not applied to the single cell data due to the lower sequencing depth per cell compared to bulk sequencing samples. Thus, each cell with at least one BMMF read detected is counted as BMMF positive. The heatmap in figure 3.44 shows an overview of the BMMF subgroups detected in this data set. Due to the larger number of total cells, only positive cells

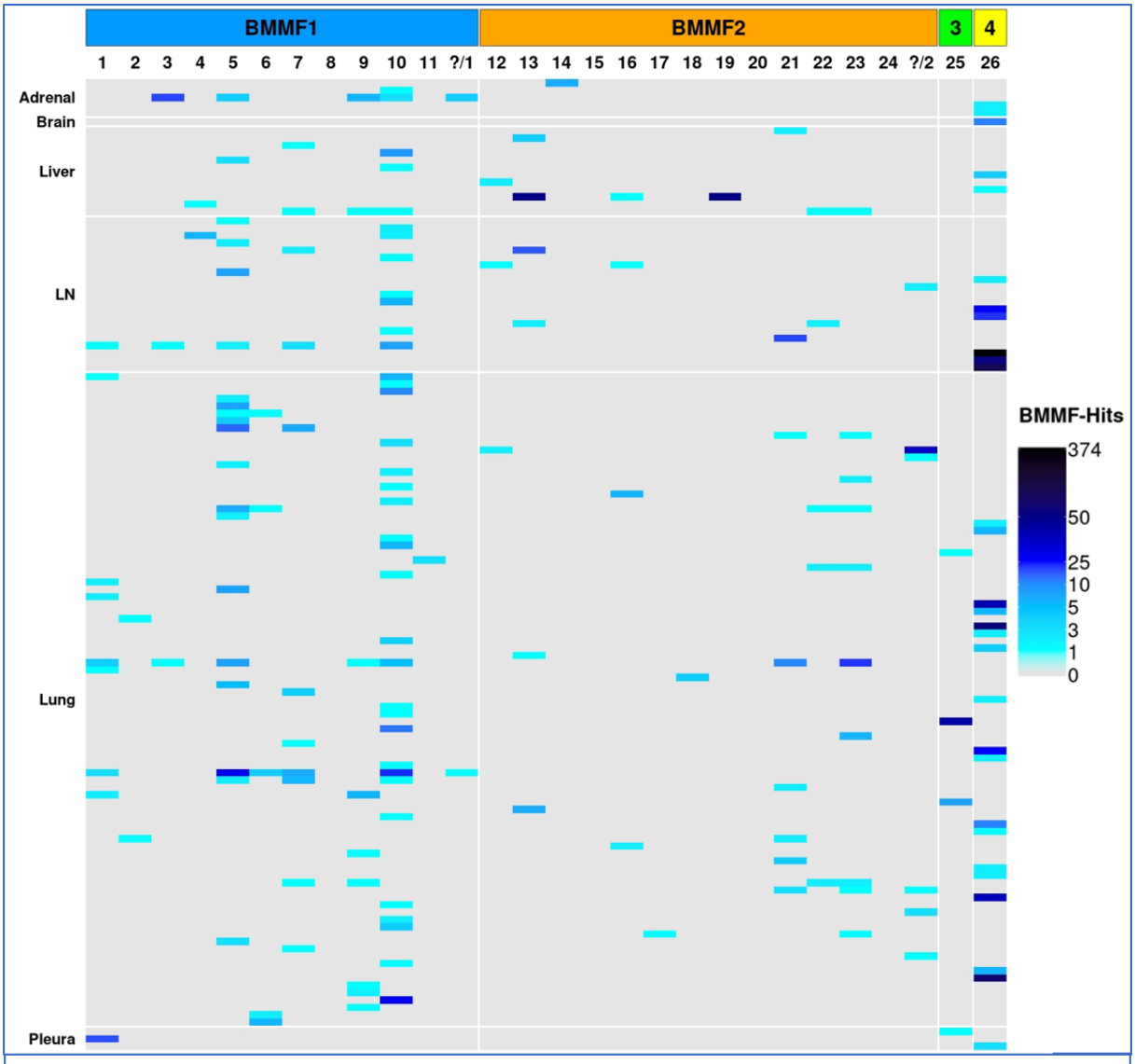


Fig. 3.44: BMMF reads detected per positive cell in single cell data of lung tumor and metastasis biopsies: Overview of BMMF positive cells detected in single cell data of lung cancer tumor and adrenal, brain, liver, lymph node and pleura metastasis biopsies. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

are summarized in this graphic. 89 BMMF positive cells – 67.9 % of all positive cells – are derived from lung tumor samples. The remaining BMMF positive cells originate from adrenal, brain, liver, lymph node and pleura metastases. 16 % of the BMMF positive cells are derived from metastases in lymph nodes and 9.2 % of the BMMF positive cell were found in liver metastases. There were only a few BMMF positive cells detected for the adrenal, brain and pleura biopsies (fig. 3.44). The patterns of the BMMF subgroups detected are similar between tumor and metastasis samples. The BMMF4 group genome MSSI1.162 is most frequently detected both in the primary tumor cells and in the cells derived from different metastases (fig. 3.44). 24.4 % of the BMMF positive cells contain MSSI1.162. BMMF1 subgroup 10 genome C1HB.4 is even detected in 27.5 % of the BMMF positive cells, however there are on average less reads detected per positive cell than for MSSI1.162. Besides of subgroups 10 and 26, BMMF1 subgroup 5 is also frequently detected. 15.6 % of all BMMF positive cells contain subgroup 5. Subgroup 7 hits are detected in 8.4 % of the positive cells and BMMF2 subgroup 23 is found in 7.6 % of the positive cells (fig. 3.44).

MSSI1.162 is the only BMMF genome assigned to BMMF group 4. In the bulk sequencing data analyzed previously, MSSI1.162 did not belong to the most frequently detected BMMF genomes. However, in the lung cancer single cell data set 57.7 % of the total BMMF reads detected cover MSSI1.162. The entire sequence of the MSSI1.162 genome is covered by reads detected in different BMMF positive cells of the single cell data set (fig. 3.45). While all regions of the BMMF4 genome are found, some parts of the genome are covered by higher numbers of reads such as the peak regions ranging from 300 to 450 bp and from 800 to 900 bp (fig. 3.45). There are

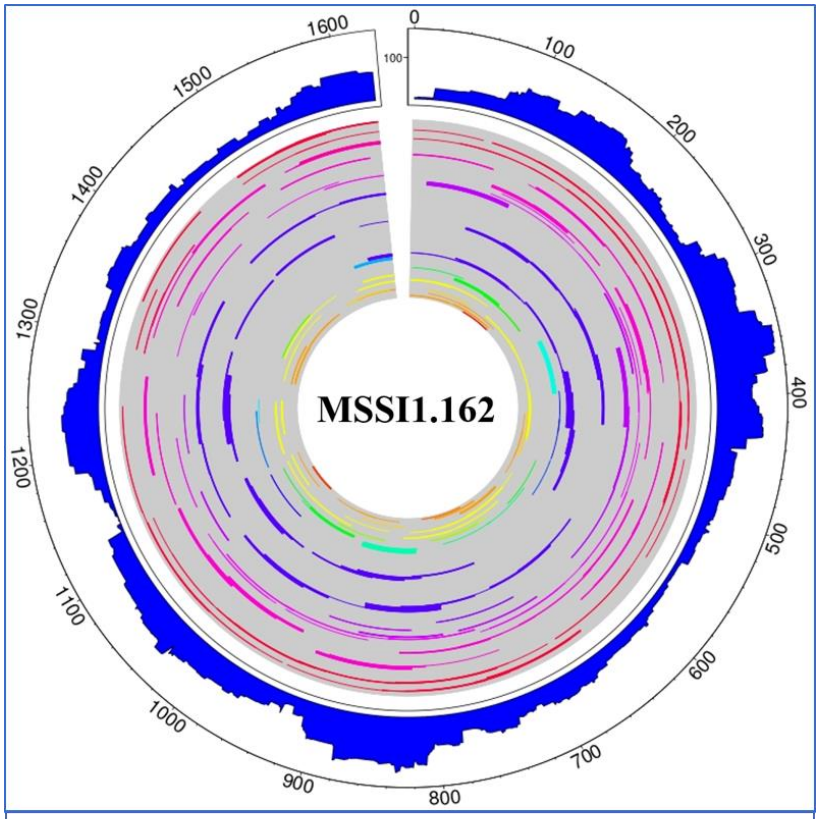


Fig. 3.45: Reads detected for BMMF group 4 isolate MSSI1.162 in single cell data of lung tumor and metastasis biopsies: Coverage of MSSI1.162 genome by reads detected in single cell data of lung cancer tumor and adrenal, brain, liver, lymph node and pleura metastasis biopsies. The outer circle shows in blue the number of detected reads per base pair of the genome of MSSI1.162. The inner circle maps the individual reads detected for the MSSI1.162 genome. Reads detected in the same cell are depicted in the same color.

several cells positive for BMMF group 4, which contain a high number of reads of MSSI1.162. Additionally, there are several cells with reads covering the full sequence of MSSI1.162 (fig. 3.45). This shows both a strong and reliable BMMF4 signal in the lung cancer single cell data cohort.

While the only BMMF1 subgroup 10 genome C1HB.4 was detected in a higher number of different cells than MSSI1.162, the total number of reads detected is much lower. C1HB.4 is covered by 157 reads – 8.6 % of the total BMMF reads detected in the entire single cell data set. The C1HB.4 reads detected cover different regions of the C1HB.4 genome, however for some parts of this genome no reads were detected at all. On the other hand, two peak regions from 100 to 200 bp and from 1250 bp to 1400 bp are standing out with the highest read coverage (fig. 3.46). The average number of C1HB.4 reads per positive cell is low, there are only a few cells with multiple reads covering larger parts of the genome.

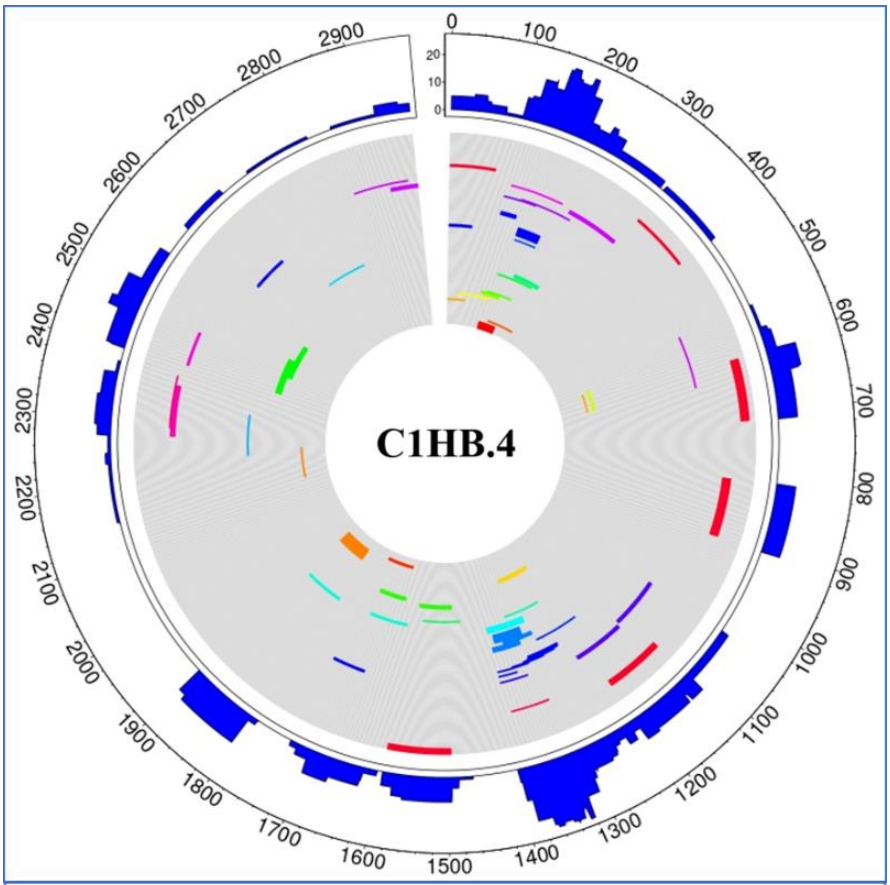


Fig. 3.46: Reads detected for BMMF group 1 isolate C1HB.4 in single cell data of lung tumor and metastasis biopsies: Coverage of BMMF1 subgroup 10 genome C1HB.4 by reads detected in single cell data of lung cancer tumor and adrenal, brain, liver, lymph node and pleura metastasis biopsies. The outer circle shows in blue the number of detected reads per base pair of the genome of C1HB.4. The inner circle maps the individual reads detected for the C1HB.4 genome. Reads detected in the same cell are depicted in the same color.

The remaining BMMF subgroups detected are found in less samples and covered by lower read numbers. BMMF1 subgroup 5 was the third most detected subgroup with most of the detected reads mapping to the isolate C1MI.3M.1 (fig. 3.47) While this BMMF template is detected in about 10 different samples, the read numbers per sample are low. The region ranging from 0 to 300 bp is not covered by any reads as well as two regions on the genome with a length of about 100 bp. Besides of a peak region at around 1650 bp, the remaining parts of the genome are covered by low numbers of overlapping reads (fig. 3.47).

The BMMF2 signal detected in the single cell sequencing data is distributed on many different subgroups as well as on different isolates within those groups. BMMF2 subgroup 23 isolate C2MI.5A.4 was detected in five different cells with reads covering different short sections of the genome (fig. 3.47). These sparse reads hint at the presence of BMMF2-related sequences in the data, however are not enough to reliable indicate the detection of a certain BMMF isolate or a BMMF subgroup. In case of other BMMF2 genomes such as Sphinx 2.36 only one very short section of the genomes is covered by many reads of two different samples. In cases like this the signal is more likely caused by artifacts or BMMF2-related bacterial plasmids instead of actual BMMF2 signal. The BMMF4 reads however are likely derived from real MSSI1.162 signal and the detected BMMF1 reads – especially in case of C1HB.4 and C1MI.3M.1 – are also likely caused by actual BMM1 sequences.

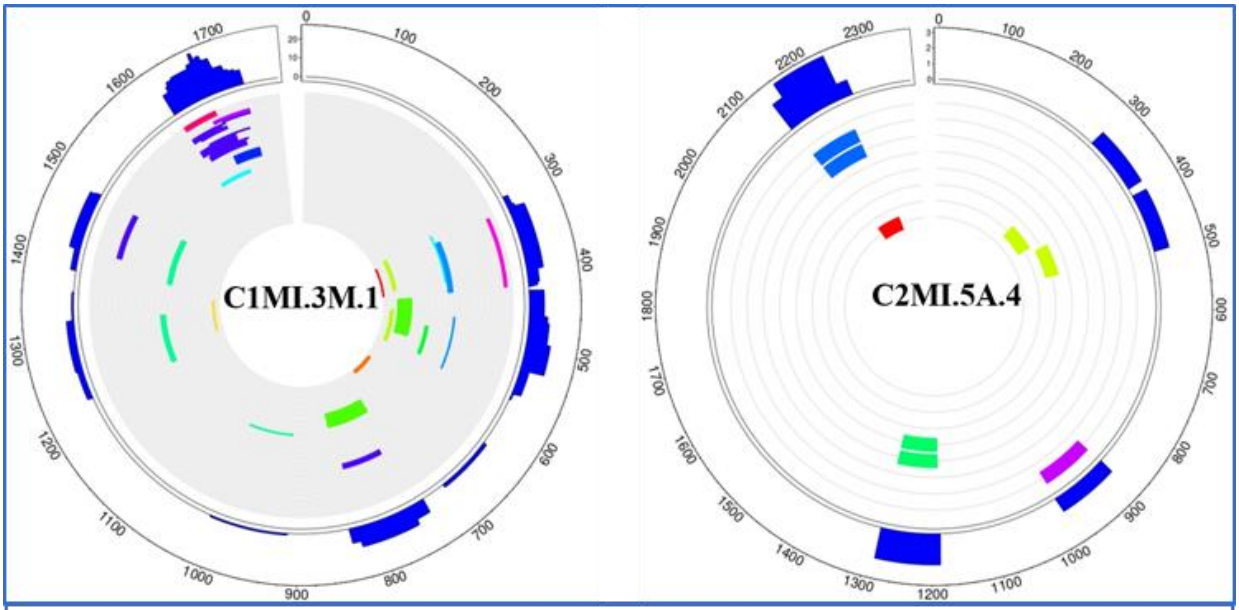


Fig. 3.47: Reads detected for BMMF group 1 isolate C1MI.3M.1 and BMMF group 2 isolate C2MI.5A.4 in single cell data of lung tumor and metastasis biopsies: Coverage of BMMF1 subgroup 5 genome C1MI.3M.1 (left) and BMMF2 subgroup 23 genome C2MI.5A.4 (right) by reads detected in single cell data of lung cancer tumor and adrenal, brain, liver, lymph node and pleura metastasis biopsies. The outer circle shows in blue the number of detected reads per base pair. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same cell are depicted in the same color.

4. Discussion

BMMFs were suggested as infectious agents, which indirectly promote carcinogenesis of colorectal cancer as well as a range of other inflammation-linked cancer types. According to this model as well as wet-lab experiments that found BMMF DNA and Rep protein in tissues adjacent to the primary tumor for colorectal cancer samples, BMMF DNA and RNA should be detectable *in silico* as well in colorectal cancer data and potentially also in other inflammation-linked cancer types. *In silico* analyses of high-throughput sequencing data of cancer types-of-interest for BMMF research could not only support wet-lab results, but also help to expand the analysis by targeting cancer types and BMMF genomes not investigated in experiments so far.

Therefore, I analyzed six sequencing data sets of publicly available sequencing projects as well as one single cell data set using the D-ViSioN algorithm. The D-ViSioN algorithm was originally designed to detect viral reads and integration sites, but the results of the screened data sets indicated, that the algorithm can also be successfully applied for BMMF detection. For this, I used a library consisting of 173 BMMF genomes to screen three WGS and RNA sequencing data sets each for BMMF detection. The RNA data sets include the cancer sequencing data provided by the PCAWG and TCGA projects as well as the healthy tissue data of the GTEx project, which was analyzed to draw a case versus control comparison between the BMMF detection in tumor and normal tissue samples. The analyzed WGS data sets comprised the PCAWG tumor and normal data sets as well as the cell line sequencing samples of the DepMap project. The PCAWG WGS data sets contain both tumor and control samples of the same cancer patient derived from blood or solid tissue samples close or distant to the tumor, which are analyzed to search for tendencies of BMMF detection pointing towards a direct or indirect impact of BMMFs on carcinogenesis. Finally, a single cell data set of lung cancer tumor and metastasis samples was screened for BMMFs to examine, if BMMFs can also be identified in single cell data using D-ViSioN.

4.1 DepMap cell line analyses indicate low background contamination rates

The WGS data of the cell lines included in the DepMap project showed no positivity upon application of the three-hits threshold. Due to the presumed indirect mode of action of BMMFs, the cell line sequencing data was expected to be BMMF negative. The DepMap data was mainly analyzed with the objection to obtain an overview of the background detection levels, that can be expected for BMMFs. The lack of BMMF detection after application of the three-hits threshold confirms the assumption, that cell line data would be BMMF negative. Additionally,

this indicates that only low levels of BMMF background contaminations have to be expected. The three-hits threshold seems to successfully account for these low contamination levels by filtering them out. On the other hand, this filtering step also drastically reduces the number of samples defined as positive in the RNA and WGS tumor and normal tissue data, which might also cause false negative samples.

4.2 Higher BMMF positivity rates and read numbers detected on DNA level than on RNA level

Comparing the BMMF detection rates in WGS and RNA data, the PCAWG WGS tumor and normal tissue/blood data sets show higher positivity rates than the TCGA, PCAWG and GTEx RNA data sets. After the application of the three-hits threshold, 8.8% of the PCAWG WGS tumor samples and 4.6% of the PCAWG WGS normal tissue/blood samples are BMMF positive. On the contrary, only 2.3% of the PCAWG RNA samples, 0.8% of the TCGA RNA samples and 2.5 % of the GTEx RNA samples are BMMF positive. Consequently, BMMFs are detected at a higher percentage in WGS tumor samples compared to WGS normal tissue/blood samples. On the other hand, the detection rates for the PCAWG RNA tumor samples are on the same level as the percentage of BMMF positive samples in the healthy tissue samples of the GTEx RNA data. Both the PCAWG RNA data set and the GTEx RNA data set contained higher positivity rates than the TCGA RNA tumor data.

The absolute read numbers detected for the respective data cohorts, show a similar pattern: The PCAWG WGS tumor data set contains the highest number of BMMF reads of all data sets with 6902 BMMF reads detected – three times as many as the 2335 BMMF reads detected in the PCAWG WGS normal tissue/blood data. The TCGA RNA data set did not only contain the lowest percentage of positive samples, but also the lowest absolute BMMF read numbers. While the GTEx RNA and PCAWG RNA data sets have similar positivity rates, the absolute read numbers detected in the healthy tissue GTEx RNA data are higher than in the tumor PCAWG RNA data.

4.2.1 Strongest normalized BMMF signal in PCAWG RNA data

After normalization for cohort size and sequencing depth, the PCAWG RNA tumor data stands out with 29.5 BMMF reads per billion mapped reads per 100 samples reported, whereas the GTEx data and the TCGA data only contained 1.8 respectively 1.5 BMMF reads per billion mapped reads per 100 samples. While this points towards higher BMMF detection in the

PCAWG RNA tumor data compared to the healthy tissue RNA data, the low BMMF signal in the TCGA RNA data contradicts this observation.

While the PCAWG WGS tumor data sets contained both higher BMMF positivity rates and higher BMMF read numbers compared to the PCAWG WGS normal tissue/blood data, the normalized BMMF signal is only slightly higher in the PCAWG WGS tumor data than in the PCAWG WGS normal tissue/blood data. This is mainly caused by the on average higher sequencing depths across the PCAWG WGS tumor samples. With 1.1 respectively 0.9 BMMF reads per billion mapped reads per 100 samples detected, both PCAWG WGS data sets in total exhibit a low BMMF signal.

The normalization for sequencing depth and cohort size, puts the observations made for positivity rate and absolute read numbers into a new context, since it elevates the BMMF detection in RNA data sets compared to the WGS data due to the lower sequencing depth in RNA sequencing samples. Additionally, this step emphasizes, that BMMF detection in WGS and RNA data cannot be compared directly due to the different experimental set-ups.

4.3 Highest BMMF detection in PCAWG RNA acute myeloid leukemia cohort

The PCAWG RNA data showed the highest positivity rates in the ovarian, stomach and uterine cancer cohorts with between 10% and 20% of the samples being BMMF positive. Upon normalization of the detected reads, the US acute myeloid leukemia cohort stood out with the strongest normalized BMMF signal across all RNA and WGS data sets. Additionally, the ovarian and uterine cancer cohorts exhibited increased BMMF signals upon normalization in the PCAWG RNA data set.

The GTEx data RNA contained less than 5% positive samples for all tissue cohorts, whereas the TCGA data contained less than 1% positive samples for all cancer types. Considering the normalized BMMF signal, the TCGA RNA data shows the highest detection in the colon and pancreatic cancer cohorts, whereas the GTEx RNA data exhibits the strongest BMMF signal in breast, uterus, liver, lung and prostate healthy tissue samples. However, the TCGA and GTEx cohorts contained much lower normalized BMMF read numbers compared to the positive PCAWG RNA cohorts

4.3.1 High BMMF signal in European breast and Canadian prostate cancer data

In the PCAWG WGS data sets the breast, lung, prostate and stomach cancer cohorts stood out with the highest normalized BMMF signals in both the tumor and the normal tissue/blood data.

On the other hand, the colon and uterine cancer cohorts only exhibited elevated BMMF signal in the PCAWG WGS normal data. The European breast and Canadian prostate cancer cohorts contained the highest absolute reads numbers across all cohorts in both the tumor samples and the non-tumor samples.

For breast, liver, ovarian, pancreatic and prostate cancer as well as for acute myeloid leukemia, there is more than one cancer cohort available within the PCAWG WGS data. Interestingly, the cohorts for the same cancer type derived from different geographic locations often showed strong discrepancies regarding their BMMF detection. Whereas the PCAWG WGS European and UK breast cancer cohorts are strongly positive in case of both the tumor and the non-tumor samples, the US breast cancer cohort is negative in the tumor samples and only weakly positive in the normal tissue/blood samples. A similar observation can be made in case of the prostate cancer cohorts: The Canadian Prostate Cancer cohort exhibited a strong BMMF signal in both tumor and non-tumor samples. The UK prostate cancer cohort showed only positivity in the tumor data, whereas only the non-tumor samples of the US prostate cancer cohorts exhibited BMMF detection. Additionally, the early-onset prostate cancer cohort shows strong BMMF detection in the non-tumor samples, while there is hardly any signal in the tumor data of this cohort. In case of both pancreatic and ovarian cancer, there is each one negative and BMMF positive cohort included in the PCAWG WGS data sets.

These strong differences between cohorts of the same cancer type could be caused by the application of different methods during surgery, sample extraction or analysis. However, it is also possible, that the BMMF library used is better adapted to the BMMF landscape found in cohorts from certain geographic locations. Since most of the BMMFs were first isolated from milk and serum samples in Germany, this might for example explain the higher levels of BMMF detection in the European and UK breast cancer samples compared to the US breast cancer data.

4.4 BMMF1 detection dominates across all data sets

In all three RNA data sets as well as in the PCAWG WGS data, BMMF1 reads make up for between 78.7% and 90.2% of the total BMMF reads detected. Consequently, BMMF1 is dominating both regarding the percentage of positive samples with BMMF1 reads detected and with regards to the normalized read numbers reported. This might indicate a generally more frequent occurrence of BMMF1 in the human host.

BMMF group 2 is the second most frequently detected BMMF group. In the PCAWG RNA data set, the US acute myeloid leukemia cohort stands out as the only cohort with BMMF2

signal, however BMMF2 detection is present in a higher number of different cohorts in the remaining data sets. The PCAWG WGS tumor samples of the US acute myeloid leukemia cohort, of the US kidney renal clear cell carcinoma cohort and of the Japanese liver cancer cohort LINC-JP stand out with high levels of BMMF2 detection. In the PCAWG WGS normal data set, the South Korean acute myeloid leukemia and French liver cancer cohorts stand out with high BMMF2 detection.

BMMF3 detection seems to play only a minor role across all data sets. BMMF group 4 is also rarely reported, however the GTEx RNA breast, prostate and blood cohorts exhibit BMMF4 detection. Additionally, BMMF4 was found in tumor and non-tumor samples of the PCAWG WGS European breast and Canadian prostate cancer cohorts.

4.4.1 BMMF clusters 1, 5, 6, 7, 10 and 21 most frequently reported

The four main BMMF groups were split into 26 different subgroups to obtain an overview which BMMF genomes within the subgroups are present in which cancer cohort. Since some BMMF genomes are so closely related, that it is virtually impossible to unambiguously assign a detected BMMF read to a specific sequence, the formation of subgroup at least helps to narrow it down to few closely related isolates. The introduction of the different subgroups improves the assignment of different BMMF1 reads to certain variants significantly. Especially in case of the BMMF group 2 subgroups, almost all detected BMMF reads can be assigned to a specific subgroup. Only a small fraction of BMMF2 reads match more than one subgroup equally and cannot be clearly sorted in one subgroup. The BMMF1 sequences are closer related than the BMMF2 isolates, which causes a larger share of unclear reads. However, between 85% and 92% of the total reads detected in the five RNA and WGS data sets can be clearly assigned to one specific subgroup. Consequently, the defined subgroups successfully help to provide a better classification of the detected BMMF reads.

In the PCAWG RNA data, BMMF1 subgroup 10 stands out as the most frequently reported BMMF cluster, followed by BMMF1 subgroups 5 and 6 and BMMF2 subgroup 21. BMMF1 subgroup 10 is also the only subgroup detected in all five TCGA RNA cohorts. In the GTEx RNA data, BMMF1 subgroup 5, 6, 7 and 10 as well as BMMF2 subgroups 13, 17 and 21 are reported in several different tissue cohorts with high normalized BMMF signals. The PCAWG WGS tumor cohorts exhibit elevated detection of BMMF1 subgroups 1, 5, 6, 7 and 10 as well as of BMMF2 subgroups 17 and 21. In the PCAWG WGS non-tumor data, subgroups 1, 5, 6, 10 and 21 stand out with both a high signal and frequent discovery in different cohorts.

Additionally, the subgroups 7, 12 and 13 show a strong BMMF signal – especially in the South Korean acute myeloid leukemia cohort. In conclusion, several subgroups are repeatedly detected in a broad range of different cohorts and data sets. This includes BMMF1 subgroups 1, 5, 6, 7 and 10 as well as BMMF2 subgroup 21.

4.5 Statistical analysis of subgroups detected in tumor and non-tumor samples

The statistical comparison of the subgroups reported for matching RNA and WGS tumor and normal tissue data cohorts was performed using the Zero-Inflated Rank Test, which is an adaption of the Wilcoxon rank sum (Mann-Whitney U) test for data sets with uneven sample sizes and a high number of zero values, that was first developed for the analysis of microorganisms in the human microbiome (Wang, 2021). Since the majority of the samples of all data sets are BMMF negative, the BMMF results contain many zero values, especially if the detected reads are additionally split into different subgroups. Consequently, the ZIR test fits the BMMF detection data better than the statistical test traditionally used to compared non-normally distributed data sets.

4.5.1 BMMF signal significantly increased in PCAWG WGS tumor samples of Canadian prostate cancer cohort

The comparison of the PCAWG WGS tumor and normal data showed only small differences regarding the normalized read numbers between these data sets for most of the cohorts and subgroups. Only five cohorts contained at least one subgroup with a statistically significant difference between the tumor and non-tumor samples. This includes the German early-onset prostate cancer cohort, which contains significantly less BMMF signal for subgroups 1, 5, 6 and “BMMF1 unclear” in the tumor samples than in the blood samples of the PCAWG WGS normal data set. On the other hand, the US kidney chromophobe cohort shows a significantly increased BMMF detection in the tumor samples for subgroups 1, 6 and “BMMF1 unclear”. The US lung squamous cell carcinoma cohort also shows a significantly higher BMMF signal in the tumor for the subgroups 6, 9 and 21. The Canadian pancreatic cancer cohort exhibits significantly elevated levels of BMMF1 subgroup 10 and BMMF2 subgroup 21 in the tumor data compared to the PCAWG WGS normal data. Additionally, the Canadian prostate cancer data set shows a significant increase of BMMF detection in the tumor data for all BMMF1 subgroups but 9 and 11 as well as for BMMF2 subgroups 17, 20, 21 and 22 and for BMMF group 4.

While the European and UK breast cancer cohorts are two of the cohorts with the highest absolute read numbers reported, the differences between the tumor and non-tumor data are too small to be statistically significant. Even splitting the origins of normal samples into blood and tissue adjacent to the primary tumor, does not lead to statistically significant differences. However, in case of the US stomach cancer cohort splitting the PCAWG WGS blood and solid tissue samples results in a significant increase of the detection of BMMF1 subgroup 1 in the tumor data compared to the blood samples of the PCAWG WGS normal data set.

The French liver cancer data set contains very high normalized read numbers for both the tumor and the normal data derived from tissue adjacent to the primary tumor. Since the presence of BMMFs is expected to be higher in the peritumor than in the tumor, normal tissue samples from regions close to the primary tumor are not suited for a case versus control comparison of BMMF detection. In fact, samples derived from tissue adjacent to the primary tumor are expected to be positive, if they were taken from tissue close enough to the primary tumor to be counted as peri-tumor samples.

4.5.2 Case versus control comparison of healthy tissue and tumor RNA data

The statistical comparison of the GTEx healthy tissue data with the corresponding TCGA and PCAWG RNA cohorts was performed to discriminate BMMF types, that might be present in both healthy and tumor samples, from potential high-risk BMMF types, which might occur at increased frequencies in cancer samples. Since high- and low-risk profiles might differ for each cancer type, a pairwise comparison of tumor and matched healthy tissue data is conducted for all cohorts and subgroups.

In general, except for the TCGA pancreatic cancer cohort, there were subgroups identified with significantly increased detection in the healthy tissue data for all other cohorts in the TCGA RNA data set. Especially BMMF1 subgroups 5, 6, 7 and 10 show increased detection in the GTEx data compared to the TCGA data. There was no significantly increased BMMF signal in the tumor data reported for the TCGA-RNA data set.

The comparison of the PCAWG RNA data with the GTEx RNA data also revealed a set of significantly increased BMMF subgroups in the GTEx data for most cohorts. On the other hand, subgroup 6 in the acute myeloid leukemia and ovarian cancer cohorts, subgroup 5 in the stomach cancer data, subgroup 8 in the uterine cancer cohort and subgroup 21 in the acute myeloid leukemia data show significantly increased detection in the tumor compared to the healthy tissue data. Consequently, these BMMF subgroups might be potential candidates for

high-risk BMMF types. However, the very uneven samples size between the PCAWG RNA and the GTEx RNA data sets might limit the performance of the statistical test in case of some comparisons. The acute myeloid leukemia cohort is for example nearly 40 times smaller than the respective GTEx blood cohort it is compared to. While the ZIR test is described to be adapted for uneven sample sizes, this huge difference might still affect the statistical testing.

4.6 Detection of potential high-risk BMMF genomes

The quality assessment of the detected BMMF reads of genomes-of-interest was performed by analyzing the sequence identity of the BMMF reads compared to the respective BMMF library sequences as well as by visualization of the locations on the BMMF genomes the reads map can be aligned to. Generally, the detected BMMF reads match the respective library sequences with a very high sequence identity. Reads that could not be unambiguously assigned to one subgroup share on average even a higher sequence identity to the respective sequences of the BMMF library than clear reads. This indicates, that reads are not classified as unclear reads because of lack of similarity to the BMMF genomes they map to, but because they frequently match highly conserved sequences that are highly similar or even identical between BMMF genomes across different subgroups.

Considering both absolute and normalized numbers of BMMF reads detected as well as the statistical comparisons, BMMF1 subgroups 1, 5, 6 and 10 as well as BMMF2 subgroup 21 are frequently standing out compared to the remaining BMMF subgroups. Consequently, the isolates of these subgroups are especially interesting for BMMF research. This includes C1MI.3M.1 (subgroup 5), H1MSB.1 (subgroup 1), C1HB.4 (subgroup 10) and Sphinx 2.36 (subgroup 21).

4.6.1 Circular coverage plots can distinguish specific BMMF detection from artifacts

In spite of the mostly high sequence identities of the BMMF reads, it is still important additionally analyze the coverage of the respective BMMF genomes by the detected reads. While the combined clear and unclear reads of some of the most frequently found sequences such as C1MI.3M.1, H1MSB.1, or Sphinx 2.36 cover the entire genomes of the BMMF isolate, there are other BMMF isolates, such as HCBI8.215, for which only a very short section is covered by a high number of reads. The latter example is most likely an artifact and not a real BMMF signal.

The BMMF genome C1MI.1 (subgroup 7), which was most frequently detected in the TCGA RNA and GTEx RNA data also shows a similar peak region in the reads detected in the PCAWG WGS tumor and blood samples of one single Canadian prostate cancer patient. However, in contrast to the very short, steep peaks in case of the BMMF3 genome, the C1MI.2 peak comprises about 150 bp. Since the data of this Canadian prostate cancer patient contains an unusually high number of different BMMF reads including of several other BMMF1 isolates, it seems likely that these reads are caused by a real BMMF1 signal, but probably by a genome with shares a high sequence identity with C1MI.2 in this region and not by C1MI.2 itself.

Subgroup 10 genome C1HB.4 was consistently found in all RNA, WGS and single cell data sets. While there is generally the entire genome or at least large fractions of it detected, some of the data sets such as the GTEx RNA or PCAWG WGS tumor data sets show two short peak regions, which are covered by higher read numbers. On the other hand, one patient in the PCAWG RNA data set with very high C1HB.4 read numbers shows a gap of about 200 bp. This could indicate the presence of a truncated version of C1HB.4.

The comparison of the detected BMMF reads with the predicted open reading frames generally did not reveal any clear correlation between the location of the BMMF reads and the open reading frames.

4.7 BMMF detection in single cell data

While I mainly focused on the analysis of WGS and bulk RNA sequencing data, I also screened a single cell data set of metastatic lung cancer patients to investigate the feasibility of BMMF detection in single cell data using D-ViSioN. Since the sequencing depth per single cell is lower than in a bulk sequencing sample, the three-hits threshold was not applied during this analysis. In total, 0.56% of the screened cells were found to be BMMF positive. This shows, that the D-ViSioN workflow is also suited to analyze single cell data sets. The detection of BMMFs in single cell data provides new opportunities for future analysis to for example examine the colocalization of BMMFs and specific immune cells, that has been described in wet lab experiments (Nikitina *et al.*, 2023c).

4.8 Limitations of the analyses

The biggest limitation to the analyses for *in silico* BMMF detection is the lack of data availability of peri-tumor samples. Wet-lab analyses previously showed especially in case of colorectal cancer, that the highest BMMF prevalence can be detected in the periphery of tumors,

but not in the primary tumor sample (Nikitina *et al.*, 2023c). However, publicly available sequencing projects usually just contain primary tumor samples.

Additionally, the analysis is limited by the BMMF library, which mostly consists of genomes isolated from tissue, serum or milk samples derived from Germany. Consequently, the library might not be ideally adapted to detect BMMFs in sequencing data derived from other geographic locations. Since the detected BMMF reads usually map to the library with a very high sequence identity, the D-ViSioN workflow seems to be well suited to detect known BMMF genomes in sequencing data. However, the algorithm is not detecting unknown isolates, that are more distantly related to the known genomes.

Another problem during screenings for BMMFs, is caused by the high sequence identity between BMMF genomes to *A. baumanii* plasmids. While the detected reads match the library sequences very well, there might still be highly conserved regions on bacterial plasmids, that could overlap and interfere with BMMF detection. To address this problem, the BMMF detection workflow could be expanded to use more than virus detection pipeline to validate the detected BMMF reads.

4.9 Conclusion & Outlook

The analysis of different data sets of WGS, bulk RNA and single cell sequencing data revealed, that Bovine Meat and milk Factors can be detected in a large range of cancer types as well as in healthy tissue samples derived from different organs. While the analysis of the DepMap cell line data as negative control did not indicate the frequent presence of BMMF contaminations, I still applied a filter step using a threshold to focus on samples with a reliable BMMF signal. The number of reads detected per sample is on average low, however there are still multiple samples across the different data sets analyzed, which contained higher read numbers covering large fractions or even the entire sequence of the respective BMMF genomes.

The statistical comparison of the PCAWG WGS tumor and normal data revealed an increased detection of certain subgroups in the tumor data of kidney, lung, pancreatic, prostate and stomach cancer cohorts. On the other hand, the PCAWG WGS data sets included more than one cohort for several of these cancer types, which showed very different levels of BMMF detection. Consequently, there is further research and sequencing data needed to examine the role of BMMFs for these cancer types and to understand the different BMMF signals detected in different cohorts of the same cancer type. The case versus control comparison of the GTEx RNA data with the TCGA RNA and PCAWG RNA data, showed mostly an increased BMMF

signal in the healthy tissue data for several different BMMF subgroups. Only the PCAWG RNA acute myeloid leukemia, ovarian, stomach and uterine cancer cohorts show at least one subgroup with higher BMMF detection in the tumor data.

In general, it is no surprise to find BMMF sequences present in a range of different healthy tissue sites. After all, the exposure to BMMFs during long periods of latency is expected to contribute to cancer development, which would mean a constant presence of BMMFs in healthy tissue prior to malignant transformation. The presumed indirect mode of BMMFs also means, that the highest prevalence of BMMFs is not expected for the tumor tissue itself, but rather for the peritumor. The PCAWG WGS normal samples are mainly obtained from blood samples, solid tissue samples from locations distant to the primary tumor and adjacent tissue of the primary tumor. Consequently, BMMF detection can be expected in the samples derived from adjacent regions to the primary tumor and to a certain extent also in blood samples. Several liver and breast cancer samples derived from tissue adjacent to the primary tumor were found to be BMMF positive, but for most cancer types, peritumor samples are not available within the analyzed data set.

Across all analyses performed, there are several cancer types standing out in certain ways. While breast cancer, liver cancer, prostate and pancreatic cancer have been investigated by wet lab experiments, stomach and ovarian cancer have not been experimentally examined so far. Consequently, these two cancer types might be of interest for future research. The colon cancer data did neither indicate a high BMMF signal nor any bias towards detection in tumor samples. Since colorectal cancer has always been a main target of BMMF research, these findings were somewhat expected. Wet-lab experiments previously showed, that there is little BMMF detection to expect in pure tumor samples, while there is increased BMMF presence in the peri-tumor (Nikitina *et al.*, 2023a; Nikitina *et al.*, 2023c).

Across all data sets analyzed, the detection of BMMF1 subgroups 1, 5, 6, 7 and 10 as well as BMMF2 subgroup 21 frequently stood out. Looking at the genomes detected of these subgroups, subgroup 1 genome H1MSB.1, subgroup 5 genome C1MI.3M.1 as well as H1MSB.2, C1MI.2 and C1HB.4 which are the only genomes of subgroups 6, 7 and 10 have been frequently detected with a convincing read coverage. Additionally, BMMF2 subgroup 21 genome Sphinx2.36 has been found in some of the data sets analyzed. The main target of the wet lab analyses so far has been H1MSB.1. While this genome is among the most frequently detected BMMF genomes in the *in silico* analyses, the results also suggest to focus on additional genomes such as C1MI.3M.1 or C1HB.4.

The successful use of D-ViSioN to detect BMMFs in RNA, WGS and single cell sequencing data, suggests that this pipeline can be a useful tool for further *in silico* analyses of BMMFs. Additionally, the expansion of the BMMF library could improve the BMMF detection of D-ViSioN. Since the start of my thesis new BMMF genomes have been isolated from different origins. Future *in silico* analyses for BMMFs should include these new isolates to broaden the BMMF detection. Furthermore, early results of a collaboration with the DKFZ research group "Viral Hepatitis and Liver Cancer" indicate, that BMMFs might just form an evolutionary branch of a large family of bacterial plasmids or plasmid-like sequences. These findings suggest, that the current library of BMMF sequences might only contain a small extract of a larger group of sequences, whose origin, function and impact on the human body are yet to be understood.

7. References

Ahluwalia, N., Andreeva, V.A., Kesse-Guyot, E., and Hercberg, S. (2013). Dietary patterns, inflammation and the metabolic syndrome. Diabetes Metab 39, 99-110. 10.1016/j.diabet.2012.08.007.

Aho, A.V., Kernighan, B. W., Weinberger, P. J. (1987). The AWK programming language (Addison-Wesley).

Alisson-Silva, F., Kawanishi, K., and Varki, A. (2016). Human risk of diseases associated with red meat intake: Analysis of current theories and proposed role for metabolic incorporation of a non-human sialic acid. Mol Aspects Med 51, 16-30. 10.1016/j.mam.2016.07.002.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215, 403-410. 10.1016/S0022-2836(05)80360-2.

Anaconda Software Distribution (2021). Computer software, Version: anaconda3/2021.05 (Anaconda).

Aune, D., Lau, R., Chan, D.S.M., Vieira, R., Greenwood, D.C., Kampman, E., and Norat, T. (2012). Dairy products and colorectal cancer risk: a systematic review and meta-analysis of cohort studies. Ann Oncol 23, 37-45. 10.1093/annonc/mdr269.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603-607. 10.1038/nature11003.

Botsios, S., and Manuelidis, L. (2016). CJD and Scrapie Require Agent-Associated Nucleic Acids for Infection. J Cell Biochem 117, 1947-1958. 10.1002/jcb.25495.

Bouvard, V., Loomis, D., Guyton, K.Z., Grosse, Y., Ghissassi, F.E., Benbrahim-Tallaa, L., Guha, N., Mattock, H., Straif, K., and International Agency for Research on Cancer Monograph Working, G. (2015). Carcinogenicity of consumption of red and processed meat. Lancet Oncol 16, 1599-1600. 10.1016/S1470-2045(15)00444-1.

Brown, K.F., Rumgay, H., Dunlop, C., Ryan, M., Quartly, F., Cox, A., Deas, A., Elliss-Brookes, L., Gavin, A., Hounsome, L., et al. (2018). The fraction of cancer attributable to modifiable risk factors in England, Wales, Scotland, Northern Ireland, and the United Kingdom in 2015. Br J Cancer 118, 1130-1141. 10.1038/s41416-018-0029-6.

Bund, T., Nikitina, E., Chakraborty, D., Ernst, C., Gunst, K., Boneva, B., Tessmer, C., Volk, N., Brobeil, A., Weber, A., et al. (2021). Analysis of chronic inflammatory lesions of the colon for BMMF Rep antigen expression and CD68 macrophage interactions. Proc Natl Acad Sci U S A 118. 10.1073/pnas.2025830118.

Chang, Y., Moore, P.S., and Weiss, R.A. (2017). Human oncogenic viruses: nature and discovery. Philos Trans R Soc Lond B Biol Sci 372. 10.1098/rstb.2016.0264.

Charif, D., Lobry J.R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Structural approaches to sequence evolution: Molecules, networks, populations, U.P. Bastolla, M.; Roman, H.E.; Vendruscolo, M., ed. (Springer Verlag), pp. 207-232.

Chavakis, T., Alexaki, V.I., and Ferrante, A.W., Jr. (2023). Macrophage function in adipose tissue homeostasis and metabolic inflammation. Nat Immunol 24, 757-766. 10.1038/s41590-023-01479-0.

Cheng, L., Wang, Y., and Du, J. (2020). Human Papillomavirus Vaccines: An Updated Review. Vaccines (Basel) 8. 10.3390/vaccines8030391.

Christiansen, T., Foy, B. d., Wall, L.,Orwant, J. (2012). Programming Perl, 4 Edition (O'Reilly).

Clinton, S.K., Giovannucci, E.L., and Hursting, S.D. (2020). The World Cancer Research Fund/American Institute for Cancer Research Third Expert Report on Diet, Nutrition, Physical Activity, and Cancer: Impact and Future Directions. J Nutr 150, 663-671. 10.1093/jn/nxz268.

Correa, P. (1992). Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res 52, 6735-6740.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. Gigascience 10. 10.1093/gigascience/giab008.

de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G.M. (2020). Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. Lancet Glob Health 8, e180-e190. 10.1016/S2214-109X(19)30488-7.

de Villiers, E.M., Gunst, K., Chakraborty, D., Ernst, C., Bund, T., and zur Hausen, H. (2019). A specific class of infectious agents isolated from bovine serum and dairy products and peritumoral colon cancer tissue. Emerg Microbes Infect 8, 1205-1218. 10.1080/22221751.2019.1651620.

de Villiers, E.M., and zur Hausen, H. (2021). Bovine Meat and Milk Factors (BMMFs): Their Proposed Role in Common Human Cancers and Type 2 Diabetes Mellitus. Cancers (Basel) 13. 10.3390/cancers13215407.

Dheilly, N.M., Ewald, P.W., Brindley, P.J., Fichorova, R.N., and Thomas, F. (2019). Parasite-microbe-host interactions and cancer risk. PLoS Pathog 15, e1007912. 10.1371/journal.ppat.1007912.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21. 10.1093/bioinformatics/bts635.

Eilebrecht, S., Hotz-Wagenblatt, A., Sarachaga, V., Burk, A., Falida, K., Chakraborty, D., Nikitina, E., Tessmer, C., Whitley, C., Sauerland, C., et al. (2018). Expression and replication of virus-like circular DNA in human cells. Sci Rep 8, 2851. 10.1038/s41598-018-21317-w.

Emini, E.A., Ellis, R.W., Miller, W.J., McAleer, W.J., Scolnick, E.M., and Gerety, R.J. (1986). Production and immunological analysis of recombinant hepatitis B vaccine. J Infect 13 Suppl A, 3-9. 10.1016/s0163-4453(86)92563-6.

Epstein, A. (2015). Why and How Epstein-Barr Virus Was Discovered 50 Years Ago. Curr Top Microbiol Immunol 390, 3-15. 10.1007/978-3-319-22822-8_1.

Falida, K., Eilebrecht, S., Gunst, K., zur Hausen, H., and de Villiers, E.M. (2017). Isolation of Two Virus-Like Circular DNAs from Commercially Available Milk Samples. Genome Announc 5. 10.1128/genomeA.00266-17.

Farrow, B., and Evers, B.M. (2002). Inflammation and the development of pancreatic cancer. Surg Oncol 10, 153-169. 10.1016/s0960-7404(02)00015-4.

Faupel-Badger, J.M., Arcaro, K.F., Balkam, J.J., Eliassen, A.H., Hassiotou, F., Lebrilla, C.B., Michels, K.B., Palmer, J.R., Schedin, P., Stuebe, A.M., et al. (2013). Postpartum remodeling, lactation, and breast cancer risk: summary of a National Cancer Institute-sponsored workshop. J Natl Cancer Inst 105, 166-174. 10.1093/jnci/djs505.

Flores, J.E., Thompson, A.J., Ryan, M., and Howell, J. (2022). The Global Impact of Hepatitis B Vaccination on Hepatocellular Carcinoma. Vaccines (Basel) 10. 10.3390/vaccines10050793.

Fraser, G.E., Jaceldo-Siegl, K., Orlich, M., Mashchak, A., Sirirat, R., and Knutsen, S. (2020). Dairy, soy, and risk of breast cancer: those confounded milks. Int J Epidemiol 49, 1526-1537. 10.1093/ije/dyaa007.

Funk, M., Gunst, K., Lucansky, V., Muller, H., zur Hausen, H., and de Villiers, E.M. (2014). Isolation of protein-associated circular DNA from healthy cattle serum. Genome Announc 2. 10.1128/genomeA.00846-14.

Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S., Franceschi, C., Ferrucci, L., Gilroy, D.W., Fasano, A., Miller, G.W., et al. (2019). Chronic inflammation in the etiology of disease across the life span. Nat Med 25, 1822-1832. 10.1038/s41591-019-0675-0.

Gao, G.F., Parker, J.S., Reynolds, S.M., Silva, T.C., Wang, L.B., Zhou, W., Akbani, R., Bailey, M., Balu, S., Berman, B.P., et al. (2019). Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. Cell Syst 9, 24-34 e10. 10.1016/j.cels.2019.06.006.

Graham, S.V. (2017). The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review. Clin Sci (Lond) 131, 2201-2221. 10.1042/CS20160786.

Grant, B.J., Rodrigues, A.P., ElSawy, K.M., McCammon, J.A., and Caves, L.S. (2006). Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22, 2695-2696. 10.1093/bioinformatics/btl461.

Grosso, G., Bella, F., Godos, J., Sciacca, S., Del Rio, D., Ray, S., Galvano, F., and Giovannucci, E.L. (2017). Possible role of diet in cancer: systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. Nutr Rev 75, 405-419. 10.1093/nutrit/nux012.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature 550, 204-213. 10.1038/nature24277.

Gu, Z. (2022). Package "colorRamp2" - Generate Color Mapping Functions.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847-2849. 10.1093/bioinformatics/btw313.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize Implements and enhances circular visualization in R. Bioinformatics 30, 2811-2812. 10.1093/bioinformatics/btu393.

Gunst, K., zur Hausen, H., and de Villiers, E.M. (2014). Isolation of bacterial plasmid-related replication-associated circular DNA from a serum sample of a multiple sclerosis patient. Genome Announc 2. 10.1128/genomeA.00847-14.

Habermann, D., Klempt, M., and Franz, C.M.A.P. (2023). Identification and Characterization of Novel SPHINX/BMMF-like DNA Sequences Isolated from Non-Bovine Foods. Genes 14. 10.3390/genes14071307.

Häfele, L. (2020). Detection of Bovine Meat and Milk Factors in colorectal and breast cancer patients using D-VISION. Master Programme Molecular Biotechnology (Ruprecht-Karls-Universität Heidelberg).

Hansen, J.P., Ali, W.M., Sivadasan, R., and Rajeeve, K. (2021). Bacteria-Cancer Interface: Awaiting the Perfect Storm. Pathogens 10. 10.3390/pathogens10101321.

Harmon, B.E., Wirth, M.D., Boushey, C.J., Wilkens, L.R., Draluck, E., Shivappa, N., Steck, S.E., Hofseth, L., Haiman, C.A., Le Marchand, L., and Hebert, J.R. (2017). The Dietary Inflammatory Index

Is Associated with Colorectal Cancer Risk in the Multiethnic Cohort. J Nutr 147, 430-438. 10.3945/jn.116.242529.

Heredia-Torres, T.G., Rincon-Sanchez, A.R., Lozano-Sepulveda, S.A., Galan-Huerta, K., Arellanos-Soto, D., Garcia-Hernandez, M., Garza-Juarez, A.J., and Rivas-Estilla, A.M. (2022). Unraveling the Molecular Mechanisms Involved in HCV-Induced Carcinogenesis. Viruses 14. 10.3390/v14122762.

Hess, J.M., Stephensen, C.B., Kratz, M., and Bolling, B.W. (2021). Exploring the Links between Diet and Inflammation: Dairy Foods as Case Studies. Adv Nutr 12, 1S-13S. 10.1093/advances/nmab108.

Horak, P., Uhrig, S., Witzel, M., Gil-Farina, I., Hutter, B., Rath, T., Gieldon, L., Balasubramanian, G.P., Pastor, X., Heilig, C.E., et al. (2020). Comprehensive genomic characterization of gene therapy-induced T-cell acute lymphoblastic leukemia. Leukemia 34, 2785-2789. 10.1038/s41375-020-0779-z.

Horny, K. (2018). Enhancement of D-VISION procedure and analysis of viral gene expression in human breast and cervical cancer data sets. Master Biochemistry (Ruprecht-Karls-Universität Heidelberg).

Hutter, C., and Zenklusen, J.C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. Cell 173, 283-285. 10.1016/j.cell.2018.03.042.

Illah, O., and Olaitan, A. (2023). Updates on HPV Vaccination. Diagnostics (Basel) 13. 10.3390/diagnostics13020243.

Ji, J., Sundquist, J., and Sundquist, K. (2015). Lactose intolerance and risk of lung, breast and ovarian cancers: aetiological clues from a population-based study in Sweden. Br J Cancer 112, 149-152. 10.1038/bjc.2014.544.

Kassambara, A. (2023). Package "ggpubr" - "ggplot2" Based Publication Ready Plots.

Key, T.J., Bradbury, K.E., Perez-Cornago, A., Sinha, R., Tsilidis, K.K., and Tsugane, S. (2020). Diet, nutrition, and cancer risk: what do we know and what is the way forward? BMJ 368, m511. 10.1136/bmj.m511.

Kilic, T., Popov, A.N., Burk-Korner, A., Koromyslova, A., zur Hausen, H., Bund, T., and Hansman, G.S. (2019). Structural analysis of a replication protein encoded by a plasmid isolated from a multiple sclerosis patient. Acta Crystallogr D Struct Biol 75, 498-504. 10.1107/S2059798319003991.

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., Jupyter Development Team, (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows.

König, M.T., Frölich, K., Jandowsky, A., Knauf-Witzens, T., Langner, C., Dietrich, R., Märtlbauer, E., and Didier, A. (2023). First Insights into the Occurrence of Circular Single-Stranded DNA Genomes in Asian and African Cattle. Animals (Basel) 13. 10.3390/ani13091492.

König, M.T., Fux, R., Link, E., Sutter, G., Märtlbauer, E., and Didier, A. (2021a). Circular Rep-Encoding Single-Stranded DNA Sequences in Milk from Water Buffaloes (Bubalus arnee f. bubalis). Viruses 13. 10.3390/v13061088.

König, M.T., Fux, R., Link, E., Sutter, G., Märtlbauer, E., and Didier, A. (2021b). Identification and Characterization of Circular Single-Stranded DNA Genomes in Sheep and Goat Milk. Viruses 13. 10.3390/v13112176.

Kratz, M., Coats, B.R., Hisert, K.B., Hagman, D., Mutskov, V., Peris, E., Schoenfelt, K.Q., Kuzma, J.N., Larson, I., Billing, P.S., et al. (2014). Metabolic dysfunction drives a mechanistically distinct proinflammatory phenotype in adipose tissue macrophages. Cell Metab 20, 614-625. 10.1016/j.cmet.2014.08.010.

Krump, N.A., and You, J. (2018). Molecular mechanisms of viral oncogenesis in humans. Nat Rev Microbiol 16, 684-698. 10.1038/s41579-018-0064-6.

Lamberto, I., Gunst, K., Muller, H., zur Hausen, H., and de Villiers, E.M. (2014). Mycovirus-like DNA virus sequences from cattle serum and human brain and serum samples from multiple sclerosis patients. Genome Announc 2. 10.1128/genomeA.00848-14.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol 9, e1003118. 10.1371/journal.pcbi.1003118.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760. 10.1093/bioinformatics/btp324.

Li, X.M., Ganmaa, D., and Sato, A. (2003). The experience of Japan as a clue to the etiology of breast and ovarian cancers: relationship between death from both malignancies and dietary practices. Med Hypotheses 60, 268-275. 10.1016/s0306-9877(02)00385-7.

Liu, L., Nishihara, R., Qian, Z.R., Tabung, F.K., Nevo, D., Zhang, X., Song, M., Cao, Y., Mima, K., Masugi, Y., et al. (2017). Association Between Inflammatory Diet Pattern and Risk of Colorectal Carcinoma Subtypes Classified by Immune Responses to Tumor. Gastroenterology 153, 1517-1530 e1514. 10.1053/j.gastro.2017.08.045.

LoConte, N.K., Brewster, A.M., Kaur, J.S., Merrill, J.K., and Alberg, A.J. (2018). Alcohol and Cancer: A Statement of the American Society of Clinical Oncology. J Clin Oncol 36, 83-93. 10.1200/JCO.2017.76.1155.

Lowenfels, A.B., Maisonneuve, P., Cavallini, G., Ammann, R.W., Lankisch, P.G., Andersen, J.R., Dimagno, E.P., Andren-Sandberg, A., and Domellof, L. (1993). Pancreatitis and the risk of pancreatic

cancer. International Pancreatitis Study Group. N Engl J Med 328, 1433-1437. 10.1056/NEJM199305203282001.

Lu, B., and Li, M. (2014). Helicobacter pylori eradication for preventing gastric cancer. World J Gastroenterol 20, 5660-5665. 10.3748/wjg.v20.i19.5660.

Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A.R.N., Lopez, R., and Butcher, S. (2024). The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. Nucleic Acids Res 52, W521-W525. 10.1093/nar/gkae241.

Manuelidis, L. (2011). Nuclease resistant circular DNAs copurify with infectivity in scrapie and CJD. J Neurovirol 17, 131-145. 10.1007/s13365-010-0007-0.

Manuelidis, L. (2013). Infectious particles, stress, and induced prion amyloids: a unifying perspective. Virulence 4, 373-383. 10.4161/viru.24838.

Manuelidis, L. (2019). Prokaryotic SPHINX 1.8 REP protein is tissue-specific and expressed in human germline cells. J Cell Biochem 120, 6198-6208. 10.1002/jcb.27907.

Maynard, A., McCoach, C.E., Rotow, J.K., Harris, L., Haderk, F., Kerr, D.L., Yu, E.A., Schenk, E.L., Tan, W., Zee, A., et al. (2020). Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. Cell 182, 1232-1251 e1222. 10.1016/j.cell.2020.07.017.

Moorman, P.G., and Terry, P.D. (2004). Consumption of dairy products and the risk of breast cancer: a review of the literature. Am J Clin Nutr 80, 5-14. 10.1093/ajcn/80.1.5.

Morgan, M., Obenchain, V, Hester, J, Pagès, H (2024a). Package "SummarizedExperiment" - SummarizedExperiment container.

Morgan, M., Wang, J, Obenchain, V, Lang, M, Thompson, R; Turaga, N (2024b). Package "BiocParallel" - Bioconductor facilities for parallel evaluation.

Morin, C.R., Baeva, M.E., Hollenberg, M.D., and Brain, M.C. (2024). Milk and multiple sclerosis: A possible link? Mult Scler Relat Disord 83, 105477. 10.1016/j.msard.2024.105477.

Müller, K., Wickham, H (2023). tibble: Simple Data Frames.

Murphy, K. (2012). Janeway's Immunobiology, 8 Edition (Garland Science, Taylor & Francis Group, LLC).

Na, B.K., Pak, J.H., and Hong, S.J. (2020). Clonorchis sinensis and clonorchiasis. Acta Trop 203, 105309. 10.1016/j.actatropica.2019.105309.

Nayak, S.P., Sasi, M.P., Sreejayan, M.P., and Mandal, S. (2009). A case-control study of roles of diet in colorectal carcinoma in a South Indian Population. Asian Pac J Cancer Prev 10, 565-568.

Nieman, K.M., Anderson, B.D., and Cifelli, C.J. (2021). The Effects of Dairy Product and Dairy Protein Intake on Inflammation: A Systematic Review of the Literature. J Am Coll Nutr 40, 571-582. 10.1080/07315724.2020.1800532.

Nikitina, E., Alikhanyan, K., Nessling, M., Richter, K., Kaden, S., Ernst, C., Seitz, S., Chuprikova, L., Hafele, L., Gunst, K., et al. (2023a). Structural expression of bovine milk and meat factors in tissues of colorectal, lung and pancreatic cancer patients. Int J Cancer 153, 173-182. 10.1002/ijc.34374.

Nikitina, E., Alikhanyan, K., Nessling, M., Richter, K., Kaden, S., Ernst, C., Seitz, S., Chuprikova, L., Häfele, L., Gunst, K., et al. (2023b). Structural expression of bovine milk and meat factors in tissues of colorectal, lung and pancreatic cancer patients. Int J Cancer 153, 173-182. 10.1002/ijc.34374.

Nikitina, E., Burk-Korner, A., Wiesenfarth, M., Alwers, E., Heide, D., Tessmer, C., Ernst, C., Krunic, D., Schrotz-King, P., Chang-Claude, J., et al. (2023c). Bovine meat and milk factor protein expression in tumor-free mucosa of colorectal cancer patients coincides with macrophages and might interfere with patient survival. Mol Oncol. 10.1002/1878-0261.13390.

Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S. (2024). Package "Biostrings" - Efficient manipulation of biological strings.

Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N., and Sibley, R.K. (1991). Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med 325, 1127-1131. 10.1056/NEJM199110173251603.

Pattyn, J., Hendrickx, G., Vorsters, A., and Van Damme, P. (2021). Hepatitis B Vaccines. The Journal of Infectious Diseases 224, S343-S351. 10.1093/infdis/jiaa668.

Pedersen, T.L. (2024). Package "patchwork" - The Composer of Plots.

Peneau, C., Imbeaud, S., La Bella, T., Hirsch, T.Z., Caruso, S., Calderaro, J., Paradis, V., Blanc, J.F., Letouze, E., Nault, J.C., et al. (2022). Hepatitis B virus integrations promote local and distant oncogenic driver alterations in hepatocellular carcinoma. Gut 71, 616-626. 10.1136/gutjnl-2020-323153.

Plummer, M., de Martel, C., Vignat, J., Ferlay, J., Bray, F., and Franceschi, S. (2016). Global burden of cancers attributable to infections in 2012: a synthetic analysis. Lancet Glob Health 4, e609-616. 10.1016/S2214-109X(16)30143-7.

Pohl, S., Habermann, D., Link, E.K., Fux, R., Boldt, C.L., Franz, C.M.A.P., Holzel, C., and Klempt, M. (2022). Detection of DNA sequences attributed to bovine meat and milk factors (BMMF/SPHINX) in food-related samples. Food Control 135. ARTN 108779 10.1016/j.foodcont.2021.108779.

Pöschl, G., and Seitz, H.K. (2004). Alcohol and cancer. Alcohol Alcohol 39, 155-165. 10.1093/alcalc/agh057.

Praud, D., Rota, M., Rehm, J., Shield, K., Zatonski, W., Hashibe, M., La Vecchia, C., and Boffetta, P. (2016). Cancer incidence and mortality attributable to alcohol consumption. Int J Cancer 138, 1380-1387. 10.1002/ijc.29890.

Proulx, J., Ghaly, M., Park, I.W., and Borgmann, K. (2022). HIV-1-Mediated Acceleration of Oncovirus-Related Non-AIDS-Defining Cancers. Biomedicines 10. 10.3390/biomedicines10040768.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842. 10.1093/bioinformatics/btq033.

R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Ramey, C., Fox, B. (2003). GNU Bash Reference Manual (Network Theory LTD).

Rochford, R., and Moormann, A.M. (2015). Burkitt's Lymphoma. Curr Top Microbiol Immunol 390, 267-285. 10.1007/978-3-319-22822-8_11.

Rothwell, P.M., Fowkes, F.G., Belch, J.F., Ogawa, H., Warlow, C.P., and Meade, T.W. (2011). Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. Lancet 377, 31-41. 10.1016/S0140-6736(10)62110-1.

Rous, P. (1911). The Relations of Embryonic Tissue and Tumor in Mixed Grafts. J Exp Med 13, 239-247. 10.1084/jem.13.2.239.

Scheckel, C., and Aguzzi, A. (2018). Prions, prionoids and protein misfolding disorders. Nat Rev Genet 19, 405-418. 10.1038/s41576-018-0011-4.

Shannon-Lowe, C., Rickinson, A.B., and Bell, A.I. (2017). Epstein-Barr virus-associated lymphomas. Philos Trans R Soc Lond B Biol Sci 372. 10.1098/rstb.2016.0271.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12, 1611-1618. 10.1101/gr.361602.

Steck, S.E., and Murphy, E.A. (2020). Dietary patterns and cancer risk. Nat Rev Cancer 20, 125-138. 10.1038/s41568-019-0227-4.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71, 209-249. 10.3322/caac.21660.

Tabung, F.K., Liu, L., Wang, W., Fung, T.T., Wu, K., Smith-Warner, S.A., Cao, Y., Hu, F.B., Ogino, S., Fuchs, C.S., and Giovannucci, E.L. (2018). Association of Dietary Inflammatory Potential With Colorectal Cancer Risk in Men and Women. JAMA Oncol 4, 366-373. 10.1001/jamaoncol.2017.4844.

Tamura, K., Stecher, G., and Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38, 3022-3027. 10.1093/molbev/msab120.

The Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113-1120. 10.1038/ng.2764.

The GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580-585. 10.1038/ng.2653.

The GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330. 10.1126/science.aaz1776.

The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature 578, 82-93. 10.1038/s41586-020-1969-6.

Thorley-Lawson, D.A. (2015). EBV Persistence--Introducing the Virus. Curr Top Microbiol Immunol 390, 151-209. 10.1007/978-3-319-22822-8_8.

Todoric, J., Antonucci, L., and Karin, M. (2016). Targeting Inflammation in Cancer Prevention and Therapy. Cancer Prev Res (Phila) 9, 895-905. 10.1158/1940-6207.CAPR-16-0209.

Troshin, P.V., Procter, J.B., and Barton, G.J. (2011). Java bioinformatics analysis web services for multiple sequence alignment--JABAWS:MSA. Bioinformatics 27, 2001-2002. 10.1093/bioinformatics/btr304.

Troshin, P.V., Procter, J.B., Sherstnev, A., Barton, D.L., Madeira, F., and Barton, G.J. (2018). JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure. Bioinformatics 34, 1939-1940. 10.1093/bioinformatics/bty045.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell 170, 564-576 e516. 10.1016/j.cell.2017.06.010.

van der Pols, J.C., Bain, C., Gunnell, D., Smith, G.D., Frobisher, C., and Martin, R.M. (2007). Childhood dairy intake and adult cancer risk: 65-y follow-up of the Boyd Orr cohort. Am J Clin Nutr 86, 1722-1729. 10.1093/ajcn/86.5.1722.

van Rossum, G., Drake Jr, F.L. (2009). Python 3 Reference Manual (CreateSpace).

Vondran, A. (2022). Comparative computational Analysis of Bovine Meat and Milk factors in NGS data from the Genotype Tissue Expression Project. German Cancer Research Center, Department of Applied Bioinformatics.

Wang, L.G., Lam, T.T., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C.W., Jones, B.R., Bradley, T., et al. (2020). Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol Biol Evol 37, 599-603. 10.1093/molbev/msz240.

Wang, W., Chen, E., Li H. (2021). Truncated Rank-Based Tests for Two-Part Models with Excessive Zeros and Applications to Microbiome Data. arXiv.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 25, 1189-1191. 10.1093/bioinformatics/btp033.

White, M.K., Pagano, J.S., and Khalili, K. (2014). Viruses and human cancers: a long road of discovery of molecular paradigms. Clin Microbiol Rev 27, 463-481. 10.1128/CMR.00124-13.

Whitley, C., Gunst, K., Muller, H., Funk, M., zur Hausen, H., and de Villiers, E.M. (2014). Novel replication-competent circular DNA molecules from healthy cattle serum and milk and multiple sclerosis-affected human brain tissue. Genome Announc 2. 10.1128/genomeA.00849-14.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis, 2 Edition (Springer Verlag).

Wickham, H. (2023a). Package "stringr" - Simple, Consistent Wrappers for Common String Operations.

Wickham, H., Bryan, J. (2023b). Package "readxl" - Read Excel Files.

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D. (2023c). dplyr: A Grammar of Data Manipulation.

World Cancer Research Fund/American Institute for Cancer Research (2018). Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. http://www.dietandcancerreport.org/.

Wu, Y., Huang, R., Wang, M., Bernstein, L., Bethea, T.N., Chen, C., Chen, Y., Eliassen, A.H., Freedman, N.D., Gaudet, M.M., et al. (2021). Dairy foods, calcium, and risk of breast cancer overall and for subtypes defined by estrogen receptor status: a pooled analysis of 21 cohort studies. Am J Clin Nutr 114, 450-461. 10.1093/ajcn/nqab097.

Xu, S., Chen, M., Feng, T., Zhan, L., Zhou, L., and Yu, G. (2021). Use ggbreak to Effectively Utilize Plotting Space to Deal With Large Datasets and Outliers. Front Genet 12, 774846. 10.3389/fgene.2021.774846.

Yu, G. (2023). Data integration, manipulation and visualization of phylogenetic trees, 1st Edition (Chapman & Hall/CRC Press).

Zerr, I., Ladogana, A., Mead, S., Hermann, P., Forloni, G., and Appleby, B.S. (2024). Creutzfeldt-Jakob disease and other prion diseases. Nat Rev Dis Primers 10, 14. 10.1038/s41572-024-00497-y.

zur Hausen, H. (2001). Oncogenic DNA viruses. Oncogene 20, 7820 - 7823.

zur Hausen, H. (2002). Papillomaviruses and cancer: from basic studies to clinical application. Nat Rev Cancer 2, 342-350. 10.1038/nrc798.

zur Hausen, H. (2012). Red meat consumption and cancer: reasons to suspect involvement of bovine infectious factors in colorectal cancer. Int J Cancer 130, 2475-2483. 10.1002/ijc.27413.

zur Hausen, H. (2015). Risk factors: What do breast and CRC cancers and MS have in common? Nat Rev Clin Oncol 12, 569-570. 10.1038/nrclinonc.2015.154.

zur Hausen, H., Bund, T., and de Villiers, E.M. (2019). Specific nutritional infections early in life as risk factors for human colon and breast cancers several decades later. Int J Cancer 144, 1574-1583. 10.1002/ijc.31882.

zur Hausen, H., and de Villiers, E.M. (2014). Cancer "causation" by infections--individual contributions and synergistic networks. Semin Oncol 41, 860-875. 10.1053/j.seminoncol.2014.10.003.

zur Hausen, H., and de Villiers, E.M. (2015a). Dairy cattle serum and milk factors contributing to the risk of colon and breast cancers. Int J Cancer 137, 959-967. 10.1002/ijc.29466.

zur Hausen, H., and de Villiers, E.M. (2015b). Reprint of: cancer "causation" by infections--individual contributions and synergistic networks. Semin Oncol 42, 207-222. 10.1053/j.seminoncol.2015.02.019.

zur Hausen, H.B., T.; de Villiers, E.M. (2017a). Infectious Agents in Bovine Red Meat and Milk and Their Potential Role in Cancer and Other Chronic Diseases. Current Topics in Microbiology and Immunology 407, 83-116.

zur Hausen, H.B., T.; de Villiers, E.M. (2017b). Infectious Agents in Bovine Red Meat and Milk and Their Potential Role in Cancer and Other Chronic Diseases. In Viruses, Genes, and Cancer, K.B. Eric Hunter, ed. (Springer International Publishing AG), pp. 83-116.

6. Supplementary Materials

S1: BMMF library

GENOME	BMMF GROUP	LENGTH	ORIGIN OF SEQUENCE	GENEBANK IDENTIFIER
4ABAYE	2	2726	Acinetobacter baumannii str. AYE plasmid p4ABAYE	CU459139.1
Ac.Bau.DS002.16S.rRNA	-	1496	Acinetobacter sp. DS002 16S ribosomal RNA gene, partial sequence	JX519565.1
C1HB.3	1	1086	cattle group1 healthy bovine isolate 3	LK931495.1
C1HB.4	1	2958	cattle group1 healthy bovine isolate 4	LK931496.1
C1HB.5	1	1723	cattle group1 healthy bovine isolate 5	LK931497.1
C1HB.6.1	1	2522	cattle group1 healthy bovine isolate 6.1	LK931493.1
C1HB.6.2	1	1591	cattle group1 healthy bovine isolate 6.2	LK931494.1
C1MI.1	1	2523	cattle group 1 milk isolate 1	LK931487.1
C1MI.15M.1	1	2040	cattle group 1 milk isolate, sample 15, primer M, isolate 1	LR215494.1
C1MI.15M.2	1	2041	cattle group 1 milk isolate, sample 15, primer M, isolate 1	LR215495.1
C1MI.2	1	2148	cattle group 1 milk isolate 2	LK931488.1
C1MI.3	1	1687	cattle group 1 milk isolate 3	LK931489.1
C1MI.3M.1	1	1767	cattle group 1 milk isolate, sample 3, primer M, isolate 1	LR215499.1
C1MI.4	1	1583	cattle group 1 milk isolate 4	LK931490.1
C1MI.5.1	1	1706	cattle group 1 milk isolate 5.1	LT715554.1
C1MI.5.2	1	2406	cattle group 1 milk isolate 5.2	LT715555.1
C1MI.9M.1	1	1935	cattle group 1 milk isolate, sample 9, primer M, isolate 1	LR215496.1
C1MI.9M.2	1	1934	cattle group 1 milk isolate, sample 9, primer M, isolate 1	LR215497.1
C1MIs.3M.1	1	461	cattle group 1 milk isolate, sample 3, primer M, small isolate 1	LR215498.1
C2HB1	2	2251	cattle group2 healthy bovine isolate 1	LK931499.1
C2HB2	2	1407	cattle group2 healthy bovine isolate 2	LK931500.1
C2HB7	2	2280	cattle group2 healthy bovine isolate 7	LK931498.1
C2MI.10A.1	2	2505	cattle group 2 milk isolate, sample 10, primer A, isolate 1	LR215592.1
C2MI.10A.2	2	2504	cattle group 2 milk isolate, sample 10, primer A, isolate 2	LR215593.1
C2MI.10As.1	2	697	cattle group 2 milk isolate, sample 10, primer A, small isolate 1	LR215597.1
C2MI.13B.1	2	2301	cattle group 2 milk isolate, sample 13, primer B, isolate 1	LR215594.1
C2MI.13B.2	2	2300	cattle group 2 milk isolate, sample 13, primer B, isolate 2	LR215595.1
C2MI.13B.3	2	2301	cattle group 2 milk isolate, sample 13, primer B, isolate 3	LR215596.1
C2MI.15B.1	2	2275	cattle group 2 milk isolate, sample 15, primer B, isolate 1	LR215553.1
C2MI.15B.10	2	2590	cattle group 2 milk isolate, sample 15, primer B, isolate 10	LR215562.1
C2MI.15B.11	2	2362	cattle group 2 milk isolate, sample 15, primer B, isolate 11	LR215563.1
C2MI.15B.12	2	2362	cattle group 2 milk isolate, sample 15, primer B, isolate 12	LR215564.1
C2MI.15B.13	2	2278	cattle group 2 milk isolate, sample 15, primer B, isolate 13	LR215565.1
C2MI.15B.14	2	2776	cattle group 2 milk isolate, sample 15, primer B, isolate 14	LR215566.1
C2MI.15B.15	2	2775	cattle group 2 milk isolate, sample 15, primer B, isolate 15	LR215567.1

C2MI.15B.16	2	2824	cattle group 2 milk isolate, sample 15, primer B, isolate 16	LR215568.1
C2MI.15B.17	2	2952	cattle group 2 milk isolate, sample 15, primer B, isolate 17	LR215569.1
C2MI.15B.18	2	2778	cattle group 2 milk isolate, sample 15, primer B, isolate 18	LR215570.1
C2MI.15B.2	2	2312	cattle group 2 milk isolate, sample 15, primer B, isolate 2	LR215554.1
C2MI.15B.3	2	2362	cattle group 2 milk isolate, sample 15, primer B, isolate 3	LR215555.1
C2MI.15B.4	2	2362	cattle group 2 milk isolate, sample 15, primer B, isolate 4	LR215556.1
C2MI.15B.5	2	2279	cattle group 2 milk isolate, sample 15, primer B, isolate 5	LR215557.1
C2MI.15B.6	2	2279	cattle group 2 milk isolate, sample 15, primer B, isolate 6	LR215558.1
C2MI.15B.7	2	2279	cattle group 2 milk isolate, sample 15, primer B, isolate 7	LR215559.1
C2MI.15B.8	2	2277	cattle group 2 milk isolate, sample 15, primer B, isolate 8	LR215560.1
C2MI.15B.9	2	2590	cattle group 2 milk isolate, sample 15, primer B, isolate 9	LR215561.1
C2MI.16B.1	2	2234	cattle group 2 milk isolate, sample 16, primer B, isolate 1	LR215571.1
C2MI.16B.10	2	2234	cattle group 2 milk isolate, sample 16, primer B, isolate 10	LR215580.1
C2MI.16B.11	2	2782	cattle group 2 milk isolate, sample 16, primer B, isolate 11	LR215581.1
C2MI.16B.12	2	2479	cattle group 2 milk isolate, sample 16, primer B, isolate 12	LR215582.1
C2MI.16B.2	2	2233	cattle group 2 milk isolate, sample 16, primer B, isolate 2	LR215572.1
C2MI.16B.3	2	2478	cattle group 2 milk isolate, sample 16, primer B, isolate 3	LR215573.1
C2MI.16B.4	2	2478	cattle group 2 milk isolate, sample 16, primer B, isolate 4	LR215574.1
C2MI.16B.5	2	2428	cattle group 2 milk isolate, sample 16, primer B, isolate 5	LR215575.1
C2MI.16B.6	2	2478	cattle group 2 milk isolate, sample 16, primer B, isolate 6	LR215576.1
C2MI.16B.7	2	2566	cattle group 2 milk isolate, sample 16, primer B, isolate 7	LR215577.1
C2MI16B.8	2	2234	cattle group 2 milk isolate, sample 16, primer B, isolate 8	LR215578.1
C2MI.16B.9	2	2777	cattle group 2 milk isolate, sample 16, primer B, isolate 9	LR215579.1
C2MI.1A.1	2	2293	cattle group 2 milk isolate, sample 1, primer A, isolate 1	LR215583.1
C2MI.1A.2	2	2296	cattle group 2 milk isolate, sample 1, primer A, isolate 2	LR215584.1
C2MI.1A.3	2	2294	cattle group 2 milk isolate, sample 1, primer A, isolate 3	LR215585.1
C2MI.1A.4	2	2296	cattle group 2 milk isolate, sample 1, primer A, isolate 4	LR215586.1
C2MI.3A.1	2	2356	cattle group 2 milk isolate, sample 3, primer A, isolate 1	LR215587.1
C2MI.3A.2	2	2356	cattle group 2 milk isolate, sample 3, primer A, isolate 2	LR215588.1
C2MI.4A.1	2	2661	cattle group 2 milk isolate, sample 4, primer A, isolate 1	LR215589.1
C2MI.4A.2	2	2661	cattle group 2 milk isolate, sample 4, primer A, isolate 2	LR215590.1
C2MI.4B.3	2	2661	cattle group 2 milk isolate, sample 4, primer B, isolate 3	LR215591.1

C2MI.5A.1	2	2257	cattle group 2 milk isolate, sample 5, primer A, isolate 1	LR215500.1
C2MI.5A.2	2	2363	cattle group 2 milk isolate, sample 5, primer A, isolate 2	LR215501.1
C2MI.5A.3	2	2257	cattle group 2 milk isolate, sample 5, primer A, isolate 3	LR215502.1
C2MI.5A.4	2	2363	cattle group 2 milk isolate, sample 5, primer A, isolate 4	LR215503.1
C2MI.5As.1	2	697	cattle group 2 milk isolate, sample 5, primer A, small isolate 1	LR215598.1
C2MI.5B.10	2	2736	cattle group 2 milk isolate, sample 5, primer B, isolate 10	LR215509.1
C2MI.5B.11	2	2735	cattle group 2 milk isolate, sample 5, primer B, isolate 11	LR215510.1
C2MI.5B.12	2	2567	cattle group 2 milk isolate, sample 5, primer B, isolate 12	LR215511.1
C2MI.5B.13	2	2392	cattle group 2 milk isolate, sample 5, primer B, isolate 13	LR215512.1
C2MI.5B.5	2	2257	cattle group 2 milk isolate, sample 5, primer B, isolate 5	LR215504.1
C2MI.5B.6	2	2376	cattle group 2 milk isolate, sample 5, primer B, isolate 6	LR215505.1
C2MI.5B.7	2	2375	cattle group 2 milk isolate, sample 5, primer B, isolate 7	LR215506.1
C2MI.5B.8	2	2257	cattle group 2 milk isolate, sample 5, primer B, isolate 8	LR215507.1
C2MI.5B.9	2	2736	cattle group 2 milk isolate, sample 5, primer B, isolate 9	LR215508.1
C2MI.7A.1	2	2103	cattle group 2 milk isolate, sample 7, primer A, isolate 1	LR215513.1
C2MI.7A.10	2	2460	cattle group 2 milk isolate, sample 7, primer A, isolate 10	LR215522.1
C2MI.7A.2	2	2102	cattle group 2 milk isolate, sample 7, primer A, isolate 2	LR215514.1
C2MI.7A.3	2	2102	cattle group 2 milk isolate, sample 7, primer A, isolate 3	R215515.1
C2MI.7A.4	2	2313	cattle group 2 milk isolate, sample 7, primer A, isolate 4	LR215516.1
C2MI.7A.5	2	2315	cattle group 2 milk isolate, sample 7, primer A, isolate 5	LR215517.1
C2MI.7A.6	2	2314	cattle group 2 milk isolate, sample 7, primer A, isolate 6	LR215518.1
C2MI.7A.7	2	2101	cattle group 2 milk isolate, sample 7, primer A, isolate 7	LR215519.1
C2MI.7A.8	2	2460	cattle group 2 milk isolate, sample 7, primer A, isolate 8	LR215520.1
C2MI.7A.9	2	2460	cattle group 2 milk isolate, sample 7, primer A, isolate 9	LR215521.1
C2MI.7As.1	2	697	cattle group 2 milk isolate, sample 7, primer A, small isolate 1	LR215599.1
C2MI.7B.11	2	2564	cattle group 2 milk isolate, sample 7, primer B, isolate 11	LR215523.1
C2MI.7B.12	2	2102	cattle group 2 milk isolate, sample 7, primer B, isolate 12	LR215524.1
C2MI.7B.13	2	2565	cattle group 2 milk isolate, sample 7, primer B, isolate 13	LR215525.1
C2MI.7B.14	2	2102	cattle group 2 milk isolate, sample 7, primer B, isolate 14	LR215526.1
C2MI.7B.15	2	2563	cattle group 2 milk isolate, sample 7, primer B, isolate 15	LR215527.1
C2MI.7B.16	2	2486	cattle group 2 milk isolate, sample 7, primer B, isolate 16	LR215528.1
C2MI.7B.17	2	2564	cattle group 2 milk isolate, sample 7, primer B, isolate 17	LR215529.1

C2MI.7B.18	2	2565	cattle group 2 milk isolate, sample 7, primer B, isolate 18	LR215530.1
C2MI.8A.1	2	3090	cattle group 2 milk isolate, sample 8, primer A, isolate 1	LR215531.1
C2MI.8A.2	2	3090	cattle group 2 milk isolate, sample 8, primer A, isolate 2	LR215532.1
C2MI.8A.3	2	3090	cattle group 2 milk isolate, sample 8, primer A, isolate 3	LR215533.1
C2MI.8B.4	2	2850	cattle group 2 milk isolate, sample 8, primer B, isolate 4	LR215534.1
C2MI.8B.5	2	3090	cattle group 2 milk isolate, sample 8, primer B, isolate 5	LR215535.1
C2MI.8B.6	2	3090	cattle group 2 milk isolate, sample 8, primer B, isolate 6	LR215536.1
C2MI.8B.7	2	2736	cattle group 2 milk isolate, sample 8, primer B, isolate 7	LR215537.1
C2MI.9A.1	2	2405	cattle group 2 milk isolate, sample 9, primer A, isolate 1	LR215538.1
C2MI.9A.2	2	2357	cattle group 2 milk isolate, sample 9, primer A, isolate 2	LR215539.1
C2MI.9A.3	2	2365	cattle group 2 milk isolate, sample 9, primer A, isolate 3	LR215540.1
C2MI.9As.2	2	697	cattle group 2 milk isolate, sample 9, primer A, small isolate 2	LR215600.1
C2MI.9B.10	2	2832	cattle group 2 milk isolate, sample 9, primer B, isolate 10	LR215547.1
C2MI.9B.11	2	2537	cattle group 2 milk isolate, sample 9, primer B, isolate 11	LR215548.1
C2MI.9B.12	2	2367	cattle group 2 milk isolate, sample 9, primer B, isolate 12	LR215549.1
C2MI.9B.13	2	2366	cattle group 2 milk isolate, sample 9, primer B, isolate 13	LR215550.1
C2MI.9B.14	2	2593	cattle group 2 milk isolate, sample 9, primer B, isolate 14	LR215551.1
C2MI.9B.15	2	2554	cattle group 2 milk isolate, sample 9, primer B, isolate 15	LR215552.1
C2MI.9B.4	2	2278	cattle group 2 milk isolate, sample 9, primer B, isolate 4	LR215541.1
C2MI.9B.5	2	2486	cattle group 2 milk isolate, sample 9, primer B, isolate 5	LR215542.1
C2MI.9B.6	2	2301	cattle group 2 milk isolate, sample 9, primer B, isolate 6	LR215543.1
C2MI.9B.7	2	2279	cattle group 2 milk isolate, sample 9, primer B, isolate 7	LR215544.1
C2MI.9B.8	2	2279	cattle group 2 milk isolate, sample 9, primer B, isolate 8	LR215545.1
C2MI.9B.9	2	2279	cattle group 2 milk isolate, sample 9, primer B, isolate 9	LR215546.1
C2MI.9Bs.4	2	671	cattle group 2 milk isolate, sample 9, primer B, small isolate 4	-
C2MI.9Bs.6	2	694	cattle group 2 milk isolate, sample 9, primer B, small isolate 6	-
coD08N.LD.Nn.15.12	1	1767	Colorectal tissue isolate	-
coD08N.LD.Nn.15.4	1	1767	Colorectal tissue isolate	-
coD08N.LD.Nn.15.5	1	1767	Colorectal tissue isolate	-
coD08N.LD.Nn.15.8	1	1768	Colorectal tissue isolate	-
coD08N.LD.Nn.15.9	1	1767	Colorectal tissue isolate	-
coD09N.LD.Nn.17.1	1	1767	Colorectal tissue isolate	-
coD09N.LD.No.17.1	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.10	1	1765	Colorectal tissue isolate	-
coD09N.LD.No.17.11	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.12	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.2	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.3	1	1767	Colorectal tissue isolate	-

coD09N.LD.No.17.4	1	1765	Colorectal tissue isolate	-
coD09N.LD.No.17.5	1	1765	Colorectal tissue isolate	-
coD09N.LD.No.17.6	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.7	1	1764	Colorectal tissue isolate	-
coD09N.LD.No.17.8	1	1766	Colorectal tissue isolate	-
coD09N.LD.No.17.9	1	1766	Colorectal tissue isolate	-
coD99N.LD.No.15.10	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.15.12	1	1767	Colorectal tissue isolate	-
coD99N.LD.No.15.2	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.15.3	1	1764	Colorectal tissue isolate	-
coD99N.LD.No.15.4	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.15.5	1	1766	Colorectal tissue isolate	-
coD99N.LD.No.15.6	1	1766	Colorectal tissue isolate	-
coD99N.LD.No.15.7	1	1762	Colorectal tissue isolate	-
coD99N.LD.No.15.8	1	1767	Colorectal tissue isolate	-
coD99N.LD.No.17.10	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.17.11	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.17.2	1	1766	Colorectal tissue isolate	-
coD99N.LD.No.17.3	1	1767	Colorectal tissue isolate	-
coD99N.LD.No.17.4	1	1764	Colorectal tissue isolate	-
coD99N.LD.No.17.5	1	1767	Colorectal tissue isolate	-
coD99N.LD.No.17.7	1	1766	Colorectal tissue isolate	-
coD99N.LD.No.17.8	1	1765	Colorectal tissue isolate	-
coD99N.LD.No.17.9	1	1765	Colorectal tissue isolate	-
H1MSB.1	1	1766	human group 1 multiple sclerosis brain isolate 1	LK931491.1
H1MSB.2	1	1766	human group 1 multiple sclerosis brain isolate 2	LK931492.1
HCBI8.215	3	2152	healthy cattle blood isolate 8 (2.15 kb)	NC_024689.1
HCBI9.212	3	2121	healthy cattle blood isolate 9 (2.12 kb)	LK931484.1
MSSI1.162	4	1627	multiple sclerosis serum isolate 1 (1.62 kb)	LK931486.1
MSSI2.225	3	2259	multiple sclerosis serum isolate 2 (2.25 kb)	LK931485.1
pA85	2	2726	Acinetobacter baumannii strain A85 plasmid pA85-1	CP021783.1
pHD4	1	1881	Uncultured bacterium plasmid clone	KX838913.1
pRGRH0103	2	2309	Plasmid, uncultured prokaryote from rat gut metagenome	LN852793.1
pRGRH0636	2	2579	Plasmid, uncultured prokaryote from rat gut metagenome	LN853262.1
pTS236	2	2252	Acinetobacter baumannii strain DS002 plasmid	JN872565.1
Sphinx1.76	1	1758	Slow Progressive Hidden INfections of variable (X) 1.76 kb	HQ444404.1
Sphinx2.36	2	2364	Slow Progressive Hidden INfections of variable (X) 2.36 kb	HQ444405.1

S2: Overview PCAWG RNA samples: List of cancer types and number of samples analyzed

Cohort	Cancer type	Number of samples
BLCA-US	Bladder Urothelial Cancer - TGCA (US)	27
BRCA-US	Breat Cancer – TCGA (US)	97
CESC-US	Cervical Squamous Cell Carcinoma – TCGA (US)	20
CLLE-ES	Chronic Lymphocytic Leukemia (Spain)	31
COAD-US	Colon Adenocarcinoma – TCGA (US)	44

DLBC-US	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma – TCGA (US)	7
GBM-US	Brain Glioblastoma Multiforme – TCGA (US)	29
HNSC-US	Head and Neck Squamous Cell Carcinoma – TCGA (US)	44
KICH-US	Kidney Chromophobe – TCGA (US)	64
KIRC-US	Kidney Renal Clear Cell Carcinoma – TCGA (US)	47
KIRP-US	Kidney Renal Papillary Cell Carcinoma – TCGA (US)	34
LAML-US	Acute Myeloid Leukemia – TCGA (US)	20
LGG-US	Brain Lower Grade Glioma – TCGA (US)	18
LIHC-US	Liver Hepatocellular carcinoma – TCGA (US)	56
LUAD-US	Lung Adenocarcinoma – TCGA (US)	48
LUSC-US	Lung Squamous Cell Carcinoma – TCGA (US)	50
MALY-DE	Malignant Lymphoma (GER)	95
OV-US	Ovarian Serous Cystadenocarcinoma – TCGA (US)	21
PRAD-US	Prostate Adenocarcinoma – TCGA (US)	21
READ-US	Rectum Adenocarcinoma – TCGA (US)	15
SARC-US	Sarcoma – TCGA (US)	33
SKCM-US	Skin Cutaneous Melanoma – TCGA (US)	34
STAD-US	Gastric Adenocarcinoma – TCGA (US)	31
THCA-US	Head and Neck Thyroid Carcinoma – TCGA (US)	52
UCEC-US	Uterine Corpus Endometrial Carcinoma- TCGA (US)	49

S3: Overview PCAWG WGS samples: List of cancer types and number of samples analyzed for both tumor and normal tissue data. The table lists number of donors, tumor samples downloaded and tumor samples analyzed. For most cancer cohorts within the PCAWG WGS data set, there is one tumor and one normal tissue files available for each donor. There are some exceptions to this, e.g. the UK prostate cancer cohort (PRAD-UK). The separation of tumor samples downloaded and tumor samples analyzed is due to problems with sample files of the Japanese liver cancer cohort LIR-JP, where several samples could not be successfully analyzed in repeated attempts.

Cohort	Cancer Type	Country/Region	Total Number of PCAWG Donors	Tumour samples downloaded	Tumour Samples analyzed	Normal tissue samples downloaded	Normal tissue samples analysed	Number of Tumour Donors analysed	Number of Normal Tissue Donors analysed	Total Samples analysed
BRCA-EU	Breast ER+ and HER2-Cancer	EU/ UK	79	79	79	79	79	79	79	158

BRCA-UK	Breast Triple Negative Cancer/ Lobular Cancer	UK	46	45	45	45	45	45	45	90
BRCA-US	Breat Cancer - TCGA	US	92	92	92	92	92	92	92	184
COAD-US	Colon Adenocarcinoma - TCGA	US	46	46	46	46	46	46	46	92
EOPC-DE	Early Onset Prostate Cancer	GER	44	70	70	43	43	43	43	113
ESAD-UK	Esophageal Adenocarcinoma	UK	100	99	99	99	99	99	99	198
HNSC-US	Head and Neck Squamous Cell Carcinoma - TCGA	US	44	44	44	44	44	44	44	88
KICH-US	Kidney Chromophobe - TCGA	US	49	49	49	49	49	49	49	98
KIRC-US	Kidney Renal Clear Cell Carcinoma - TCGA	US	40	40	40	40	40	40	40	80
KIRP-US	Kidney Renal Papillary Cell Carcinoma - TCGA	US	34	34	34	34	34	34	34	68
LAML-KR	Acute Myeloid Leukemia	South Korea	10	9	9	9	9	9	9	18
LAML-US	Acute Myeloid Leukemia - TCGA	US	33	33	33	33	33	33	33	66
LICA-FR	Liver Cancer	FRA	6	6	6	6	6	6	6	12
LIHC-US	Liver Hepatocellular carcinoma - TCGA	US	54	54	54	54	54	54	54	108
LINC-JP	Liver Cancer - NCC	JPN	31	31	31	31	31	31	31	62
LIRI-JP	Liver Cancer - RIKEN	JPN	257	265	204	255	254	255	255	458
LUAD-US	Lung Adenocarcinoma - TCGA	US	42	42	42	42	42	42	42	84
LUSC-US	Lung Squamous Cell Carcinoma - TCGA	US	48	48	48	48	48	48	48	96
MALY-DE	Malignant Lymphoma	GER	101	101	101	101	101	101	101	202
OV-AU	Ovarian Cancer	AUS	73	73	73	73	73	73	73	146
OV-US	Ovarian Serous Cystadenocarcinoma - TCGA	US	45	45	45	45	45	45	45	90
PACA-AU	Pancreatic Cancer Endocrine Neoplasms	AUS	97	96	96	96	96	96	96	192
PACA-CA	Pancreatic Cancer	CAN	147	149	149	147	147	147	147	296
PRAD-CA	Prostate Adenocarcinoma	CAN	124	122	122	122	122	122	122	244
PRAD-UK	Prostate Adenocarcinoma	UK	33	83	83	33	33	33	33	116
PRAD-US	Prostate Adenocarcinoma - TCGA	US	20	20	20	20	20	20	20	40

READ-US	Rectum Adenocarcinoma - TCGA	US	16	16	16	16	16	16	16	32
STAD-US	Gastric Adenocarcinoma - TCGA	US	39	39	39	39	39	39	39	78
UCEC-US	Uterine Corpus Endometrial Carcinoma- TCGA	US	51	51	51	51	51	51	51	102
Total			1801	1881	1820	1792	1791	1792	1792	3611

S4: Overview TCGA samples: List of cancer types and number of samples analyzed

Cancer cohort	Number of samples
TCGA-BRCA	1222
TCGA-COAD	546
TCGA-LIHC	424
TCGA-LUAD	595
TCGA-PAAD	182

S5: Overview GTEx samples: List of tissue cohorts and number of samples analyzed

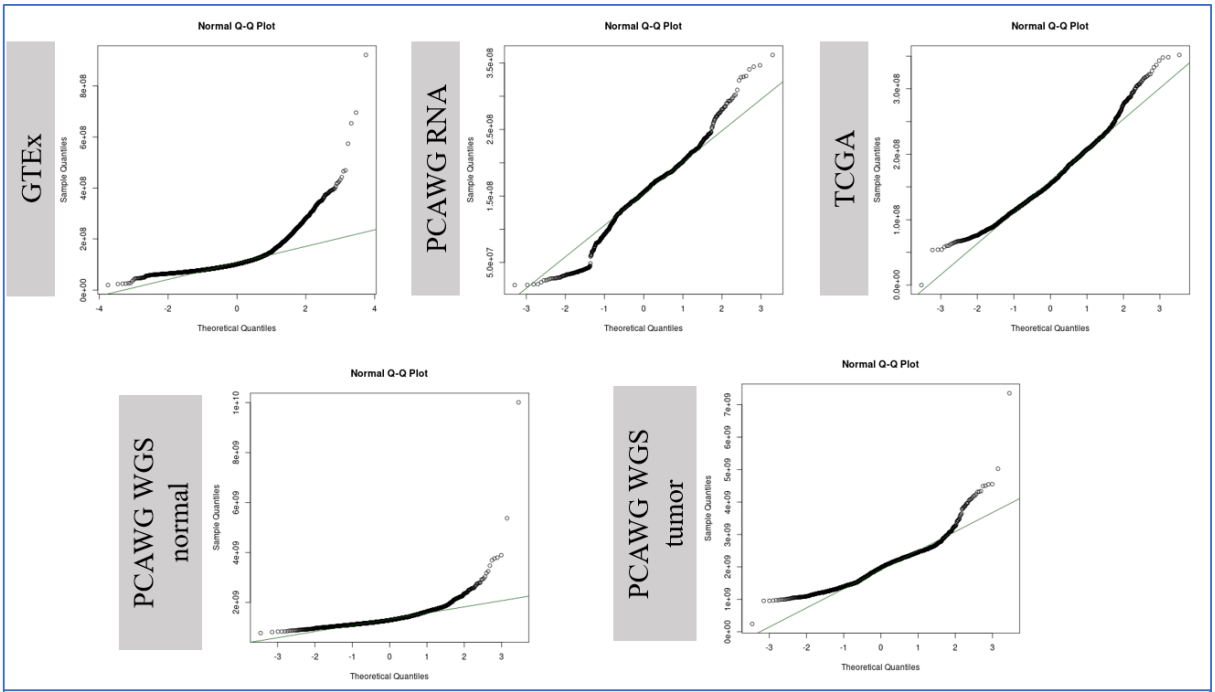
Tissue	Tissue cohort	Number of samples
Blood	Whole blood	755
Breast	Breast - mammary	457
Colon	Colon - sigmoid	373
Colon	Colon - transverse	406
Esophagus	Esophagus – gastroesophageal - junction	374
Esophagus	Esophagus - mucosa	549
Esophagus	Esophagus - muscularis	509
Kidney	Kidney	89
Liver	Liver	226
Lung	Lung	573
Ovary	Ovary	180
Pancreas	Pancreas	328

Prostate	Prostate	241
Stomach	Stomach	359
Uterus	Uterus	142

S6: Overview DepMap cell lines: List of cancer sites and number of analyzed cancer cell lines

Cancer cite	Number of cell lines
aerodigestive	5
bladder	6
Bone	1
Breast	30
Cervix	1
esophagus	8
gastric	17
glioma	15
kidney	8
leukemia	5
liver	9
lung	84
lymphoma	3
myeloma	4
neuroblastoma	2
ovary	20
pancreas	11
prostate	3
rhaboid	3
sarcoma	2
skin	17
thyroid	2

S7: Q-Q plots for sequencing depth of RNA-seq and WGS data sets

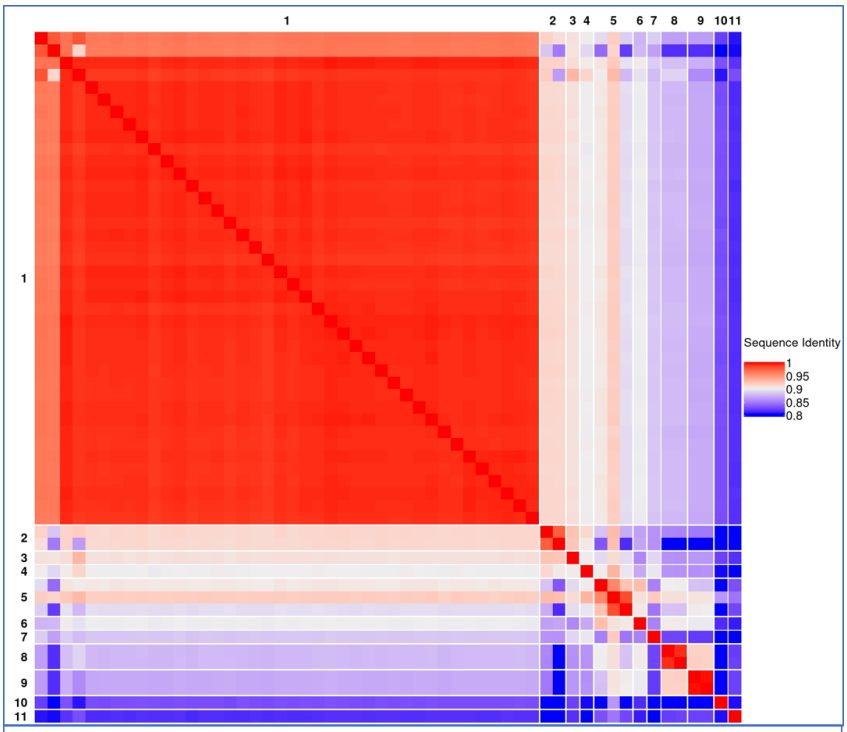


S7: Q-Q plots for sequencing depth of RNA-seq and WGS data sets: The Q-Q plots for the sequencing depth were generated for the PCAWG, TCGA and GTEx RNA seq data and for the PCAWG tumor and normal tissue/blood WGS data to verify, if the sequencing depths are normally distributed. The green line indicates the expected outcome of the Q-Q plot in case of normally distributed data. The sequencing depths were determined using samtools, the Q-Q plots were generated using R.

S8: Percentage of BMMF hits distribution to the four BMMF groups: The percentage of BMMF1, BMMF2, BMMF3 and BMMF4 hits of the BMMF reads detected in the three RNA sequencing and the PCAWG WGS data sets.

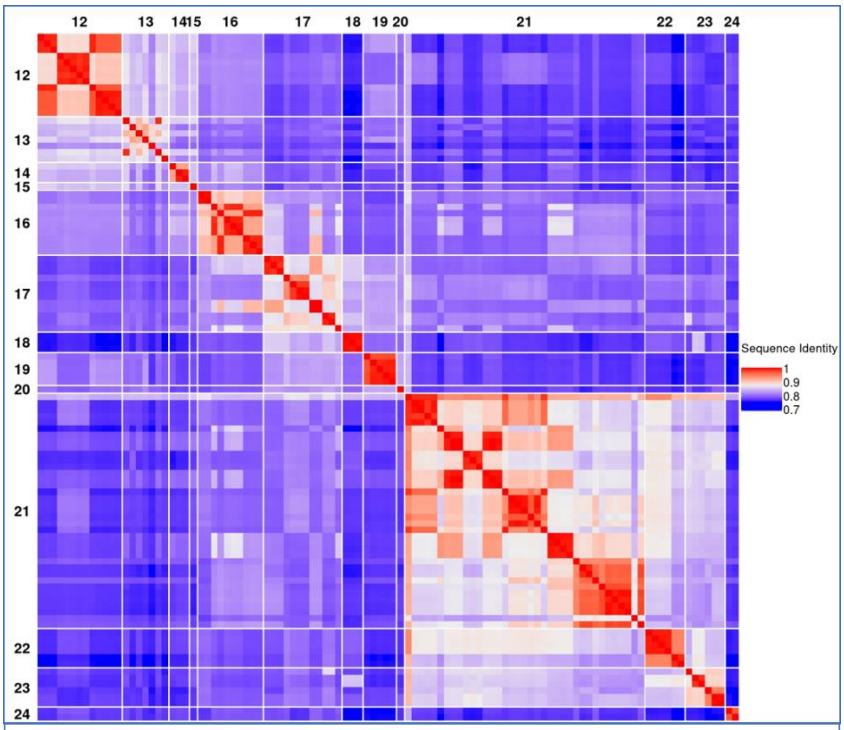
BMMF Group	TCGA RNA	PCAWG RNA	GTEx RNA	PCAWG WGS - Tumor	PCAWG WGS - Normal	DepMap	Total
BMMF1	90.16	83.03	78.67	83.50	89.51	0.00	84.23
BMMF2	9.84	16.97	19.04	13.23	9.72	0.00	13.41
BMMF3	0.00	0.00	0.78	0.04	0.00	0.00	0.12
BMMF4	0.00	0.00	1.50	3.23	0.77	0.00	2.24
Total	100.00	100.00	100.00	100.00	100.00	0.00	100.00

S9: Percent identity matrix for BMMF1



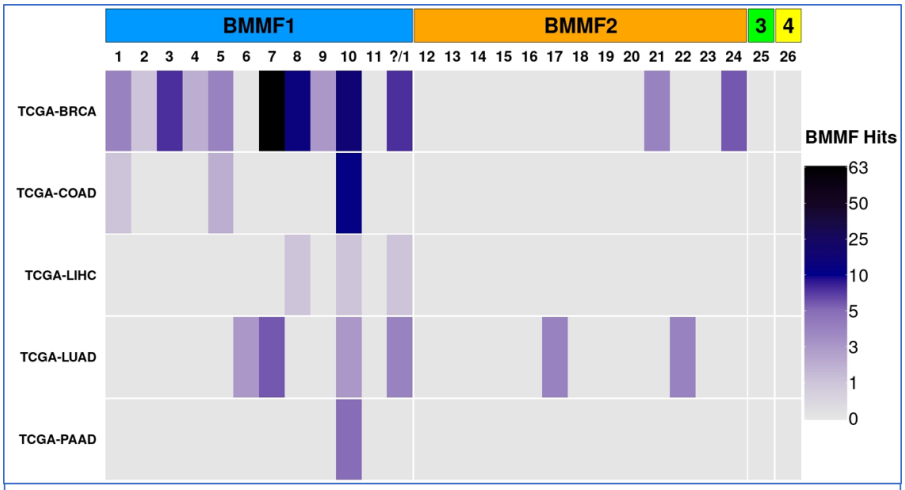
S9: Percent identity matrix for BMMF1: 11 BMMF 1 subgroups were determined based on the timetree of BMMF group 1 genomes calculated with MEGA 11. The small BMMF1 isolates C1MIs.3M.1 and C1HB.6.2 were excluded from the phylogenetic analysis due to the different genome size. The percent identity matrix was created using MEGA 11, bio3d and ComplexHeatmap.

S10: Percent identity matrix for BMMF



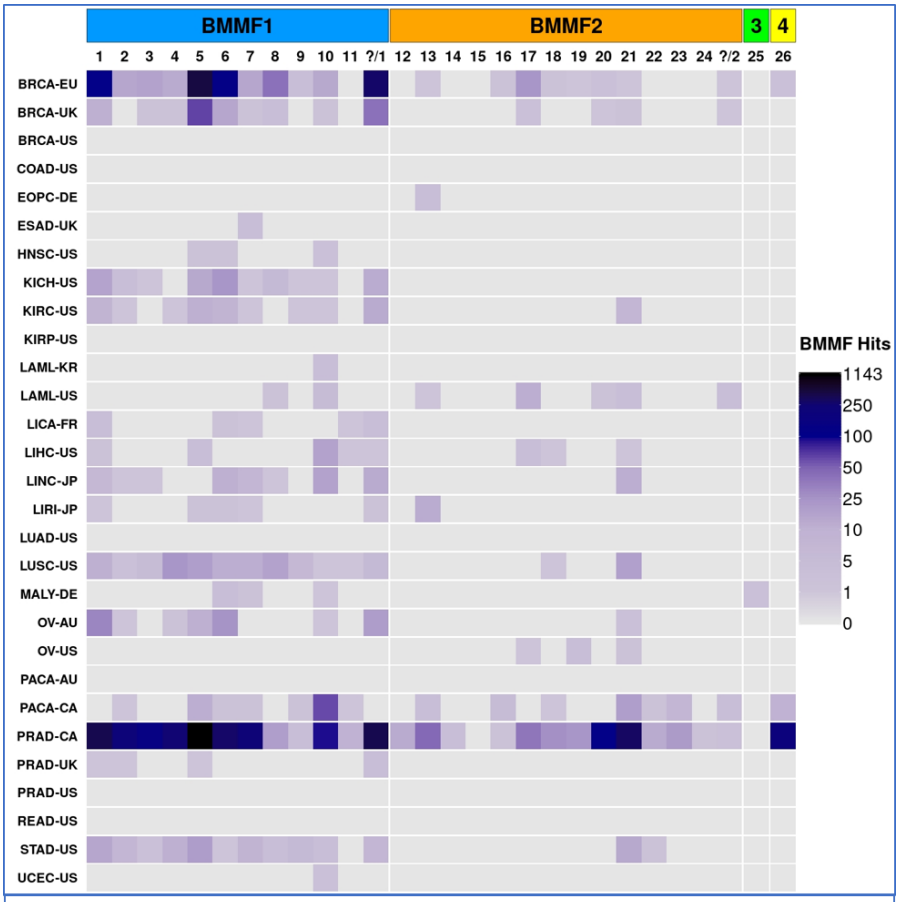
S10: Percent identity matrix for BMMF2: 13 BMMF 2 subgroups were determined based on the timetree of BMMF group 2 genomes calculated with MEGA 11. The small BMMF1 isolates C2MI.10As.1, C2MI.5As.1, C2MI.7As.1, C2MI.9As.2, C2MI.9Bs.4 and C2MI.9Bs.6 were excluded from the phylogenetic analysis due to the different genome size. The percent identity matrix was created using MEGA 11, bio3d and ComplexHeatmap.

S11: BMMF subgroups detected in all cancer cohorts of TCGA RNA data



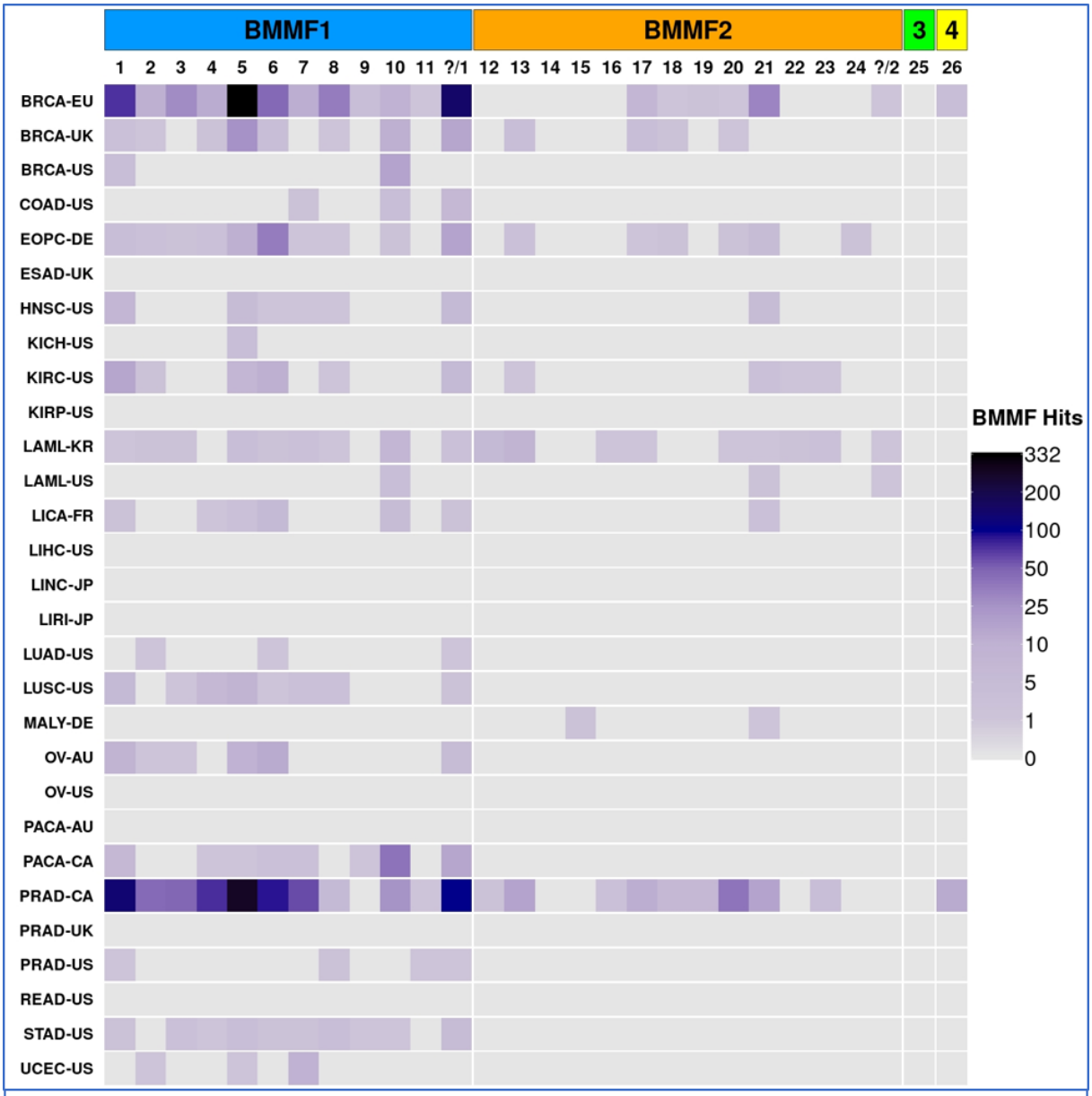
S11: BMMF subgroups detected in all cancer cohorts of TCGA RNA data: Overview of BMMF reads detected in TCGA RNA data cohorts after application of the three-hits threshold. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S12: BMMF subgroups detected in all cancer cohorts of PCAWG WGS tumor data



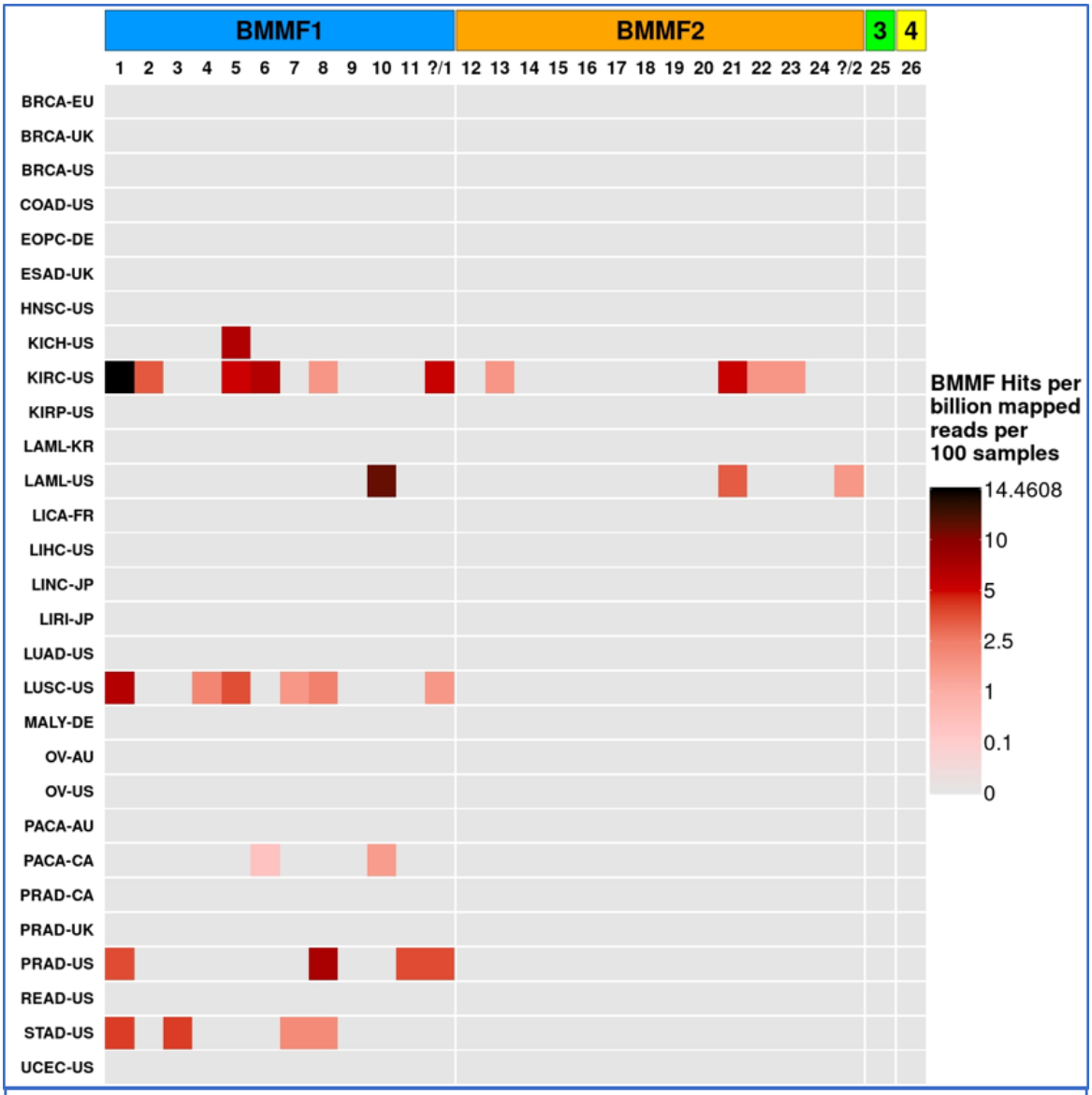
S12: BMMF subgroups detected in all cancer cohorts of PCAWG WGS tumor data: Overview of BMMF reads detected in PCAWG WGS tumor data cohorts after application of the three-hits threshold. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S13: BMMF subgroups detected in all cancer cohorts of PCAWG WGS normal tissue/blood data



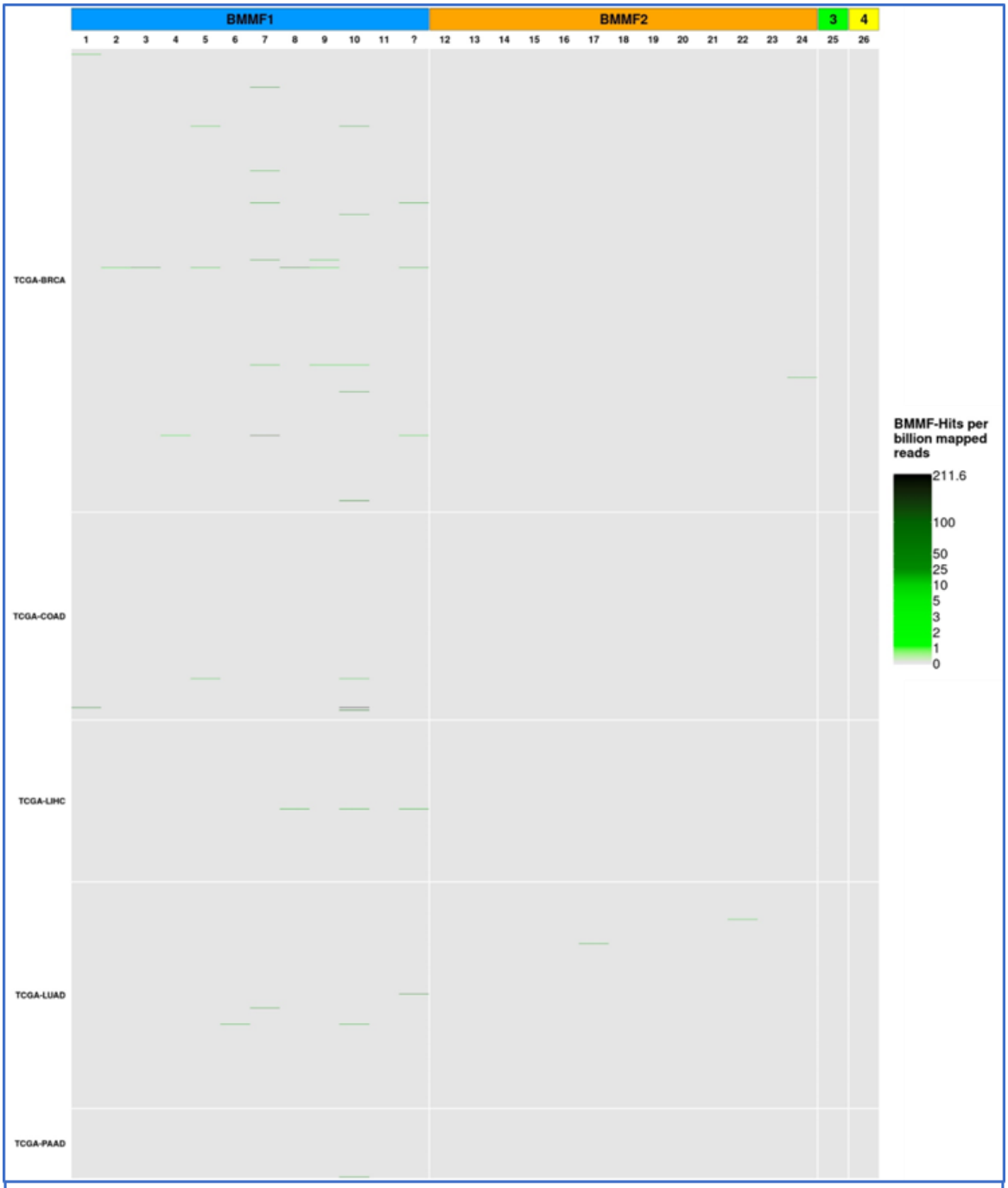
S13: BMMF subgroups detected in all cancer cohorts of PCAWG WGS normal tissue/blood data: Overview of BMMF reads detected in PCAWG WGS normal tissue/blood data cohorts after application of the three-hits threshold. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S14: Normalized BMMF reads detected for BMMF subgroups in PCAWG WGS normal solid tissue data



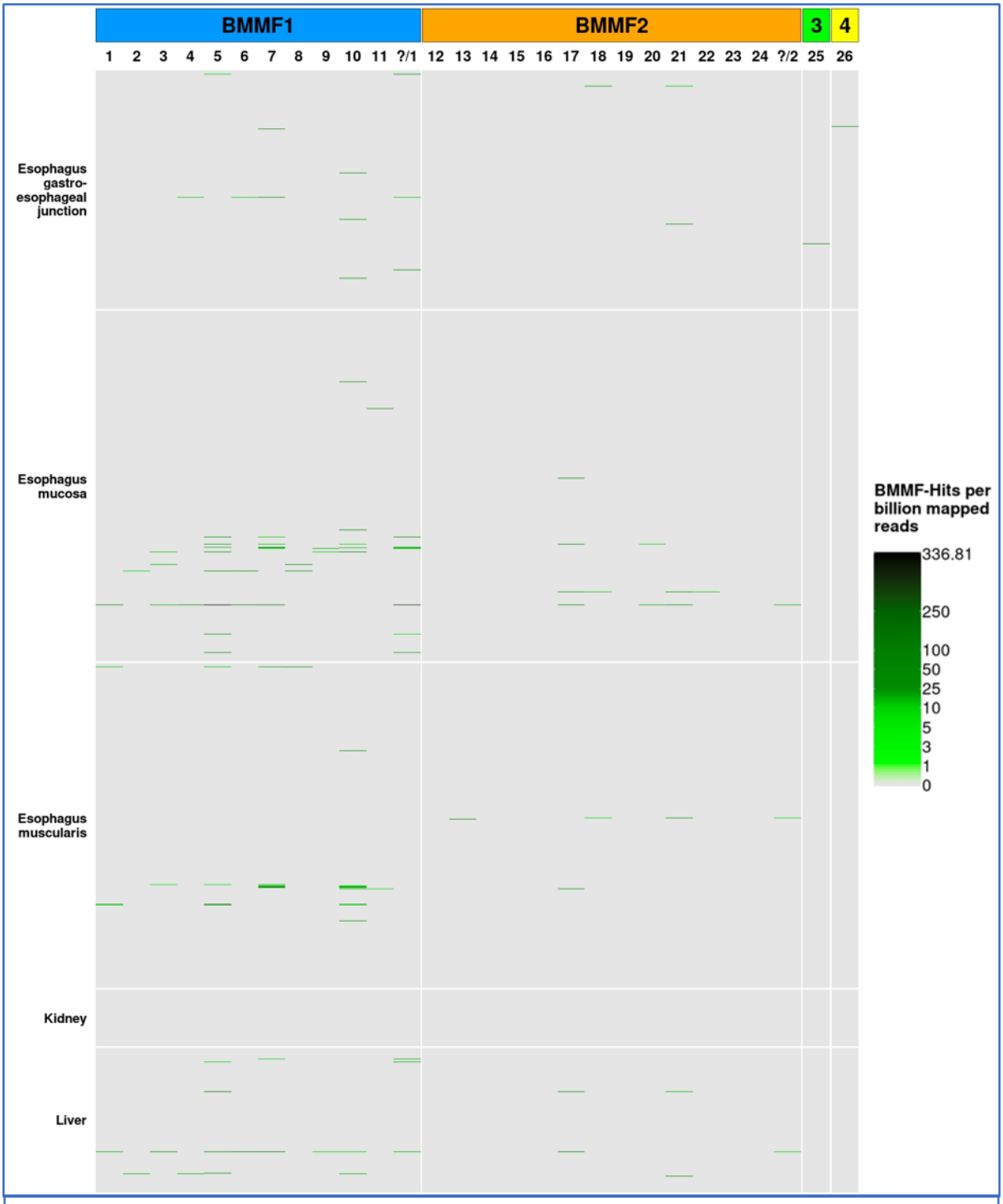
S14: Normalized BMMF reads detected for BMMF subgroups in PCAWG WGS normal solid tissue data: Overview of BMMF reads detected in PCAWG WGS normal solid tissue samples after application of the three-hits threshold and normalization for sequencing depth and assuming a cohort size of 100 samples. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S15: Normalized BMMF reads detected per patient in TCGA RNA data



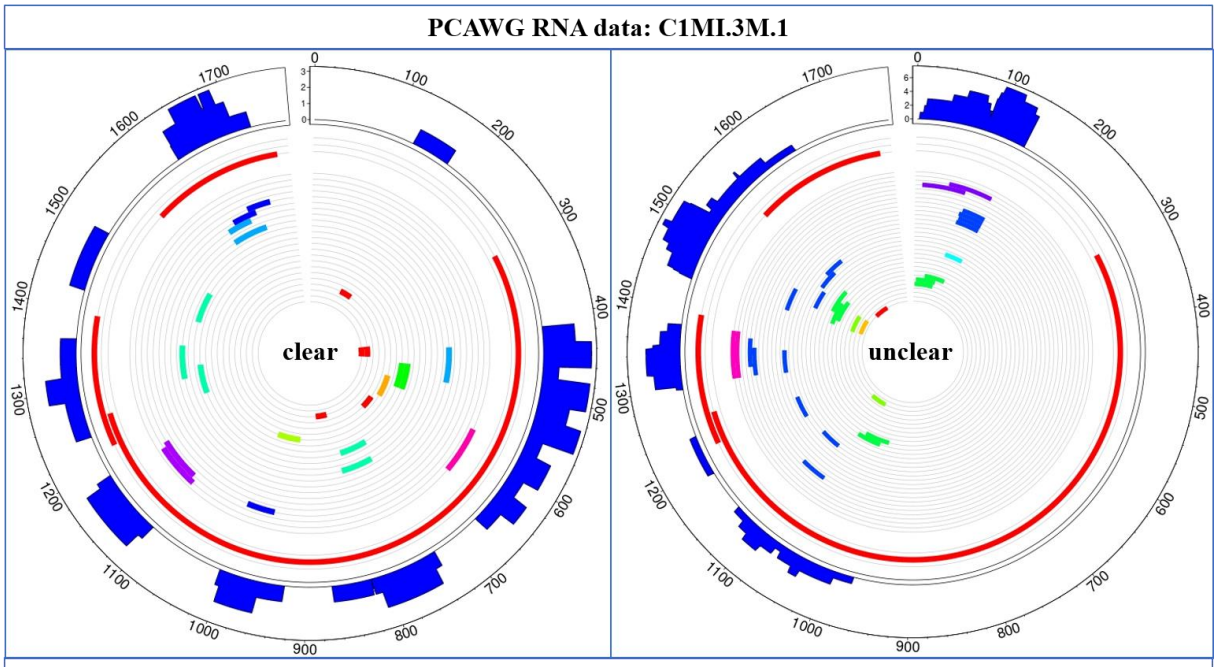
S15: Normalized BMMF reads detected per patient in TCGA RNA data: Overview of BMMF reads detected in TCGA RNA data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S16: Normalized BMMF reads detected per patient in GTEx RNA data (2)



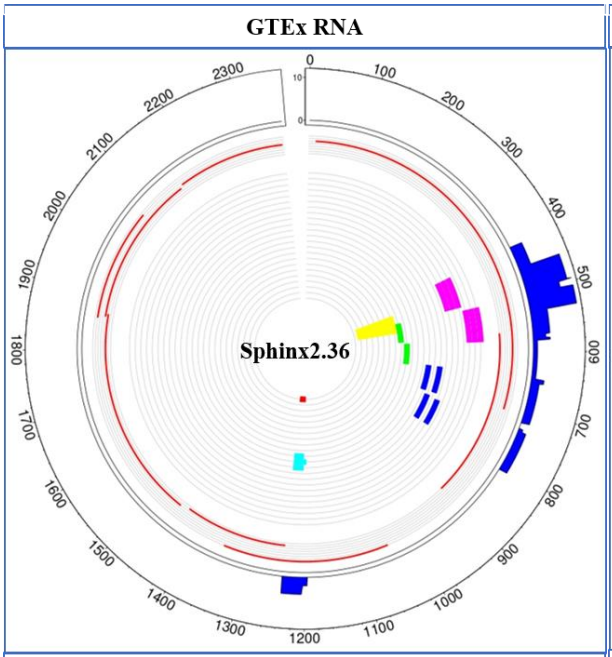
S16: Normalized BMMF reads detected per patient in GTEx RNA data (2): Overview of BMMF reads detected in GTEx RNA data after application of the three-hits threshold and normalization for sequencing depth. The BMMF group 1 and 2 isolates have been divided into 11, respectively 13 subgroups based on phylogenetic analyses. The detected BMMF reads are assigned to these subgroups based on sequence identity and e-value of the BLAST alignment to the library sequences. Reads that cannot be clearly assigned to one of the BMMF1 or BMMF2 subgroups are shown in the "?"-columns of BMMF group 1 and 2.

S17: Reads detected for BMMF group 1 isolate C1MI.3M.1 in PCAWG RNA data



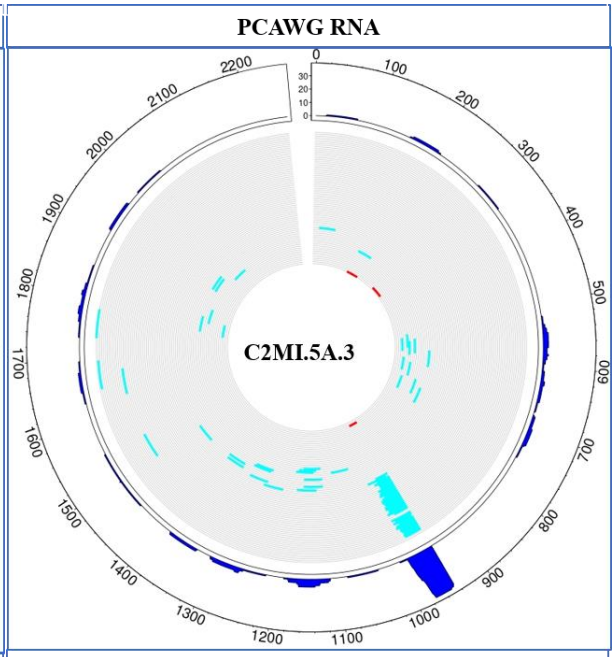
S17: Reads detected for BMMF group 1 isolate C1MI.3M.1 in PCAWG RNA data: Coverage of BMMF1 subgroup 5 genome C1MI.3M.1 by reads clearly assigned to subgroup 5 (left) and unclear reads (right). The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same cell are depicted in the same color.

S18: Reads detected for BMMF group 2 isolate Sphinx2.36 and for BMMF group 4 isolate MSSI1.162 in GTEx RNA data



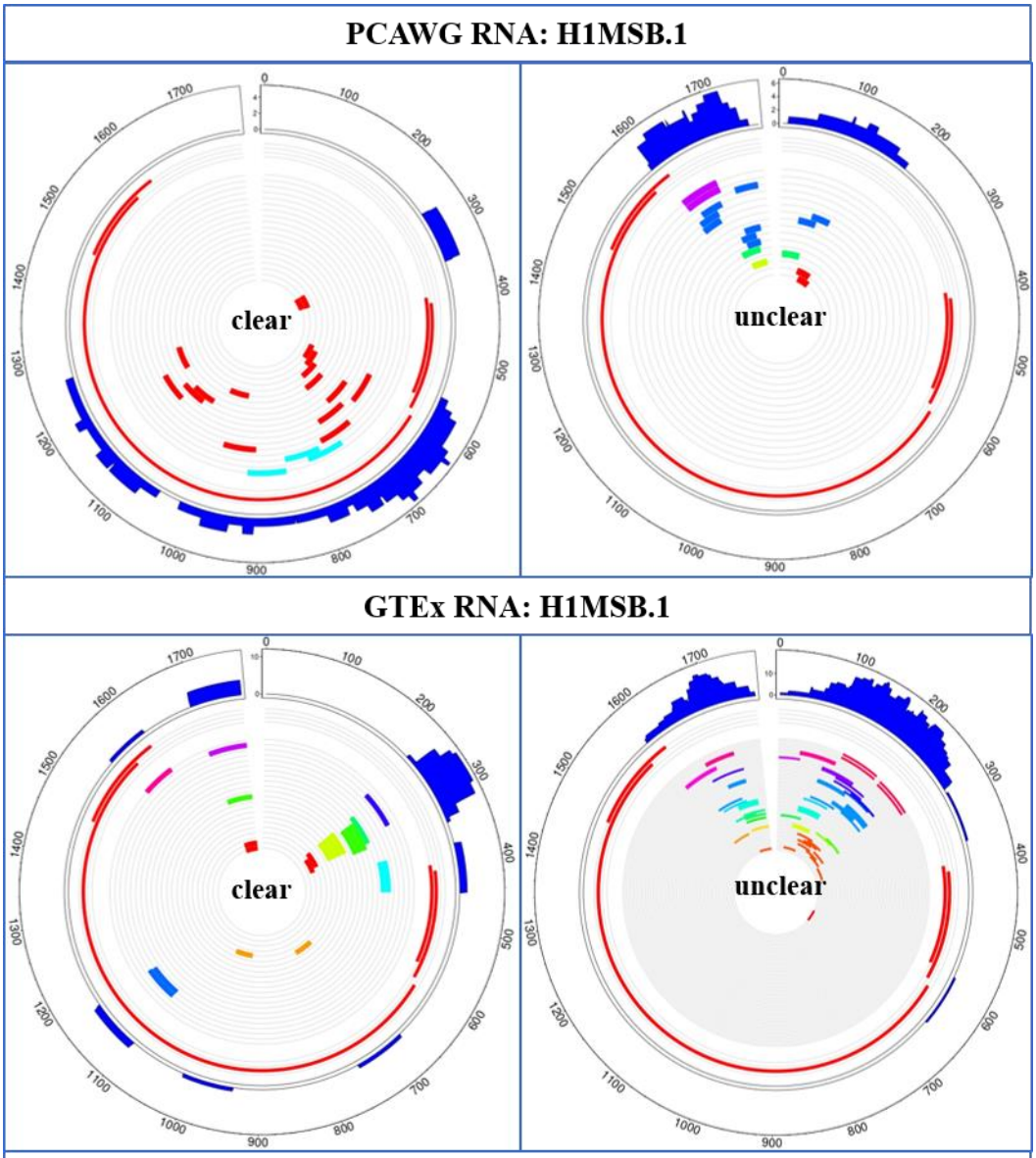
S18: Reads detected for BMMF group 2 isolate Sphinx2.36 in GTEx RNA data: Coverage of BMMF1 subgroup 21 genome Sphinx2.36 by reads clearly assigned to subgroup 21. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same cell are depicted in the same color.

S19: Reads detected for BMMF group 2 isolate C2MI.5A.3 PCAWG RNA data



S19: Reads detected for BMMF group 2 isolate C2MI.5A.3 PCAWG RNA data: Coverage of BMMF2 subgroup 21 genome C2MI.5A.3 by reads clearly assigned to subgroup 21. The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same cell are depicted in the same color.

S20: Reads detected for BMMF group 1 isolate H1MSB.1 in PCAWG and GTEx RNA data



S20: Reads detected for BMMF group 1 isolate H1MSB.1 in PCAWG and GTEx RNA data: Coverage of BMMF1 subgroup 1 genome H1MSB.1 by reads clearly assigned to subgroup 1 (left) and unclear reads (right). The outer circle shows in blue the number of detected reads per base pair. The middle circular segment shows the open readings frame predicted by ORFfinder. The inner circle maps the individual reads detected for the respective genomes. Reads detected in the same cell are depicted in the same color.