MULTI-DIMENSIONAL INTERROGATION OF DNA MUTATIONS IN CANCER

MOHAMED FEROZ BIN MOHAMMED OMAR

NATIONAL UNIVERSITY OF SINGAPORE
NUS GRADUATE SCHOOL FOR
INTEGRATIVE SCIENCES AND ENGINEERING
2015

MULTI-DIMENSIONAL INTERROGATION OF DNA MUTATIONS IN CANCER

MOHAMED FEROZ BIN MOHAMMED OMAR

B.Biotech (Hons), Flinders University

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR
INTEGRATIVE SCIENCES AND ENGINEERING
NATIONAL UNIVERSITY OF SINGAPORE

2015

DECLARATION

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

MOHAMED FEROZ B. MOHAMMED OMAR

11 AUGUST 2016

ACKNOWLEDGEMENTS

As written by John Donne, “No man is an island, entire of itself”. Concerning my PhD experience, this is so utterly true. I have to thank so many individuals for all they have done.

As a PhD student, I’ve had the privilege of working with several outstanding scientists and to these great figures, I am truly thankful. First among them, I would like to thank my initial PhD supervisor, Professor Manuel Salto-Tellez, without whom this entire journey would not have even begun. Secondly, I have to thank Prof Yoshiaki Ito, who took me on as a student when Manuel had to leave Singapore. Your wisdom has been a guidance for me throughout my time as a student, and your patience has been appreciated. To Professor Richie Soong, I am most grateful. Richie, you gave me a lifeline when I needed it most by taking me into your lab, allowing me to grow and giving me the trust and responsibility that you have. I will be eternally grateful. Dr Touati Benoukraf is the fourth scientist I would like to personally thank. He has been my mentor in the field of bioinformatics and an inspiration due to his many accomplishments as an up-and-coming scientist. I hope we will be able to work together for many years to come.

Science can only flourish in an environment filled with those who have a passion for seeking the truth and who set excellence as their standard. Doing the bulk of my research in Richie’s lab, CTRAD has given me the opportunity to work in such an environment. I am very happy to have interacted with everyone in the lab. I’d like to express a huge amount of gratitude to everyone who has helped me!

I started my PhD journey as a single man, but a couple of years in had the privilege of marrying the most wonderful woman in the world. Thank you, my dearest Aisyah,

for being patient with me for all these long years, for all those nights that I had to return home late, and for those weekends that I left us cooped up at home, rather than going out and enjoying ourselves. Thank you, my love.

To my parents, Bernard Morris and Shariffah Naina, thank you so very much for all you have done for your undeserving son. I owe you so much, hopefully, I can now concentrate on being the filial son I would like to be. And to my in-laws, Hamid Abdullah and Sarimah Kamit, thank you for being there too. You have helped me in so many ways.

To finish it off, I would like to state, Alhamdulillah, I am so happy that I have finally made it to the end of this unforgettable journey.

Sincerely,

Feroz.

TABLE OF CONTENTS

ABSTRACT

Cancers are known to develop through distinct DNA mutation events that occur during their initiation and progression. Recent studies have revealed that the analysis of point mutation patterns across cancer types can provide information on the carcinogenic origins and mutation susceptibilities of different cancers. In this study, it was postulated that additional insight can be obtained through the integrated analysis of multiple dimensions of DNA mutation events and that a tool for predicting cancer type from mutation patterns could be developed from this knowledge. The dimensions considered included the frequency, types and co-occurrence of point mutations, insertions and deletions, genes mutated and the genomic distribution of mutations. As a first step, a program was designed to provide efficient conversion of MAF files from next-generation sequencing (NGS) data into multi-dimensional mutation profiles. The programme was then applied to characterise mutation patterns from all data in The Cancer Genome Atlas (TCGA) database, comprising more than 8,000 tumours from 31 cancer types. Analysis of the results provided some interesting insights into the heterogeneity, subtypes, interrelation and biology of different cancer types. As a second step, multiple statistical and machine learning approaches were tested to determine optimal methods for building a tool to predict cancer type based on the mutation pattern of an unknown sample. Using the optimised method, close to 100% prediction accuracy was obtained in the analysis of random bootstrapped sample series from the TCGA. When applied to 5 non-TCGA NGS datasets, the prediction accuracy was 30-60%. While encouraging, the results also highlighted many issues, such as the need for standardisation of NGS protocols. In conclusion, these results have shown that multi-dimensional interrogation of DNA mutation patterns can provide novel insights into

cancer biology, and may be useful for predicting cancer types of samples of unknown origin through future development.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
CN	Count (analysis of the DNA mutations by counts in each category)
COAD	Colon adenocarcinoma
CTC	Circulating tumour cells
CUP	Cancer of unknown primary origin
ESCA	Esophageal carcinoma
fdr_bh	Benjamini & Hochberg false discovery rate analysis
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
ICGC	International cancer genome consortium
IHC	Immunohistochemistry
INDEL	Insertions and deletions
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid leukaemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
ML	Machine learning
NDT	Non-definable type: Cancers are heterogeneous across cases
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCA	Principal components analysis
PCPG	Pheochromocytoma and paraganglioma
PR	Proportions (analysis of the DNA mutations by proportions in each category)
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
RBM	restricted Boltzmann machine
SARC	Sarcoma
SCS	Shared Cancer signature: Cancers are homogeneous across cases, however profiles are not unique across cancers
SKCM	Skin cutaneous melanoma
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variants

STAD	Stomach adenocarcinoma
SV	Structural variant
TCGA	The cancer genome atlas
TGCT	Testicular germ cell tumours
THCA	Thyroid carcinoma
THYM	Thymoma
TNM	Trinucleotide mutation proportions
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UC	Unique cancer signature: Cancers are homogeneous across cases
US	Unique subtype signature: Cancers that are overall heterogeneous, however there is at least one homogeneous subgroup
UVM	Uveal melanoma
VT	Variance threshold
WES	Whole exome sequencing

Chapter I

Literature Review

1.1. Cancer: Global impact and biology

The term cancer describes a large group of diseases that are associated with uncontrolled growth and proliferation of cells within a specific region of the body that may invade into other parts of the body, either through direct expansion of a tumour mass or by metastasis to distant organs. It should be noted that not all tumours or neoplasm are cancerous as they may remain benign, however, cancers have been shown to develop from such cell masses upon specific trigger events (Qiu and Simon 2015). There are over 100 individual diseases that are considered cancer, distinguished by tissue/cell of origin and underlying mechanism of cause and disease evolution, resulting in widely varying risk factors, and respective epidemiology and impacts on health systems (Siegel, Naishadham, and Jemal 2012).

According to the GLOBACAN 2012 website (IARC 2015)., yearly, there were approximately 14.1 million new cases (Figure 1) and 8.2 million deaths worldwide due to cancer (Figure 2), estimated to represent about one in eight deaths that year (Ferlay et al. 2014). The most commonly diagnosed cancers worldwide were lung cancer, with 1.8 million cases in 2012, followed by breast (1.6 million), with colorectal in third (1.3 million) and prostate at fourth (1.1 million) (Figure 1). Lung cancer is also associated with the highest mortality rate worldwide (Figure 2) at 8.3 million cases per year. In contrast to their incidence rankings, mortality in liver (hepatic) cancer and stomach (gastric) cancers have the second and third-highest rates, with colorectal at fourth (693,000), breast at fifth (521,000) and pancreas ranked seventh (330,000). It is considered that the more aggressive nature of liver (Banerjee and Saluja 2015) (incidence of 782,000 vs mortality of 745,000) and stomach (Seo et al. 2015; Schlesinger-Raab et al. 2015) (951,000 vs 723,000) cancers underlie the higher relative mortalities compared to incidence rates.

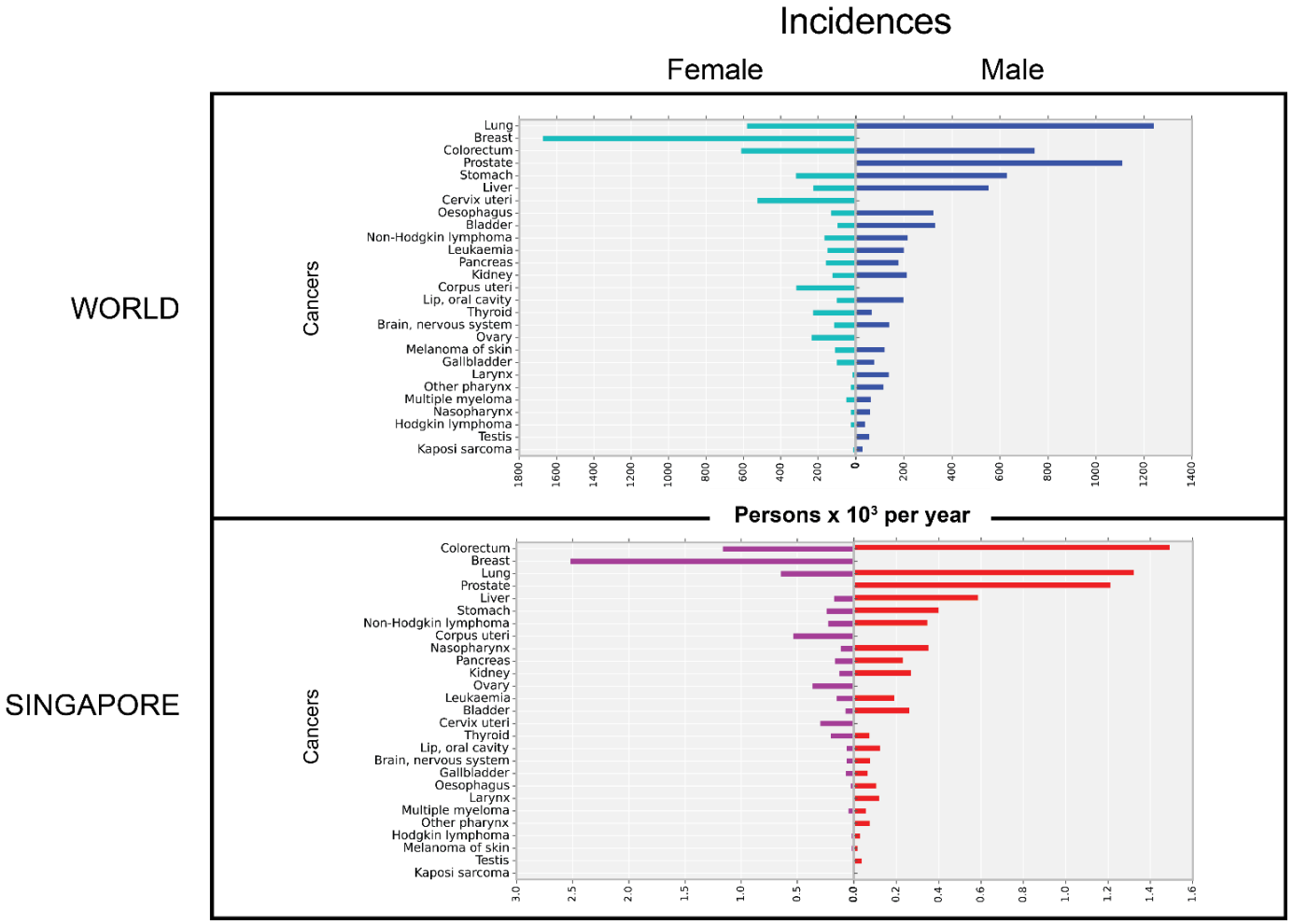

Figure 1: Incidences of the different cancers worldwide and in Singapore according to GLOBOCAN 2012.

Overall incidences of cancers in females are represented left of centre, and males are on the right. Cancers are listed from top to bottom according to the combined incidence rates in females and males.

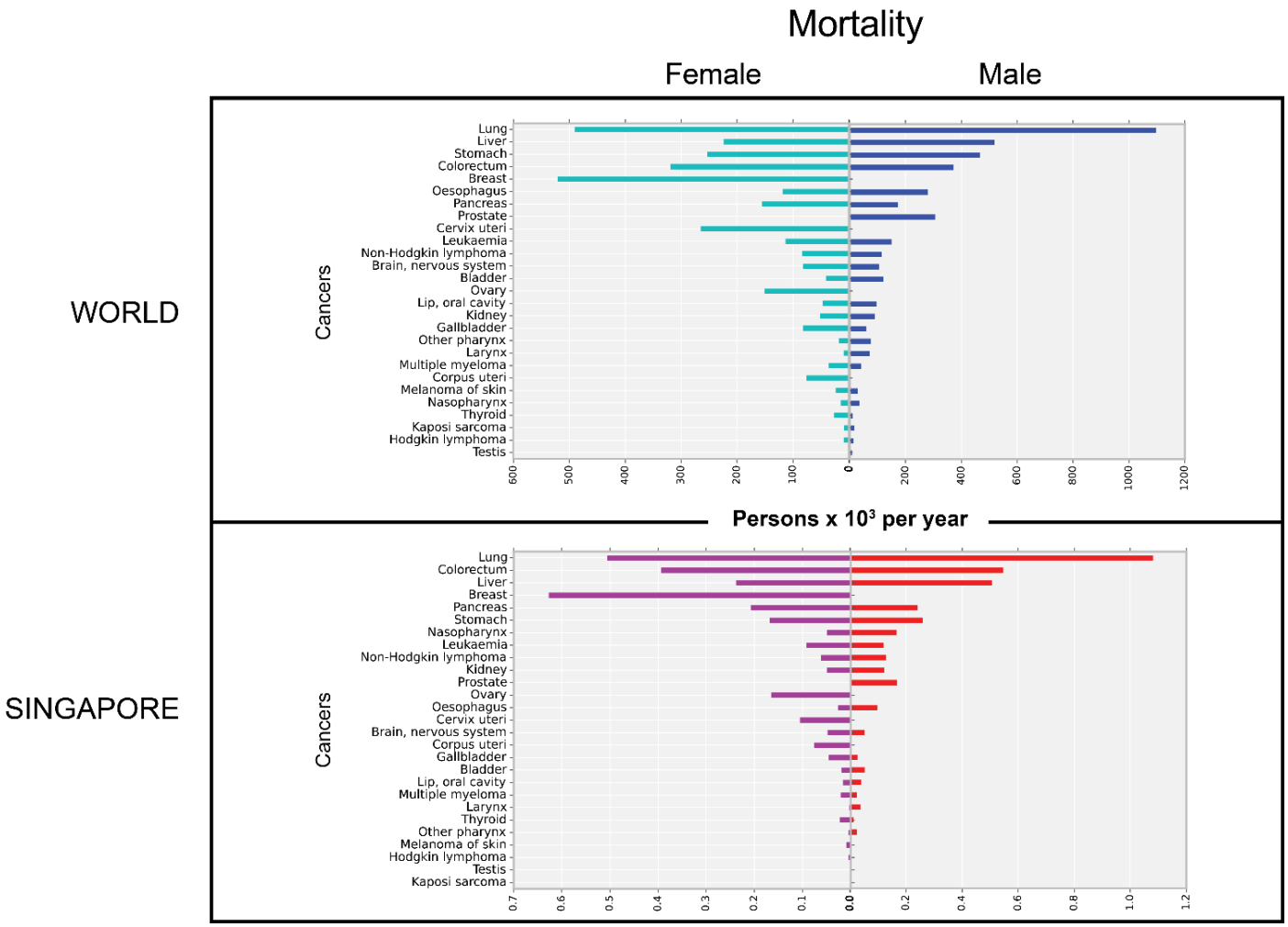

Figure 2: Mortalities of the different cancers worldwide and in Singapore according to GLOBOCAN 2012.

Overall mortalities from cancers in females are represented left of centre, and males are on the right. Cancers are listed according to combined mortality rates from the most to least frequent.

It should be noted however that the characterisation of cancers in GLOBOCAN, are based on the primary cancer site (the organ from which a cancer originates), a somewhat limiting characterisation. Cancer characterisations tend to go beyond just the primary sites and delve into cancer subtypes, which can have radically different characteristics at both the molecular level and biological level, despite having similar sites of origin. Liver cancer, for example, comprises many types of cancer, including cholangiocarcinoma (CHOL) and hepatocellular carcinoma (LIHC) both of which have very different symptoms, prognosis and treatment options (Khan et al. 2012; Kerkhofs et al. 2015). Lung cancer can be divided into small-cell carcinoma and non-small cell carcinomas (Long et al. 2015), of which the latter can be divided into adenocarcinoma (LUAD) and squamous-cell carcinoma subtypes (LUSC), all of which are histologically distinct. Subtypes can be characterised by the effects on patient survival and prognosis, or histological features (Schnitt 2010; W. D. Travis, Brambilla, and Riely 2013; Hu et al. 2012) or even by the emerging approach of molecular subtyping, which allows cancers to be interrogated and subtyped based on several molecular factors independent of histology. Examples of cancers that have been subtyped by molecular approaches include colorectal cancer (Muzny et al. 2012), urothelial cell carcinoma (Adam and DeGraff 2015) and plasma cell leukaemia (Simeon et al. 2015) for which improved diagnosis and prognosis are expected via improved patient stratification to better-suited treatment regimes. Another example of molecular characterisation is that of gastric adenocarcinomas (STAD) (Bass et al. 2014) where it has been shown that this cancer can be divided into four subgroups based on mutations, copy number aberrations, RNA expression, DNA methylation and the expression of specific proteins, allowing this cancer to be characterised with greater precision than in previous year.

There are also differences in cancer incidences and mortalities between men and women, a well-known but not well-studied phenomenon (McCann 2000) (Figure 1 and Figure *2*). There are obvious biological reasons in some cancers, such as the sex-specific organs, e.g. prostate cancer in men, and the exceedingly lower relative rates of breast cancer in men due to lower oestrogen and progesterone levels (Korde et al. 2010). However, in other cancers, this observation is not well understood (Tevfik Dorak and Karpuzoglu 2012). For example, hematologic malignancies, such as non-Hodgkin's lymphoma, are generally more common in males (McCann 2000). The two leading theories for these differences are hormonal differences and behavioural choices, e.g. differences in exposures to putative causative agents that both cause genetic and epigenetic changes. Autoimmune disorders, however, are more common in females, thus introducing the possibility that known differences in immunity may be responsible for this dichotomy. Lastly, the sex differences in genomic surveillance mechanisms may be responsible (Kirsch-Volders et al. 2010), where gender-related differences in responses to mutagens and carcinogens affect responses to chromosome damage.

There are also differences in incidences and mortalities rates between different geographical regions. For example, in Singaporean males, colorectal cancer is the most prevalent cancer type with 2662 cases and is associated with the second highest mortality (944 cases). In females, general trends in cancer incidence and mortality were more similar to the global cancer distribution. Racial and ethnic differences have been observed in a multitude of other cancers, and have been attributed to both cultural and hereditary causes. Examples include lung cancer (Haiman et al. 2006), prostate cancer (Peters and Armstrong 2005), and gastric cancer (Kuipers and Sipponen 2006). In general, colorectal cancer is found to be prevalent in East Asian populations with unique molecular characteristics (Jia et al. 2012; B. Zhang et al. 2014). In Singapore

specifically, the relatively high rate of gastric cancer in the Chinese population is attributed to specific genetic polymorphisms in addition to *Helicobacter pylori* (HP) infection status. These finding highlight the need to establish cancer statistics that are specific to different racial and ethnic groups, especially given their implications to biomarkers and patient stratification.

1.2. Cancers of unknown primary and Circulating cancer cells

Cancers of unknown primary site/origin (CUP) is an umbrella term for widely heterogeneous cancers that are distinguished by metastasis at the time of diagnosis, and for which the anatomical site of the primary tumour (origin) remains unknown, even after a detailed investigation (Briasoulis et al. 2005). This occurs in approximately 3 – 5 % of all malignancies, thus CUPSs are among the ten most frequently diagnosed cancers worldwide (Massard, Loriot, and Fizazi 2011).

CUPs were once viewed as a separate type of cancer from the other established cancer types, with the assumption that, regardless of the site of origin, these tumours shared biologic properties, which included rapid progression and dissemination. In recent years, however, this view has shifted to the notion that CUPs retain the signatures of the primary tumours, implying that treatment of CUPs could be similar to the primary cancer subtype (Varadhachary and Raber 2014). For example, global microRNA profiling has shown no significant expression differences when comparing known primary tumours with metastatic counterpart, suggesting molecular differences do not exist (Pentheroudakis et al. 2013). Despite advances in imaging, histological, and molecular profiling techniques used in identifying unknown primary cancer, there are still large challenges in identifying these primary sites. In the era of tailored therapeutic strategies, this situation presents both an opportunity and a challenge. Current methods to identify the primary sites of CUP include serological analysis of tumour antigens

(Greco, Vaughn, and Hainsworth 1986), pathological analyses (Oien and Dennis 2012) and molecular profiling (Löffler et al. 2016) with varying rates of success.

Circulating tumour cells (CTC) are cells which have migrated from primary tumour sites and entered the bloodstream, and these may have the potential to cause metastasis by invasion into other tissue. The belief that CTCs are a prerequisite to metastasis was first proposed by an Australian pathologist in 1869 when he observed that a patient with metastatic cancer had multiple identical tumours and that certain cells in the circulatory system shared a similar morphology (Ashworth 1869). The early presentation of these migratory cells in early stages of cancer progression further demonstrated the association of these cells with cancer progression, and it is currently considered that CTCs are indeed necessary for metastasis (Cristofanilli 2006). There are two main categories of methods for the detection of CTCs, namely nucleic-acid-based and cytometric approaches (Alunni-Fabbroni and Sandri 2010). Cytometric techniques are generally preferred, as they retain cellular integrity, allowing visualisation of morphology, enumeration and further molecular analysis such as protein quantification, fluorescent in situ hybridization or single cell DNA sequencing. CTC numbers have been proven to be an effective prognostic biomarker in breast (Hayes et al. 2006), colorectal (Cohen et al. 2009), lung (Hou et al. 2009) and prostate (De Bono et al. 2008) cancers. Due to the similarities between CTCs and the primary tumour, the molecular characterization of CTCs offers a unique ability to assess genotypic and phenotypic features of cancers without the need for an invasive biopsy (Krebs et al. 2010).

1.3. Cancers are caused by subtype-specific gene aberrations

The idea that cancers are molecularly driven diseases has been suggested since the early 1900s when chromosomal aberrations were microscopically observed in cancer cells. Soon after the discovery that deoxyribonucleic acid (DNA) could be responsible for heredity (Avery, Macleod, and McCarty 1944) and that it is a biopolymer of deoxyribonucleic acids arranged in a double helix (Watson and Crick 1953), there was speculation that cancers could arise directly for damage to DNA. This was evidenced by the fact that chemicals that damaged DNA tended to cause cancer and other abnormal growth characteristics (Loeb and Harris 2008). The Philadelphia translocation, which involved a translocation between chromosome 9 and 22 in chronic myeloid leukaemia, was perhaps the first specific genetic change that was associated with cancer (Rowley 1973). This translocation creates a fusion gene between ABL1, a proto-oncogene that encodes a tyrosine kinase, and BCR, a GTPase-activating protein, resulting in the formation of an oncogene. The discovery that the G > T single nucleotide variant in codon 12 of the HRAS genes (amino acid G12V) has transforming activity in bladder carcinoma cells was a significant point in establishing the idea that point mutations could have dramatic effects on cell activity and underlie cancer formation (Reddy et al. 1982). This discovery has been followed by the identification of a whole host of different cancer-associated mutations. A significant finding was published in 1990 where it was found that a significantly high proportion of tumours have TP53 protein expression and mutation aberrations (Iggo et al. 1990). Somatic mutations in TP53 have since been found in 38 – 50% of all cancers (Olivier, Hollstein, and Hainaut 2010), and have been shown to be both prognostic and predictive to disease outcome. Another significant observation is that many variants tend to be specific to certain cancers or even cancer subtypes. For example, EGFR, encoding a receptor

tyrosine kinase involved in activation of multiple cell survival pathways is mutated in approximately 10% of non-small cell lung cancers in the US and nearly 35 % of those in East Asians (Pao et al. 2004; Paez et al. 2004; Lynch et al. 2004), however, mutations are not found in other subtypes of lung cancer. These mutations occur almost exclusively within exons 18–21 of the gene, a region which encodes a portion of the EGFR kinase domain and are predictive of response to treatments with EGFR inhibitors such as gefitinib and erlotinib. As another example, *KRAS* is mutated in 15-25% of lung adenocarcinomas, however, such mutations are rare in lung squamous cell carcinomas (Brose et al. 2002). Interestingly, these *KRAS* mutations are mutually exclusive to the presence of *EGFR* mutations (Schmid et al. 2009; Omar et al. 2012) and are negative predictors of response to erlotinib and gefitinib treatment.

A number of recent breakthrough studies have identified cancer-specific recurrent mutations through the use of next-generation sequencing. For example, *ARID1A, FAT4, CDH1* and *RHOA* have been found to be recurrently mutated in gastric adenocarcinomas (Zang et al. 2012; Bass et al. 2014), twenty-four genes including *ARID1A, SOX9* and *FAM123B* have been found to be frequently mutated in colorectal cancer (Muzny et al. 2012), 11 genes disrupting *NFE2L2* and *KEAP1* of the PI3K pathway, along with *CDKN2A* and *RB1* are frequently mutated in lung squamous cell carcinomas (Hammerman et al. 2012), while *PPP6C, RAC1, SNX31, TACC1, STK19,* and *ARID2* have been found to be recurrently mutated in melanomas (Hodis et al. 2012). A recent study, which performed a statistical summary of mutations from the TCGA initiative, identified 77 significantly mutated genes including protein kinases, G-protein-coupled receptors such as *GRM8, BAI3, AGTRL1, LPHN3, GNAO1 MAP2K4* and other druggable targets in 12 cancer types (Kandoth et al. 2013). Several of the significant genes appeared to be associated with certain cancers types. Databases

such as ClinVar (Landrum et al. 2016; Landrum et al. 2014) and "My Cancer Genome", http://www.mycancergenome.org/ (Vanderbilt-Ingram Cancer Center 2010) are prominent databases of clinically relevant and phenotypically associated variants and list many cancer subtype-specific variants.

1.4. Cancers develop due to several mutational processes

Various genetic mutations types have been associated with cancer formation include single nucleotide variants (SNVs), small insertions and deletions (indels), structural variations (SV), copy number variants (CNV), aberrant DNA methylation and other epigenetic abnormalities, with the first two being the most frequently related to cancer (Kandoth et al. 2013). SNVs, also known as point mutations, are the most common and involve the substitution of a single nucleotide (A, T, C or G) with another, e.g. **T** to **A**, reciprocated in the complementary DNA strand as an **A** to **T**. For purposes of clarity, SNVs are distinct from single nucleotide polymorphism (SNP) in that SNPs denote single base differences that occur within a normal population (germline) at specific frequencies (e.g. 1%), while the term SNVs does not generally take into account population frequencies. Transitions (Ts) describe specific SNVs involving the replacement of a purine base (A or G nucleotide) with another purine or the replacement of a pyrimidine (C, T or U nucleotide) with another pyrimidine. On the other hand, transversions (Tv) are the replacement of a purine with a pyrimidine or vice versa. There are twice as many possible transversions as transitions, although transition mutations appear to be generated at a ratio (Ts:Tv) of about 2.1:1 in humans (Keller, Bensasson, and Nichols 2007), however, this ratio is not observed in all organisms.

In contrast to SNVs, which have been studied extensively, other forms of natural genetic variation in humans such as insertions and deletions (indels) have received relatively little attention (Yang et al. 2010). Indels are the second most common type

of genomic variant and the most common type of structural variant. This aberration involves changes to DNA, in which nucleotides are either added or removed without replacement (Mills et al. 2006). The aberrations can range from 1 to 10,000 base pairs (bps) in length, although the vast majority are 1–10 bps long. An estimated 0.13-0.4 million short indels are found per individual in normal cells (Mills et al. 2006). In an open reading frame, unless an indel length is a multiple of three, a frameshift mutation is resultant which would affect all amino acids in the protein after the indel position, a highly deleterious effect.

SNVs and indels can be characterised according to their location relative to coding genes, e.g., 5' untranslated region, exonic, intronic or intergenic etc., and by the genomic implication. The implications include missense mutations that can lead to changes in encoded amino acids, non-stop mutations that convert a stop codon into another codon and thereby extending the protein beyond its proper length, or an inframe insertion in which the insertion does not disrupt the codon sequence of the transcript.

1.5. Human DNA variation databases

The accumulating evidence that specific mutation processes may underlie cancers and other diseases and the growing catalogue of identified DNA variants has led to the establishment of numerous databases and DNA sequencing initiatives. The Single Nucleotide Polymorphism Database (dbSNP) (Sherry, Ward, and Sirotkin 1999) compiles a range of small molecular variation including SNPs and indels. The current version of this database is Build 144 (Jun 08, 2015), with over 85 million human variants and a host of variants from other organisms. dbVar (NCBI 2015) is a database of genomic structural variation as a counterpart to the previous database, as it includes insertions, deletions, duplications, inversions, translocations, and complex chromosomal rearrangements. The 1000 genomes project (Abecasis et al. 2012) is an

international research effort to establish the most detailed catalogue of normal human genetic variation. The Catalogue Of Somatic Mutations In Cancer (COSMIC) (S. a Forbes et al. 2011) is an online database of somatically acquired mutations found in human cancer using curated data from papers in the scientific literature and large-scale experimental screens. This database has specific sections for cell line data and mutations derived from whole genomes. The mutation databases described are essential to helping researchers put their experimental results into the context of all other studies and thus understand the significance of their findings.

Enabled by the increasing cost-effectiveness of next-generation sequencing (NGS), numerous large-scale genome studies were recently undertaken by The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, Weinstein, et al. 2013), International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium et al. 2010) and various independent laboratories. These ambitious projects have generated comprehensive catalogues of somatic mutations that are present in human cancers and revealed unprecedented insights into how various genomic contexts drive tumorigenesis. The TCGA project is a joint effort of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) and is the most comprehensive resource for NGS analyses of human cancers, including genomic sequencing, miRNA sequencing and expression, methylation data and copy number analysis, for the first time providing resources for massive interrogation of cancer signatures from 33 cancer types from over 11,000 patients. This resource is especially attractive as the different cancers are analysed using standardised pipelines, despite coming from multiple sources within the US. The International Cancer Genome Consortium (ICGC) is a similar database with the stated goal of cataloguing large-scale cancer genome studies in tumours from 50 cancer types. Unlike the TCGA, this project

is an international collaboration, with data from 16 countries making up 78 individual cancer projects with 46 cancer types so far, from approximately 9500 patients, and as such, is also a rich source of data (J. Zhang et al. 2011) that can help study the various mutation seen in cancers. Both the TCGA and ICGC contain continually expanding datasets which are available to the scientific community.

1.6. Cancer Specific DNA Signatures

The molecular interrogation of somatic DNA profiles is considered to be crucial to the actualisation of personalised oncological medicine, which is premised on utilising knowledge of the cancer mutation repertoire to classify tumours and guide the administration of anticancer agents targeting aberrant genes and pathways (Syn et al. 2016). Common frameworks for understanding these mutational profiles include (1) Gene-level analysis, which is based on the identification of recurrently and significantly mutated genes that are likely to be important in cancer pathogenesis, and was the focus of section 1.3; and (2) Analysis of mutational signatures, whereby cryptic ("gene-independent") patterns in nucleotide sequences are deciphered, and in turn may reveal potentially-actionable underlying defects in intrinsic cell-biological processes as well as environmental carcinogenic exposures.

Any individual base substitution may be represented by one of six possible transitions or transversions (i.e. C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G and T:A>G:C) (Helleday, Eshtad, and Nik-Zainal 2014; Nik-Zainal, Alexandrov, et al. 2012; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013; Alexandrov, Nik-Zainal, Wedge, Campbell, et al. 2013). However, because the flanking sequence context (i.e. neighbouring bases 5' and 3' to the substituted base) may influence mutation rates (Ellegren, Smith, and Webster 2003), and each immediately neighbouring base could

be A, C, T or G, a given base substitution could occur within any of $4 \times 6 \times 4 = 96$ trinucleotide contexts.

Another pattern in DNA sequences is the size of insertions and deletions (collectively, indels), which varies across cancers in terms of absolute counts, relative proportions and distributions (Greenman et al. 2007; Yang et al. 2010). Indels are a prominent feature in cancer genomes which may be associated with defective DNA replication and recombination repair phenotypes, such as post-replicative mismatch repair (MMR) deficiency (Bhattacharyya et al. 1994; Karran 1996; Kuraguchi et al. 2000).

The significance of studying patterns of mutations in cancers was revealed by Greenman et al (Greenman et al. 2007) in which specific mutation profiles from 8 different cancer types were identified by analysing kinome sequencing data, indicating unique carcinogenic initiators of these diseases, comprising one of the first large-scale uses of next-generation sequencing. The study revealed that overall cancer types appear to exhibit unique mutational profiles and by segregating C:G>T:A mutations into those that do or do not coincide with CpG sites, this study also highlighted the possible relationship between DNA methylation and mutations as factors in cancer development.

The importance of studying these hidden patterns and combinations of DNA sequence alterations in human cancers was again highlighted by Alexandrov and colleagues, who in 2013 published an initial set of 21 mutational signatures which has provided much guidance to the understanding of cancer development (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013). The list has since expanded to a total of 30 signatures extracted from 10,952 exomes and 1,048 whole-genomes across 40 cancer types (Wellcome Trust Sanger Institute 2016b). As alluded to earlier, the decryption of

somatic DNA signatures – however elusive – may provide major leads into mutational processes such as carcinogenic exposures and infidelities in the DNA maintenance machinery which become imprinted in the genome over the biological history of a cancer. Hence, the analysis of somatic mutational signatures may be useful to the understanding of cancer aetiology, and may lead to new possibilities for preventative, predictive and therapeutic strategies. This section presents a comprehensive summary of several well-established mutational signatures, their associated aetiologies and cancer types.

Environmental carcinogens are frequently invoked as culprits in mutagenesis and cancer development. One of the best known source of DNA damage is attributed to tobacco smoke (Hainaut and Pfeifer 2001; Pfeifer et al. 2002; Hainaut, Olivier, and Pfeifer 2001; Pleasance, Stephens, et al. 2010), which is epidemiologically associated with lung, esophageal, liver, and head and neck cancers. Compared to never-smokers, the mutational spectrum of lung cancer among smokers is characterised by significantly elevated mutational burden (~10-fold); as well as enrichment for C:G>A:T substitutions exhibiting transcriptional strand bias (Govindan et al. 2012; Imielinski et al. 2012), which corroborates experimental findings that polycyclic hydrocarbons such as benzo[a]pyrene in tobacco smoke induce bulky adduct formation on guanine (Rodin and Rodin 2005). As evidence of the potential clinical utility of DNA mutational signatures, the molecular smoking signature was recently proposed as a predictive biomarker of checkpoint blockade immunotherapy in non-small cell lung cancer (Naiyer A Rizvi et al. 2015).

Prolonged exposure to sunlight, or particularly to ultraviolet B radiation (UVB, 290 – 320 nm wavelength), is a well-established risk factor for malignant melanoma. Non-ionising ultraviolet (UV) radiation is known to generate helix-distorting

cyclobutane pyrimidine dimers which may quickly overwhelm transcription-coupled nucleotide excision repair (NER) (Pleasance, Cheetham, et al. 2010; Durbeej and Eriksson 2003; Pfeifer, You, and Besaratinia 2005; Brash et al. 1991). Consistent with the photochemistry of pyrimidine dimers, the UV-associated mutational signature is uniquely characterised by C>T transitions at dipyrimidine sequences and CC>TT dinucleotide substitutions (Hodis et al. 2012; Drobetsky, Grosovsky, and Glickman 1987; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013). Perhaps reflecting the long-term manifestation of sunlight-induced DNA damage, the age of diagnosis of melanoma is correlated with accumulation of the UV-associated signature (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013).

Some common antineoplastic agents, including cisplatin, cyclophosphamide and temozolomide, occasionally spawn secondary primary malignancies whose molecular portraits are clearly unrelated to the initially treated cancer, but instead bear distinctive signatures of chemotherapy-induced genotoxicity (Szikriszt et al. 2016; Huang et al. 2016; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013; Greenman et al. 2007; Hunter et al. 2006; Tomita-Mitchell et al. 2000). Analysis of the National Cancer Institute's Surveillance, Epidemiology, and End Results Programme (SEER) database show that approximately 1/6 of incident cancers are second- or higher-order primary neoplasms; and the risk of developing secondary cancers associated with cytotoxic chemotherapy increases with dose intensity and treatment duration (L. B. Travis 2006; A. K. Ng and Travis, n.d.). Szikriszt et al recently reported the mutagenic impact of eight widely used antineoplastic agents, including cisplatin, cyclophosphamide, hydroxyurea, gemcitabine, 5-fluorouracil, etoposide, doxorubicin and paclitaxel (Szikriszt et al. 2016). They found that cisplatin, cyclophosphamide and etoposide-induced significant quantities of base substitutions each characterised by their unique

mutational spectra. For instance, in chicken DT40 lymphoblast cell line, which has a spontaneous mutagenesis rate of ~2.3 × 10-10 per base per cell division, four cycles of cisplatin exposure was sufficient to generate mutational burden equivalent to that of common leukaemias (~0.8 mutations per megabase) (Szikriszt et al. 2016). Moreover, the mutational signature associated with each drug is likely related to their unique mechanisms of action. For instance, cisplatin-induced base substitutions (of which 57% are C:G>A:T) and short indels are primarily localised to its putative sites of purine intrastrand crosslink formation; while cyclophosphamide treatment strongly elevated T>A (possibly caused by phosphotriester adducts) and C>T (possibly related to the formation of N7-guanine monoadducts and G-G interstrand crosslinks) substitution rates (Szikriszt et al. 2016).

Aristolochic acid (AA) is a natural compound found in many Aristolochia plants that are widely used in traditional herbal remedies from China to Romania and Croatia. However, this compound is mutagenic and has been associated with increased incidences of bladder, hepatocellular, and urothelial cell carcinomas among users. The AA-associated mutational signature is characterised by strikingly high mutation burden relative to their cancer type, with approximately ~70% of mutations being A:T to T:A transversions (Poon et al. 2015; Poon et al. 2013; Hoang et al. 2013). Furthermore, these transversions display a preference for the trinucleotide motif T/CAG and significant bias for the non-transcribed strand (Poon et al. 2015; Poon et al. 2013; Hoang et al. 2013). Because this trinucleotide motif coincides with the canonical splice acceptor sequence, AA-associated tumours display upregulation in the nonsense-mediated decay machinery as a result of aberrant splicing events and implicate splice site mutations in the pathogenesis of AA-associated tumours (Hoang et al. 2013; Poon et al. 2013).

Somatic mutations accumulate over the lifetime of an individual. The pair of age-associated mutational signatures (1A/B), as defined by Alexandrov et al, are characterised by the prominence of C>T substitutions at CpG sites, ubiquitous across at least 15 cancer types, and probably reflect endogenous mechanisms present in normal and neoplastic cells alike such as the spontaneous deamination of 5-methylcytosine to give thymine (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013; Walser, Ponger, and Furano 2008).

Under normal physiological contexts, the APOBEC/AID family of cytidine deaminases have important roles in RNA editing, as well as innate and adaptive immunity (including retrovirus restriction, hypermutation and recombination of immunoglobulin genes). However, cytidine deaminases, which convert cytosine to uracil, are capable of sculpting the landscape of chromosomal and even mitochondrial DNA (Suspène et al. 2011; Shinohara et al. 2012; Burns et al. 2013; S. a Roberts et al. 2013; Nik-Zainal, Alexandrov, et al. 2012; S. A. Roberts et al. 2012). The APOBEC editing pattern is pervasive and has been described in several cancers including pancreatic, prostate, breast, head and neck cancers, and haematological malignancies including multiple myeloma, chronic lymphocytic leukaemia and B-cell lymphoma. This editing pattern is characterised by C>T, C>G and C>A mutations within TpC contexts, with kataegis showing an even higher preference for TCA or TCT trinucleotide motifs, and positively correlates with APOBEC mRNA levels (S. a Roberts et al. 2013). Interestingly, both APOBEC-induced kataegis and non-clustered substitutions tend to localise near rearrangement breakpoints, possibly because of the affinity of APOBEC enzymes for single-stranded DNA (Walker et al. 2015; Drier et al. 2013; Smith et al. 2012). As evidence of their prognostic value, APOBEC/AID signatures in multiple myelomas is associated with multiple features which predict

poorer overall survival, including MYC, t(14;16) and t(14;20) translocations, and CCND1 mutations (Walker et al. 2015).

Defective DNA mismatch repair (MMR) is operative in approximately 1/5 of colorectal cancers and 1/7 of uterine cancers, and detectable in at least 1% of cancer samples in esophageal, liver, lung, stomach, cervical, breast and kidney cancers (Alexandrov and Stratton 2014). Tumours which exhibit MMR-deficiency may possess one or more of four distinct mutational signatures, which are distinguished by the frequency of trinucleotide alterations. For instance, one of these four signatures displays prominence of C>T at NpCpG sites whereas another shows enrichment for C>T at GpCpN contexts. The former signature is correlated with the inactivation of DNA mismatch repair genes in colorectal cancer ($P = 3.3 \times 10\text{-}5$) (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013). However, one common feature among these signatures is their association with large numbers of small indels (<3bp) at mono/polynucleotide repeats, also termed as 'microsatellite instability'. Failure of post-replicative MMR could be a consequence of biallelic somatic mutations or epigenetic inactivation of MMR genes such as MLH1 (Helleday, Eshtad, and Nik-Zainal 2014; Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013; Boland and Goel 2010).

The aberrant activity of the B family DNA polymerase ε, which typically has an error rate of 1 in 10^{-7} nucleotides synthesised (Shevelev and Hübscher 2002) and is encoded by the POLE gene, result in "ultra-hypermutability" characterised by a striking pattern of C>A at TpCpT and C>T at TpCpG contexts in colorectal and endometrial cancers. This may occur a result of somatic and germline hotspot mutations in the exonuclease domain in POLE such as Pro286Arg and Val411Leu, which is suggested to lead to loss of proofreading capacity and replication fidelity (Cancer Genome Atlas Network 2012; Cancer Genome Atlas Research Network, Kandoth, et al. 2013).

The final signature which this section covers is one attributed to defective homologous-recombination-based DNA double-strand break repair, which is proposed to be a result of germline and/or somatic BRCA1/2 inactivation or abnormalities (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013), and is present in breast, ovarian and pancreatic cancers. The BRCA1 protein is responsible for the resection of DNA ends while BRCA2 mediates the loading of RAD51 onto single-stranded DNA (Helleday, Eshtad, and Nik-Zainal 2014; Moynahan et al. 1999; Moynahan, Pierce, and Jasin 2001; Bunting et al. 2010; Davies et al. 2001). Hence, their inactivation results in non-homologous end-joining repair of DNA double-strand breaks which leads to large indels (longer than 3bp and up to 50bp) with overlapping microhomology at breakpoint junctions. This signature exhibits an overall equal representation of all 96 trinucleotide alterations. Interestingly, pancreatic cancer patients who respond to platinum therapy usually exhibit the BRCA1/2–associated mutational signature (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013).

Overall these studies reveal the significance of specific variants and mutated genes in the different cancer subtypes. It is thus conceivable that mutational signatures, along with SNV and indel profiles, may be used as unique identifiers of cancer subtypes, however, there has been little investigation into whether this is possible.

1.7. Next-generation Sequencing (NGS)

NGS (also known as High-throughput sequencing (HTS) or second generation sequencing or massively parallel sequencing) technologies, have allowed the rapid and accurate sequencing of expressed genes (transcriptomes), known exons (exomes) and even complete genomes of cancer samples, a task that would take significantly greater resources by older sequencing technologies (Meyerson, Gabriel, and Getz 2010; Mertes et al. 2011; Xi, Kim, and Park 2010). NGS, as used in this study, encapsulated a set of

technologies that involve 1) the enrichment of short segments of DNA from an organism that varies from 35 to 400 bases, depending on the technology, 2) the determination of the nucleotide sequence of the enriched DNA (reads), 3) The alignment of these reads to a known genome assembly by using optimised algorithms and finally 4) the determination of characteristics of the organism's genome via comparison to the reference genome and/or to other organism's DNA. These technological advances are important for advancing our understanding of malignant neoplasms because cancer is fundamentally a disease of the genome. A wide range of genomic alterations, including point mutations, copy number changes and rearrangements can lead to the development of cancer. Most of these alterations are somatic, that is, they are present in cancer cells but not in a patient's germ line (Meyerson, Gabriel, and Getz 2010). Thus a genetic analysis comparing cancer vs. germline DNA may reveal intricate patterns of cancer development (Goh et al. 2011).

NGS experiments are either replacing or complementing other high-throughput technologies such as beadChips or microarrays (Hurd and Nelson 2009) or low-throughput approaches such as real-time PCR (Tuononen et al. 2013). The advantage over microarrays/bead chips are the larger dynamic range that NGS technologies can achieve and the fact that variant detection is not limited to specific probe locations, significantly, novel mutations can be discovered. The high-through nature of NGS allows for the analysis of larger regions of the genome than real-time PCR, at much more cost-effective rates. NGS approaches are also very versatile having applications in DNA sequencing, RNA sequencing and expression determination (Wang, Gerstein, and Snyder 2009), DNA methylation analysis (Adusumalli et al. 2015) and elucidation of chromatin biology (Wang, Gerstein, and Snyder 2009).

When carrying out sequencing using NGS, the actual approach depends on the scope of the genome to be studied. Specific regions of the genome can be analysed via targeted panel sequencing (TPS), an approach taken when a study requires the sequencing results from candidate genes or regions, usually cancer-specific or pan-cancer genes and when a great level of data clarity is required (Tomlinson 2012) and is typically performed on low to mid-range sequencers such as the MiSeq (Illumina), SOLiDv4 (Life Technologies) or Ion PGM (Thermo Fisher Scientific Inc). When the full complement of protein-coding segments of the genome is to be analysed (180,000 exons from about 30 megabases), whole exome sequencing (WES) is used (S. B. Ng et al. 2009; Asan et al. 2011; Parla et al. 2011) generally performed on mid to high-range sequencers such as the NextSeq 500 or HiSeq from Illumina or the Ion Proton (Thermo Fisher Scientific Inc). When the entire sequence of the genome is to be studied, whole genome sequencing (WGS) can be used, allowing for the analysis of functionally relevant non-coding regions (Kellis et al. 2014) including noncoding RNA which plays a role in disease regulation (M. Li et al. 2009), cis- and trans- regulatory elements which control transcription of distant genes (Visel, Rubin, and Pennacchio 2009), functional intronic sequences, pseudogenes which potentially regulate the expression of protein-coding genes (Sisu et al. 2014), and many more regions involved in genomic regulation. WGS requires high-range sequencers such as the HiSeq or if exceedingly high-throughput is required, the HiSeq X Ten system (Check Hayden 2014). Of specific relevance to the work in this thesis is WES and to a lesser extent WGS, as data generated from these sequencing approaches, available via that TCGA and ICGC databases, are utilised.

1.7.1. Bioinformatics in NGS

Current knowledge of the mutational panels present in cancers has been revealed through NGS interrogation of cancer as never before. In recent years a large number of techniques have been developed to tackle the different procedures involved in NGS analysis with the ultimate goal to reveal the variants present in disease states. Figure 3 shows the workflow of a comprehensive NGS DNA analysis from sample preparation to final analysis, highlighting several of the tools used currently.

Before DNA-based NGS can be carried out, DNA must be extracted and quality control tested to ensure high-quality DNA for the NGS analysis. For samples of an appropriate quality, library preparation can then be carried out. Briefly, this is a process that shears DNA and then anneals tagged sequences (adapter sequences) to the ends of the sheared sequences to facilitate the polymerase chain reaction (PCR) amplification procedure, which is required prior to sequencing. Sequences for identification of samples (bar codes) may also be added to allow multiplexing. The process of amplification is called cluster generation on Illumina machines. These amplified sequences are then sequenced using different mechanisms call chemistries and is dependant of the sequencing machine used. Illumina uses a system called sequencing by synthesis, while Ion based machine use the Ion semiconductor sequencing approach. As summarised in Figure 3, following sequencing, several bioinformatics (*in silico*) techniques are carried out, including the conversion of raw base calling files (BCL) (Illumina) or DAT files (Ion torrent) into the FASTQ format (Cock et al. 2010), which is a text-based format that stores both sequence information and quality score for the predicted bases and is the standardized format used by all modern short aligners.

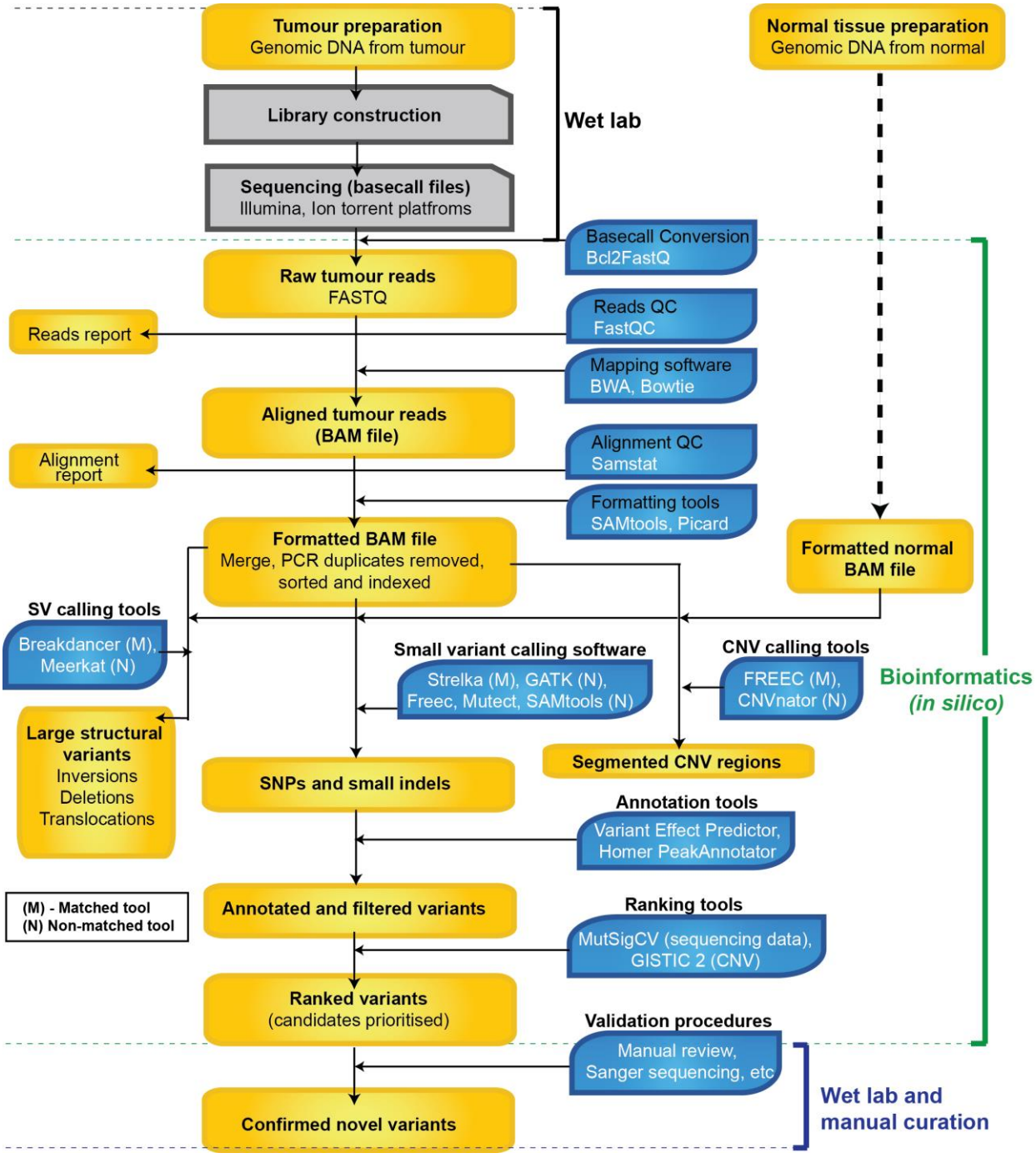

Figure 3: Workflow of DNA-based Next-Generation Sequencing

Before using the sequence reads, QC is performed to detect the proportion of low-quality bases, i.e. to evaluate suitable of the dataset. The quality of called bases is evaluated via a Phred score (B Ewing et al. 1998; Brent Ewing and Green 1998), which is a measure of the quality of the identification of the bases. This valuation will then direct whether reads should be filtered out entirely or trimmed, so as to avoid the use of low-quality segments. There are several tools available for this task (Babraham/Bioinformatics 2010; T. Zhang et al. 2011; Martínez-Alcántara et al. 2009).

Sequence alignment is the method of arranging the sequences of DNA, produced by the sequencing procedure, to the corresponding region in the genome, based on nucleic acid arrangement similarity. The actual alignment process involves software packages that use algorithms that are optimized for short reads alignments such as the Burrows–Wheeler transform (BWT) as used in BWA (H. Li and Durbin 2009) and Bowtie (Langmead et al. 2009), i.e. methods that align millions to billions of short segments of DNA to a longer reference in a way that allows for mismatches and gaps that would naturally occur due to SNV and indels. The practical output of these processes is an aligned reads file in the either the bam or sam formats (H. Li et al. 2009).

1.7.2. Variant calling

As described in section 1.3, DNA mutations can be broadly divided into four categories, namely, single nucleotide variants (SNV), insertion and deletions (indel), structural variants (SV) and finally copy number variants (CNV). The different characteristics of these mutations may require different tools for their discovery. The two variant types used as part of this thesis work are SNVs and indels. Although SNV and indels are different biological events, in practical terms both these events are often derived by the same software packages that specialise in small variants detection. These packages can be divided into two groups, germline variant callers and somatic variant

callers. The prior group details variant callers that compare the aligned sam/bam file to a reference genome and thus reveals both germline variants (differences to the reference genome, most often SNPs) and somatic variants that may be associated with a diseased state, if existent. Somatic callers on the hand specifically derive somatic mutations from a sample with a diseased state (e.g. cancer) by comparison against both a matched nondiseased sample and the reference genome. In this way, mutations present in both the sample and matched nondiseased will be ignored, thus leaving only true somatic mutations.

GATK UnifiedGenotyper (McKenna et al. 2010; DePristo et al. 2011) is one of the most popular germline callers that identifies both SNVs and indels by the use of a bayesian genotype likelihood model to simultaneously estimate the most likely genotypes and allele frequencies. Another popular caller is SAMtools mpileup (H. Li et al. 2009), which uses hidden Markov models (HMM) to estimate the mutations. Somatic callers use joint probability-based statistical approaches to determine the true somatic variants. Three of the most commonly used somatic caller are Strelka (Saunders et al. 2012), MuTect (Cibulskis et al. 2013) and VarScan 2 somatic (Koboldt et al. 2012). It should be noted that the TCGA studies use both MuTect and Strelka to derive mutational data and prior to that, BWA for alignment, while the ICGC project uses a variety of tools, depending on the different centres involved. Both databases use somatic variant caller to derive cancer-specific somatic mutations.

Figure 4 explains several variants calling processes in detail, as well as comparing the possible outcome from somatic and germline variant callers. The top panel represents the alignment seen in a normal (matched) sample within a specific region of DNA. The bottom represents alignment in a cancer sample within the same genomic coordinate region of DNA. The reference sequence is shown between these two

alignments. This figure diagrammatically represents the concepts behind variant calling from aligned reads. Some examples highlight how comparing cancer to a match normal by the use of a somatic variant caller can have superior outcomes to just using the tumour sample with a germline caller. The numbers below corresponds to the numbering in Figure 4.

1) Both the normal and tumour have more than one nucleotide (A and T) aligned to this position. This would suggest that this nucleotide position is heterozygous, therefore a somatic variant caller would not report a variant from this position, a germline caller, however, would report this as a potential A > T variant.

2) The heterozygosity seen in the normal sample (C and G) is not observed in the tumour samples. A somatic caller such as VarScan would report this as a loss of heterozygosity, but a germline caller would report this as just a variant.

3) A deletion (GGT) is identified where a sequence gap is detected in the aligned reads. The observance of these gaps is dependent on the gap tolerance of the alignment algorithm.

4) Insertions (AG) are detected when nucleotides on the aligned read do not have counterparts on the reference genome.

5) The variant allele C is a minor constituent of the bases in both the normal and tumour, a matched caller is unlikely to call this a variant, however, a germline may call it, depending on the variant calling thresholds.

6) This is a high-confidence variant as all the aligned nucleotides in the tumour represent a change (A > G) while all the normal nucleotides are the same as the reference.

7) At this position, a somatic caller would rank this (C > A) as a low confidence variant as the A > C mutation is seen in both the normal and tumour, but at a slightly

higher frequency in the tumour. A germline variant caller would simply rank this a high-confidence variant.

8) This variant (G > T) may not be called or may be called as a low confidence variant whether using a Somatic or germline caller, depending on the variant calling threshold, as there are very few bases supporting the variant.

1.7.3. Annotation of Variants

The implications of mutation data is largely uninformative without greater analysis and therefore have to be annotated to be meaningful. The annotation process includes assigning details such as the affected gene, the region of the gene that has been affected (exon, intron, intergenic etc.), whether or not the variant has been found previously, done found by referencing databases such as COSMIC (S. A. Forbes et al. 2015) and 1000 genomes project database (Abecasis et al. 2012) and many more potentially beneficial details. Variant Effect Predictor (McLaren et al. 2010) is an annotation approach that references the latest Ensembl database and is provided as both an online tool, as well as a downloadable Perl script. SnpEff (Cingolani et al. 2012) and Oncotator (Ramos et al. 2015) reference a list of databases to annotated variants in a similar manner to VEP. Homer Peakfinder (Heinz et al. 2010) is an annotation tool that was actually intended for DNA motif finding, and although less comprehensive than VEP and SnpEff it is significantly faster and as such, desirable for basic annotation. The TCGA MAF file is a standardised format for listing annotated variants used by the TCGA database and be generated by Oncotator and is used as part of this thesis.

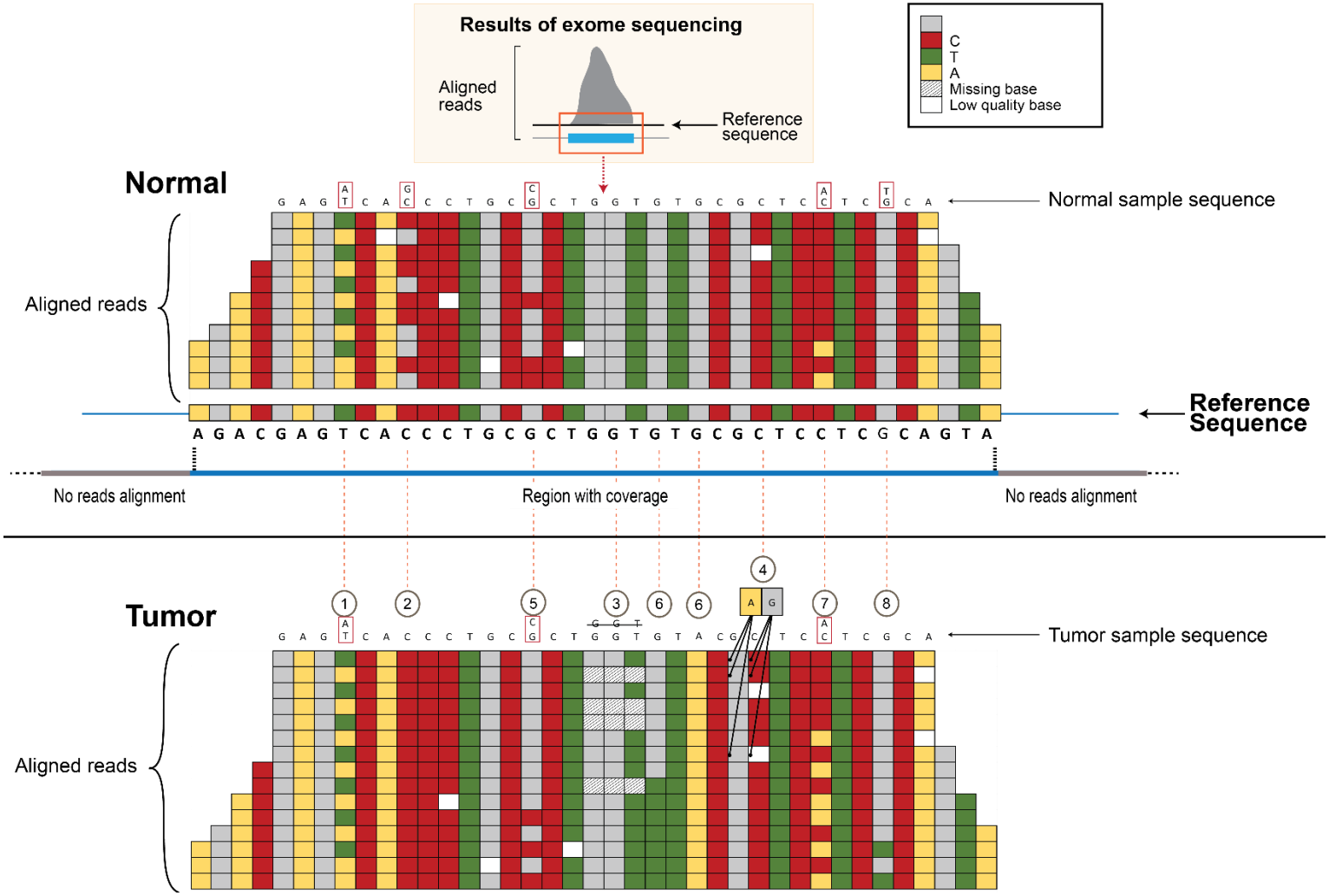

Figure 4: Comparing Somatic and Germline SNV and indel Variant Callers

1.8. Machine Learning: Application in biology

The term machine learning (ML) refers to a set of topics dealing with the creation and evaluation of algorithms that facilitate pattern recognition, classification, and prediction, based on models derived from existing data (Tarca et al. 2007). In a practical sense, it can be thought of as the process of utilising algorithms in a computing environment to solve a problem. The choice of algorithm is based on several considerations, most importantly the question/problem at hand, the type of data being used, the time frame available to derive the results and the computational processing infrastructure available. The underlying principal of ML is to use a technique that is able to carry out pattern matching and/or predict results on a given dataset and thus reveal underlying truths that are not initially observable. The advantage of using machine learning techniques is that once properly implemented, models can be evaluated and iterated over several modified parameters until an optimised solution is found (Heffernan et al. 2015; Demirci et al. 2016; Cheng et al. 2015). ML has in previous years been utilized in a myriad of application (Angermueller et al. 2016) such as natural language interpretation (Burger et al. 2016; Napolitano et al. 2016), weather and pollution prediction (Pandey, Zhang, and Jian 2013; Liu et al. 2016), host-pathogen interactions (Sen, Nayak, and De 2016) among many other applications.

Machine learning can be divided and subdivided into several categories, but perhaps the most practical considering when approaching machine learning is to understand whether the analysis process requires a supervised or unsupervised learning approach. Broadly, supervised learning involves classification of data based on labelled categories and there is a pre-existing assumption that the data can, in some way be grouped based on the labelled categories. Unsupervised learning is an unbiased approach to classification, where inferences are made from unlabeled datasets where

no pre-existing assumptions are made about how the data should be sorted. A simple example related to biology would be as follows. Assuming there are several normal and tumour samples that have undergone sequencing and variant calling. A supervised approach would be to separate the samples into normal and tumour samples and then determine which variants are associated with tumour status. An unsupervised approach would be to analyse all variants, and then determine if the variants are associated with tumour status, or any other classifier such as age, sex or ethnicity etc. Another consideration is the data type, i.e. if data variables are continuous, categorical, binary or even a mixture of multiple data types because different ML algorithms are optimised for different data types and dynamic ranges.

Recent publications have highlighted how machine learning methodologies have been utilised to help solve biological problems (Tarca et al. 2007; Kourou et al. 2015). For example, a few ML techniques, including feature extraction, which is a subset of dimensionality reduction, supervised learning, adaptive boosting, and random forest classifiers have been used for microscopic image analysis (Sommer and Gerlich 2013). The microscopy images are converted into a data representations suitable for machine learning, and then various state-of-the-art machine-learning algorithms are introduced. Applications include the detection of cell bounders, areas of tumour tissue as opposed to non-tumour tissue, cell type detection and fluorescence intensity levels.

Hidden Markov models (HMM) have been used extensively for the detection of copy number variations (Seiser and Innocenti 2015) by programs such as PennCNV. HMMs are full probabilistic models that function to determine an unknown sequence of states based upon a sequence of observations. Markov models model stochastic processes in which known sequences are produced from a finite number of discrete

states, where each new state of a sequence is only dependent upon the previous state, with the copy number state being of interest for these software packages.

ML approaches have even been used for cancer prognosis and prediction (Kourou et al. 2015). Examples of susceptibility prediction are in breast cancer by the analysis of mammograms using artificial neural networks (ANN) (Ayer et al. 2010) and in colon cancer by the analysis of clinical and pathological data by Bayesian Networks (BN) (Stojadinovic et al. 2011). Some examples of cancer recurrence prediction after remission, include in breast cancer by the analysis of clinical, pathologic, epidemiologic data by Support Vector Machine (SVM) (W. Kim et al. 2012) and oral cancer by clinical and imaging data using BN (Exarchos, Goletsis, and Fotiadis 2012). Lastly, examples of survival prediction include using ANN in lung cancers to analyse clinical data and gene expression (Chen, Ke, and Chiu 2014) and decision tree classifiers for the analysis of breast cancer surveillance, epidemiology and end results data (J. Kim and Shin 2013). The work in chapter 3 attempts to use several machine learning algorithms including several considered to be the most influential in data mining (Wu et al. 2008) to attempt to make a site of origin prediction of cancers based on NGS results.

1.9. Hypotheses and Aims

Recently, there has been interest in the study of DNA mutational change profiles of cancers, as these potentially allow classification of the diseases according to carcinogen and treatment categories. In 2007, Greenman et al. reported that the eight cancers types included in the study were associated with distinct SNV patterns (Greenman et al. 2007). In 2013, a more comprehensive study established that 21 trinucleotide (context) mutational signatures are associated with clinical features and potential carcinogen exposure (Alexandrov, Nik-Zainal, Wedge, Aparicio, et al. 2013). The inclusion of a greater number of samples and cancer types has recently increased

the signature count to 30 (Wellcome Trust). Distinctive signatures of indel mutations have also been observed in different cancers and may vary in both length and frequency (Yang et al. 2010). A recent analysis of six cancers showed that mutations may not be distributed uniformly throughout the genome, but instead the genomic density of variants according to chromosomal location can vary by up to fivefold depending on cancer subtype (Polak et al. 2015).

These studies in themselves have revealed unique features of different cancer types, however, each study only investigates one of the aspects of mutational changes. The study of indels and genomic densities have also used very small sample sizes with few cancer types. An integrated and comprehensive analysis of all the different mutational features or dimensions is lacking. Furthermore, an assessment of the exclusivity of these mutational signatures to specific cancer types and whether these signatures are capable of definitely distinguishing any cancer subtypes is also unknown.

The aim of the work presented is to develop a methodology that can distinguish the subtype/site of origin of cancers through an integrated and comprehensive analysis of the hitherto identified mutational dimensions (trinucleotide mutations, indel, variants, genes and mutational genomic densities) and to then use identified mutational signatures to develop a prediction approach capable of reliably determining the subtype of a given cancer.

This study works on the hypothesis that cancer subtypes, in part, have distinguishing mutational signatures that can be revealed at one or more of the mutational dimensions and that these signatures are sufficiently distinct from each other to establish a framework that is necessary to determine the specific cancer subtype of a patient based exclusively on mutational signatures.

In chapter 2, subtype distinct mutational signatures are revealed via a multidimensional analysis of trinucleotide mutations, indels, mutated genes, individual variants and genomic densities of mutations found in the 31 cancer subtypes available from the TCGA database. Firstly, the interrelatedness of the cancer subtypes are determined via a comparison of the consensus mutational signatures of these different cancer subtypes, secondly, the full spectrum of mutational signatures are determined by the analysis mutational signatures of all individual patients.

In chapter 3, a methodology was developed to summarise the nucleotide change signatures of cancer samples of unknown cancer types and then compare these signatures to the mutational signatures of known cancers from the TCGA database, and subsequently make a reliable cancer subtype/site of origin prediction, by the implementation of an optimised machine learning algorithm. As part of this methodology, MutProfiler was created, a tool that allows users to carry out the prediction by the analysis of NGS results in the form of a MAF file. The hope is that this tool will supplement existing techniques for site/tissue of origin prediction especially as NGS analyses continue to transition from a pure research tool to standardised clinical applications (Shen et al. 2015).

Chapter II

Multidimensional Mutation profiles of different cancers

2.1. Introduction

In this chapter, the mutation data from all available cases in the TCGA are analysed as 10 different mutational aspects, called dimensions. Firstly a consensus version of each cancer is created, then, all the cancer consensus versions are compared to each other, separately according to 10 dimensions. This analysis reveals the relatedness of the different cancers, thus identifying overall if similar mutation mechanisms exist among the different cancer types.

Secondly, the mutation profiles of individual cases are compared to each other, also according to the 10 dimensions. This analysis focuses of trying to determine if cases within certain cancers have defining mutation characteristics. This part of the chapter sets the ground work to investigate if it is possible to identify cancer specific mutational profiles and may have practical applications in the study of CTCs or CUPs.

2.2. Methods

2.2.1. Mutation reference (TCGA database)

Annotated somatic variants from WES of 31 cancers available in the TCGA Data Portal were downloaded as mutation alignment format (MAF) files (TCGA 2013) (The Cancer Genome Atlas 2013) (Table 1) from all available studies from the data portal. The most recent download was performed on 10 June 2015. The cancer types' abbreviations, as used in Table 1, are used throughout this thesis e.g. BLCA for Bladder urothelial carcinoma. Prior to variant calling, bioinformatics was carried out using Broad Institute's Cancer Genome Analysis (CGA) bioinformatics tools managed by the Firehose pipeline (CGA 2013). Library preparation kits include the Agilent SureSelect Human All Exon 5Mb kit, NimbleGen CCDS Solution Probes and NimbleGen SeqCap EZ Exome 2.0 Solution. Alignment tools include BFAST (Homer, Merriman, and

Nelson 2009), MAQ (H. Li, Ruan, and Durbin 2008) and BWA. Variant calling was performed using a combination of MuTect, Strelka and indel locator (Chapman et al. 2011) depending on the study. Annotation and MAF file generation was performed using Oncotator in several of the studies. Different versions of the variant analyses were available from the different studies and these were united into a combined MAF file using python scripts preventing variant and case duplications. Before mutational analysis was performed, cases from colon adenocarcinomas (COAD), stomach adenocarcinomas (STAD) and uterine corpus endometrial carcinomas (UCEC) were segregated into, either non-MSI-high (NonMSIH), comprising microsatellite stable (MSS) and MSI-low cases, or MSI-high (MSIH) as determined by the MSI Analysis System (Promega).

Table 1: Cancer types for which data was obtained from the TCGA database with associated descriptors

Cancer full name	Abbreviation	Cases	MSI-high	Cancer cell type	Primary site	Primary Germ layer	Updated	Downloaded
Adrenocortical carcinoma	ACC	91		Carcinoma	Adrenal gland	Ectoderm	07/02/15	10/06/2015
Bladder urothelial carcinoma	BLCA	412		Carcinoma	Bladder	Endoderm	07/02/15	10/06/2015
Breast invasive carcinoma	BRCA	988		Carcinoma	Breast	Mesoderm	07/02/15	10/06/2015
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	198		Carcinoma	Cervix	Mesoderm	06/29/15	10/06/2015
Cholangiocarcinoma	CHOL	36		Carcinoma	Bile duct	Endoderm	06/26/15	10/06/2015
Colon adenocarcinoma	COAD	269	52	Carcinoma	Colon	Endoderm	07/02/15	10/06/2015
Esophageal carcinoma	ESCA	183	2	Carcinoma	Oesophagus	Endoderm	07/02/15	10/06/2015
Glioblastoma multiforme	GBM	291		Blastoma	Brain	Ectoderm	07/02/15	10/06/2015
Head and neck squamous cell carcinoma	HNSC	526		Carcinoma	Head and neck	Endoderm	06/26/15	10/06/2015
Kidney chromophobe	KICH	66		Carcinoma	Kidney	Mesoderm	07/02/15	10/06/2015
Kidney renal clear cell carcinoma	KIRC	451		Carcinoma	Kidney	Mesoderm	06/26/15	10/06/2015
Kidney renal papillary cell carcinoma	KIRP	169		Carcinoma	Kidney	Mesoderm	06/26/15	10/06/2015
Acute myeloid leukemia	LAML	192		leukaemia	White blood cells	Mesoderm	04/29/15	10/06/2015
Brain lower grade glioma	LGG	515		Glioma	Brain	Ectoderm	07/02/15	10/06/2015
Liver hepatocellular carcinoma	LIHC	202		Carcinoma	Liver	Endoderm	06/30/15	10/06/2015
Lung adenocarcinoma	LUAD	546		Carcinoma	Lung	Endoderm	07/02/15	10/06/2015
Lung squamous cell carcinoma	LUSC	177		Carcinoma	Lung	Endoderm	06/26/15	10/06/2015
Ovarian serous cystadenocarcinoma	OV	463		Carcinoma	Ovary	Mesoderm	06/26/15	10/06/2015
Pancreatic adenocarcinoma	PAAD	178		Carcinoma	Pancreas	Endoderm	07/02/15	10/06/2015
Pheochromocytoma and paraganglioma	PCPG	179		Glioma	Adrenal gland	Ectoderm	06/26/15	10/06/2015
Prostate adenocarcinoma	PRAD	425		Carcinoma	Prostate	Endoderm	07/02/15	10/06/2015
Rectum adenocarcinoma	READ	115	3	Carcinoma	Rectum	Endoderm	06/26/15	10/06/2015
Sarcoma	SARC	258		Sarcoma	Unknown	Mesoderm	07/02/15	10/06/2015
Skin cutaneous melanoma	SKCM	370		Melanoma	Skin	Ectoderm	07/02/15	10/06/2015
Stomach adenocarcinoma	STAD	421	77	Carcinoma	Stomach	Endoderm	07/03/15	10/06/2015
Testicular germ cell tumours	TGCT	150		Germ cell tumour	Testes	Mesoderm	07/02/15	10/06/2015
Thyroid carcinoma	THCA	441		Carcinoma	Thyroid	Endoderm	07/02/15	10/06/2015
Thymoma	THYM	123		Carcinoma	Thymus	Endoderm	06/26/15	10/06/2015
Uterine corpus endometrial carcinoma	UCEC	248	71	Carcinoma	Uterus	Mesoderm	06/26/15	10/06/2015
Uterine carcinosarcoma	UCS	57	2	Sarcoma	Uterus	Mesoderm	06/26/15	10/06/2015
Uveal melanoma	UVM	80		Melanoma	Eye	Ectoderm	06/29/15	10/06/2015

2.2.2. Construction of mutation type datasets

Mutations downloaded from the TCGA database were annotated according to 8 single dimension and 2 multidimensional mutation types as follows:

1. **Trinucleotide mutations (counts)**: Counts of the trinucleotide mutations, i.e. SNVs studied in the context of flanking bases.
2. **Trinucleotide mutations (proportions)**: Proportions of the respective types of trinucleotide mutations i.e. normalised to a sum of 1 for all categories
3. **Indel mutations (counts)**: Counts of the different indels lengths ranging from 1 to greater than 5
4. **Indel mutations (proportions)**: Proportions of the respective indels lengths ranging from 1 to greater than 5, then normalised to a sum of 1 for all categories
5. **Mutated genes**: Genes with coding mutations i.e. binary classifier with 1 corresponding a mutation in a given gene and 0 indicating indication no mutation
6. **Recurrent variants**: Coding variants with a frequency of 4 or more in the entire dataset, designated by 1 indicating a mutation and 0 indicating the wildtype allele
7. **Genomic distribution (counts)**: The counts of mutations per megabase (Mb) in the genome
8. **Genomic distribution (proportions)**: The proportions of mutations per each megabase (Mb) in the genome, normalised to a sum of 1 for all categories
9. **Multidimensional (counts)**: A combination of all mutational dimensions using counts for trinucleotide mutations, indel sizes and genomic distributions.
10. **Multidimensional (proportions)**: A combination of all mutational dimensions using proportions for trinucleotide mutations, indel sizes and genomic distributions.

Methods concerning the various dimensions are described in the following sections. The mutational signature analyses and summary statistics described in this chapter were performed using an integrated pipeline created in the Python programming language v3.4.1 using py-postgresql v1.1.0 for interfacing with PostgreSQL, Pandas v0.16.2 and Numpy v1.9.2 for data manipulation, Scipy v0.14.0 for hierarchical

clustering and statistical analysis, Matplotlib v1.4.3 for visualization and Statsmodels v0.6.1 for false discovery rate (FDR) and familywise error rate (FWER) multiple testing.

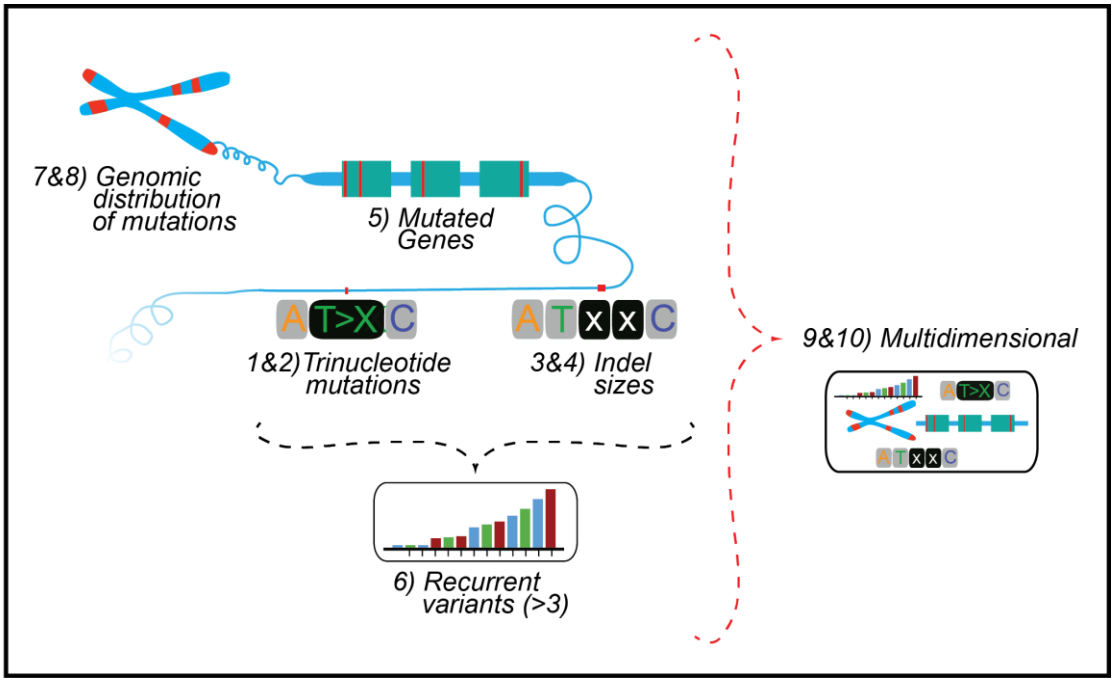

Figure 5: Analysis of the many dimensions of DNA mutations may elucidate cancer specific characteristics

The studied mutations involve a broad range of mutational phenomena, called dimensions, representing regions of the genome ranging in several orders of magnitude. As seen in Figure 5, the dimensions are, specific single base SNVs, studied as trinucleotides (1&2), indels, which involve 1 or more nucleotides (3&4), the entirety of coding mutation within genes, which may include regions of up to a hundred thousand nucleotides (5), the recurrence of variants (6) and the distribution of these mutations throughout entire chromosomes within 1 Mb (1 million base) bins.

2.2.2.1. Annotations of trinucleotide mutations

There are 12 possible SNVs (A>C, A>G, A>T, C>A, C>G, C>T, G>A, G>C, G>T, T>A, T>C, T>G). When studied in the context of flanking sequences, there are 192 possible trinucleotide changes, namely four possible 5' bases × 12 possible SNVs × four possible 3' bases. The number of trinucleotide changes can be reduced to 96 (192 ÷ 2) by assimilating reverse complement pairs (Table 2). To derive the counts of the 96 possible trinucleotide changes, the flanking bases (nucleotides 5' and 3' of the mutation site) of each SNV were determined by comparison to a PostgreSQL database composed of the entire genome sequences from UCSC genome builds hg18 (Assembly date: Mar. 2006), hg19 (Assembly date: Dec. 2013, GenBank accession ID: GCA_000001405.1) and hg38 (Assembly date: Dec. 2013, GenBank accession ID: GCA_000001305.2) (Kent et al. 2002). The proportion of each trinucleotide change was then determined by dividing the counts for each change by the total counts of all trinucleotide changes.

Table 2: All possible trinucleotide mutations arranged by SNV

Mutation	Trinucleotide primary annotation	Trinucleotide reverse complement		Mutation	Trinucleotide primary annotation	Trinucleotide reverse complement		Mutation	Trinucleotide primary annotation	Trinucleotide reverse complement
C>A	ACA>AAA	TGT>TTT		**C>T**	ACA>ATA	TGT>TAT		**T>C**	ATA>ACA	TAT>TGT
	ACC>AAC	GGT>GTT			ACC>ATC	GGT>GAT			ATC>ACC	GAT>GGT
	ACG>AAG	CGT>CTT			ACG>ATG	CGT>CAT			ATG>ACG	CAT>CGT
	ACT>AAT	AGT>ATT			ACT>ATT	AGT>AAT			ATT>ACT	AAT>AGT
	CCA>CAA	TGG>TTG			CCA>CTA	TGG>TAG			CTA>CCA	TAG>TGG
	CCC>CAC	GGG>GTG			CCC>CTC	GGG>GAG			CTC>CCC	GAG>GGG
	CCG>CAG	CGG>CTG			CCG>CTG	CGG>CAG			CTG>CCG	CAG>CGG
	CCT>CAT	AGG>ATG			CCT>CTT	AGG>AAG			CTT>CCT	AAG>AGG
	GCA>GAA	TGC>TTC			GCA>GTA	TGC>TAC			GTA>GCA	TAC>TGC
	GCC>GAC	GGC>GTC			GCC>GTC	GGC>GAC			GTC>GCC	GAC>GGC
	GCG>GAG	CGC>CTC			GCG>GTG	CGC>CAC			GTG>GCG	CAC>CGC
	GCT>GAT	AGC>ATC			GCT>GTT	AGC>AAC			GTT>GCT	AAC>AGC
	TCA>TAA	TGA>TTA			TCA>TTA	TGA>TAA			TTA>TCA	TAA>TGA
	TCC>TAC	GGA>GTA			TCC>TTC	GGA>GAA			TTC>TCC	GAA>GGA
	TCG>TAG	CGA>CTA			TCG>TTG	CGA>CAA			TTG>TCG	CAA>CGA
	TCT>TAT	AGA>ATA			TCT>TTT	AGA>AAA			TTT>TCT	AAA>AGA
C>G	ACA>AGA	TGT>TCT		**T>A**	ATA>AAA	TAT>TTT		**T>G**	ATA>AGA	TAT>TCT
	ACC>AGC	GGT>GCT			ATC>AAC	GAT>GTT			ATC>AGC	GAT>GCT
	ACG>AGG	CGT>CCT			ATG>AAG	CAT>CTT			ATG>AGG	CAT>CCT
	ACT>AGT	AGT>ACT			ATT>AAT	AAT>ATT			ATT>AGT	AAT>ACT
	CCA>CGA	TGG>TCG			CTA>CAA	TAG>TTG			CTA>CGA	TAG>TCG
	CCC>CGC	GGG>GCG			CTC>CAC	GAG>GTG			CTC>CGC	GAG>GCG
	CCG>CGG	CGG>CCG			CTG>CAG	CAG>CTG			CTG>CGG	CAG>CCG
	CCT>CGT	AGG>ACG			CTT>CAT	AAG>ATG			CTT>CGT	AAG>ACG
	GCA>GGA	TGC>TCC			GTA>GAA	TAC>TTC			GTA>GGA	TAC>TCC
	GCC>GGC	GGC>GCC			GTC>GAC	GAC>GTC			GTC>GGC	GAC>GCC
	GCG>GGG	CGC>CCC			GTG>GAG	CAC>CTC			GTG>GGG	CAC>CCC
	GCT>GGT	AGC>ACC			GTT>GAT	AAC>ATC			GTT>GGT	AAC>ACC
	TCA>TGA	TGA>TCA			TTA>TAA	TAA>TTA			TTA>TGA	TAA>TCA
	TCC>TGC	GGA>GCA			TTC>TAC	GAA>GTA			TTC>TGC	GAA>GCA
	TCG>TGG	CGA>CCA			TTG>TAG	CAA>CTA			TTG>TGG	CAA>CCA
	TCT>TGT	AGA>ACA			TTT>TAT	AAA>ATA			TTT>TGT	AAA>ACA

2.2.2.2. Annotation of indels

The lengths of insertions were derived from the alternate (variant) sequence in accordance with standard MAF file classification, while the lengths of deletions were determined based on the reference sequence. Insertions and deletions were separately characterised into lengths of 1 to 5, and greater than 5. The proportion of each indel mutation was then determined by dividing the counts for each change by the sum of all indel mutations.

2.2.2.3. Annotation of genes and variants

Functional coding variants were defined with the appropriate MAF classifications of "Frame_Shift_Del", "Frame_Shift_Ins", "In_Frame_Del", "In_Frame_Ins", "Missense_Mutation", "Nonsense_Mutation", "Nonstop_Mutation", "Splice_Site", and "Translation_Start_Site". For annotation of mutated genes, a binary (boolean) matrix was generated to denote cases associated with genes that had a functional variant. For annotation of variants, a similar procedure was carried out for individual variants, denoted by both nucleotide position and the specific mutation, e.g. chr10:100017453-100017453:T>***G*** is different from chr10:100017453-100017453:T>***A***. In order to standardise the variants results between genome builds, coordinates in hg18 were converted to hg19 using the UCSC LiftOver tool (Rosenbloom et al. 2015), which utilises chain liftOver conversion files to convert between the different genome builds. Due to computational limitations, the variants matrix was limited to variants with at least 4 occurrences within the dataset.

2.2.2.4. Annotation of the genomic distribution of mutations

The genomic mutation frequencies per megabase (Mb) was determined by separately dividing the twenty-two autosomes and X and Y chromosomes into sequential 1 Mb segments (e.g. chromosome 1 nucleotide 1 to 1,000,000), and then annotating the amount of both SNVs and indels in each segment. This approach resulted in the creation of 3089 1Mb segments based on hg19. Mitochondrial mutations were not included in this analysis. As with the other dimensions, proportions within each category were determined by dividing the mutation frequencies in each segment by the total number of mutations in all segments.

2.2.2.5. Characterisation of cancers according to mutational signatures

To examine the relatedness of different cancers, a consensus version of each cancer subtype was created for each dimension. Trinucleotide changes, indels and genomic distributions were represented by medians. Genes mutated and variants were represented by mean values, as the median values for these dimensions were often zero. Unsupervised agglomerative hierarchical clustering was firstly performed on the consensus datasets to give an overall understanding of the inter-relatedness of the various cancers present in the TCGA database. This was followed by clustering of individual cases to provide an understanding of the inter-relatedness of the different cancers, the heterogeneity of each cancer type, and whether any nucleotide changes signatures were associated with the cancer site of origin.

To elucidate whether possible biological mechanisms or carcinogenic exposures were associated with the mutational signatures, several descriptive features were aligned with the hierarchical clustering. Age at diagnosis, gender, MSI status and overall survival information were obtained from the TCGA data portal (The Cancer Genome Atlas 2013). Embryological origins (Kimelman and Bjornson 2004; Pansky

1982), body/organ systems and cancer cell type were based on a literature review (Table 1 and Table 3).

The association of trinucleotide change distributions to Alexandrov signatures (Alexandrov and Stratton 2014) was performed by cosine similarity using the Alexandrov signatures available via the "Signatures of Mutational Processes in Human Cancer" website (Wellcome Trust Sanger Institute 2016b). Trinucleotide changes or indel sizes that were significantly different from others were determined using an independent samples t-test, adjusted for false discovery using the method of Benjamini & Hochberg (fdr_bh) (Benjamini and Hochberg 1995). Determination of statistically significant genomic regions performed similarly to the trinucleotide changes analysis, except that a Holm-Šídák multiple correction test was performed. Determination of differential DNase-seq expression was performed by the use of Student's one-sample T-test and fdr_bh for multiple correction.

Table 3: TCGA cancers with the organ system annotations

Abbreviation	Primary organ system*	blood	digestive	endocrine	female	integumentary	lymphatic	male	nervous	reproductive	respiratory	urinary
ACC	Endocrine			1								
BLCA	Urinary											1
BRCA	Female				1							
CESC	Female				1							
CHOL	Digestive		1									
COAD	Digestive		1									
ESCA	Digestive		1									
GBM	Brain								1			
HNSC	Respiratory										1	
KICH	Urinary											1
KIRC	Urinary											1
KIRP	Urinary											1
LAML	Blood	1										
LGG	Brain								1			
LIHC	Digestive		1									
LUAD	Respiratory										1	
LUSC	Respiratory										1	
OV	Female			1	1					1		
PAAD	Endocrine		1	1								
PCPG	Endocrine			1								
PRAD	Male							1		1		
READ	Digestive		1									
SARC	undetermined											
SKCM	Skin					1						
STAD	Digestive		1	1								
TGCT	Male			1				1				
THCA	Endocrine			1								
THYM	Immune	1		1			1					
UCEC	Female				1					1		
UCS	Female				1					1		
UVM	Eye								1			
1 – Organ is a component of this body system												

2.3. Results

2.3.1. TCGA dataset summary description

Table 1summarises details of the cancer types and data obtained from the TCGA after combining several versions of the MAF files. The number of cases (patients) per cancer ranged from 36 in cholangiocarcinoma (CHOL) to 988 in breast invasive carcinoma (BRCA). Table 6 shows the number of cases seen in the 25 different tissues of origin. The most common sites of cancer origin, as revealed by GLOBOCAN (Figure 1), including lung (LUAD and LUSC), breast (BRCA), colorectum (divided into colon (COAD) and rectum (READ)), prostate (PRAD), amongst others, are included in this dataset. Table 7 shows the data divided into the cell type of origin. The vast majority of cancer cases were of epithelial origin, as expected. 52 of 269 (19%) COAD, 77 of 421 (18%) STAD and 71 of 248 (29%) UCEC were MSI.

The SNV mutation rates from the 34 cancers are represented in Figure 6, revealing a large range of values among the SNVs. This observation was previously reported (Lawrence et al. 2013), indicating that several cancers such as melanoma (SKCM), lung adenocarcinoma (LUSC), lung squamous carcinoma (LUAD), bladder (BLCA), stomach (STAD) and colorectal (COAD) have comparatively much higher mutation rates than other cancers, while acute myeloid leukemia's have very low rates. For each cancer, the amount of spread of mutation rate was associated with sample size, i.e. within this dataset, there does not seem to be any association between the spread of mutation data and cancer type.

The TCGA dataset has grown since the 2013 publication reference above, in terms of the samples size and the specifics of the analysis pipeline. The mutation data has gone through curation and modification of the variant calling. Despite these changes, the general trends seen are similar. It should be noted that compared to the previous

study, here, segregation of the MSI and non-MSI cases has been performed, thus revealing MSI cases from uterine corpus endometrial carcinoma (UCEC), colon adenocarcinoma (COAD) and stomach adenocarcinoma (STAD) are in fact the cancers with the highest mutation rates when MSI is taken into account. This study is the first report of mutation rates in testicular germ cell tumours (TGCT), which are found to have comparatively normative mutation rates, with a median of 100.

In terms of the trinucleotide mutations, overall C > T mutations represent the vast majority of mutations within the entire dataset (Figure 7), especially occurring at CpG sites, a trend observed previously (Greenman et al. 2007). Mutations originating from C or G (C > T, C > G or C >T) appear to be significantly more likely than mutations for A or T (C > T, C > G or C >T).

Of 8820 cases in this study, 505 cases did not have any indels, this absence observed primarily in ovarian serous cystadenocarcinoma (OV), thyroid carcinoma (THCA) and acute myeloid leukaemia (LAML) where 38%, 35% and 24% of cases did not have indels (Table 4). As may be expected, the highest median number of indels was seen in the MSI cases (Figure 8), interestingly a high median number of deletions (82.5) was also seen in the pancreatic adenocarcinomas (PAAD) with only one case devoid of indels. The lung squamous cell carcinoma (LUSC) unlike the lung adenocarcinoma (LUAD) had very few indels, while both these cancer types had very high rates of SNV, suggesting that perhaps these tumour types may be distinguished by indel mechanisms. Deletions ranged in size from 1 to 197 bases, while insertions varied from 1 to 108 bases in length (Figure 9). The relative frequencies of larger indels are low compared to the smaller indels, therefore the frequency of deletions and insertions larger than 5 were aggregated into a single category for further analysis so as increase inference from these categories and to reduce computational overhead.

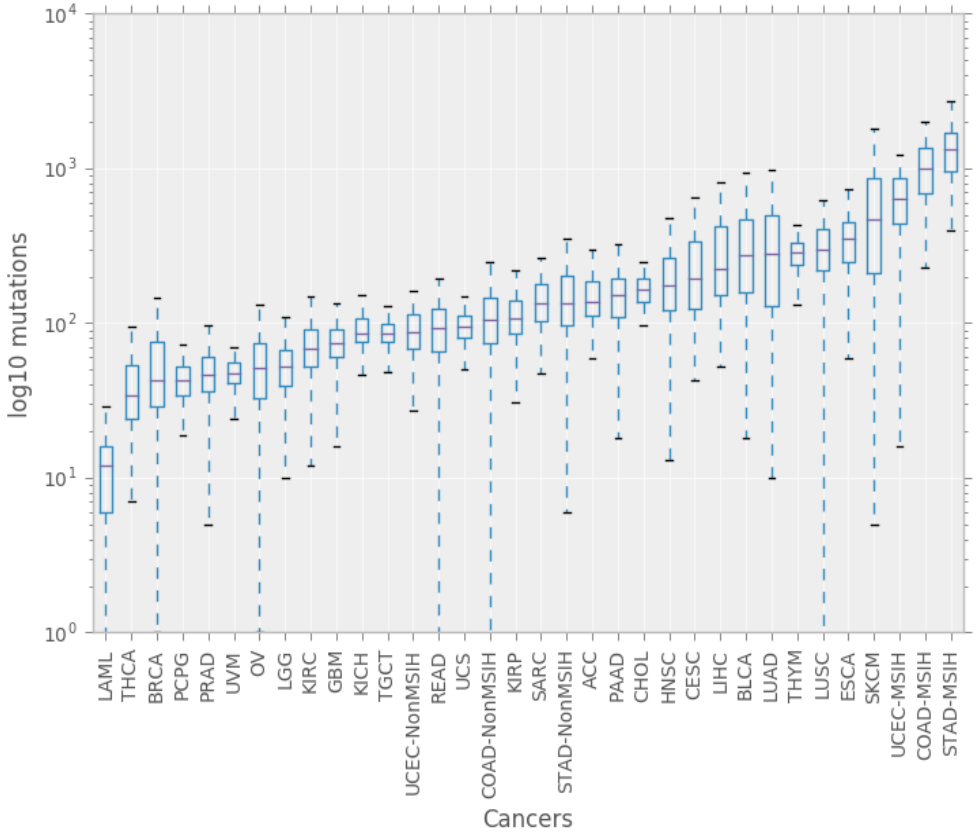

Figure 6: Distribution of SNV mutations rates for the 34 cancers used in this study

The cancers are listed along the x-axis, while the log10 transformed mutation rates are shown on the y-axis. The cancers are sorted by median mutation rate.

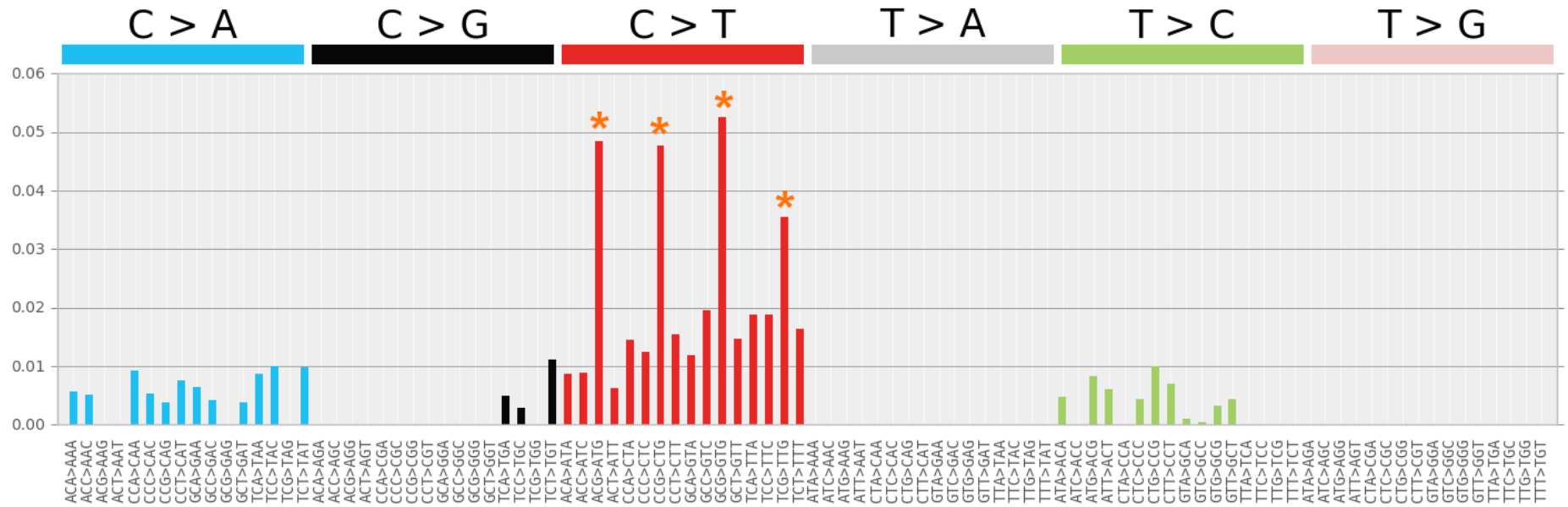

Figure 7: The proportions of the trinucleotide mutations across all cases in this study

The 96 possible nucleotide changes are represented across the x-axis, with the median proportions of each change across all cases represented in the y-axis. The coloured headings show the primary SNV type corresponding to the trinucleotide mutations. C>T mutations at a CpG are the most common type of mutation observed (indicated with *). C>G, T>A and T>G are rare events throughout the dataset.

Table 4: Insertion and deletion rates for all cancers

Cancer	Cases	Cases without deletions	Median insertions	Median deletions
ACC	91	0 (0%)	5	12
BLCA	412	2 (0%)	3	8
BRCA	988	21 (2%)	2	3.5
CESC	198	1 (1%)	3	5
CHOL	36	0 (0%)	5	10
COAD-MSIH	52	0 (0%)	53	145
COAD-NonMSIH	218	17 (8%)	3	4
ESCA	183	0 (0%)	6	11
GBM	291	13 (4%)	1	3
HNSC	526	1 (0%)	4	9
KICH	66	0 (0%)	3	5
KIRC	451	24 (5%)	3	7
KIRP	169	0 (0%)	3	10
LAML	197	47 (24%)	1	0
LGG	515	1 (0%)	2	5
LIHC	202	0 (0%)	6	15
LUAD	546	0 (0%)	4	11
LUSC	178	15 (8%)	1	4
OV	463	178 (38%)	0	1
PAAD	178	1 (1%)	22	82.5
PCPG	179	0 (0%)	1	4
PRAD	425	3 (1%)	2	4
READ	116	21 (18%)	3	3
SARC	258	0 (0%)	9	17
SKCM	370	1 (0%)	2	6
STAD-MSIH	77	0 (0%)	99	281
STAD-NonMSIH	344	2 (1%)	3	6
TGCT	150	0 (0%)	3	7
THCA	441	153 (35%)	0	1
THYM	123	0 (0%)	4	11
UCEC-MSIH	71	0 (0%)	25	80
UCEC-NonMSIH	177	1 (1%)	4	6
UCS	57	0 (0%)	5	11
UVM	80	3 (4%)	1	3

Ovarian serous cystadenocarcinoma (OV), thyroid carcinoma (THCA) and acute myeloid leukaemia (LAML) have high proportions of cases without indels. The MSI cases and pancreatic adenocarcinomas showed the greatest number of indels based on the overall median for all cases.

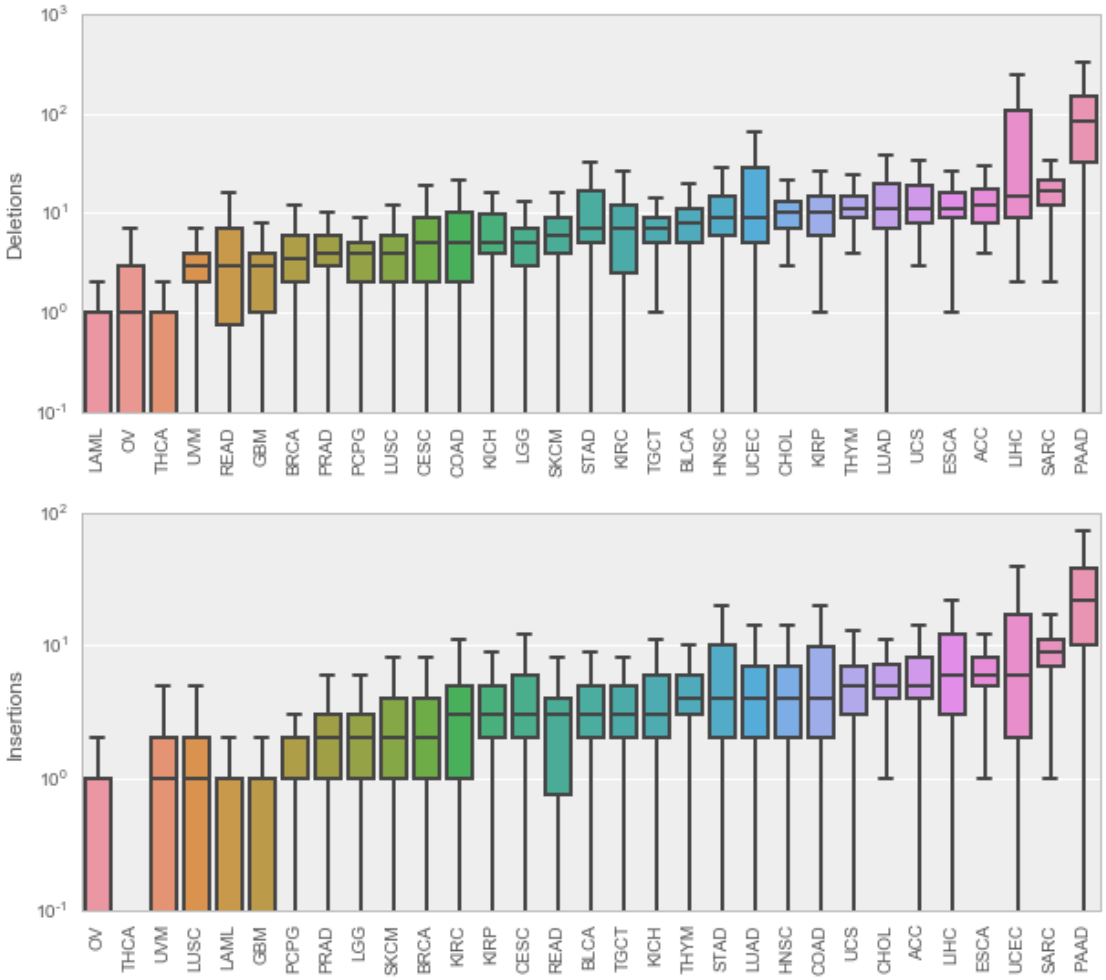

Figure 8: Distribution of indels rates for the 34 cancers used in this study

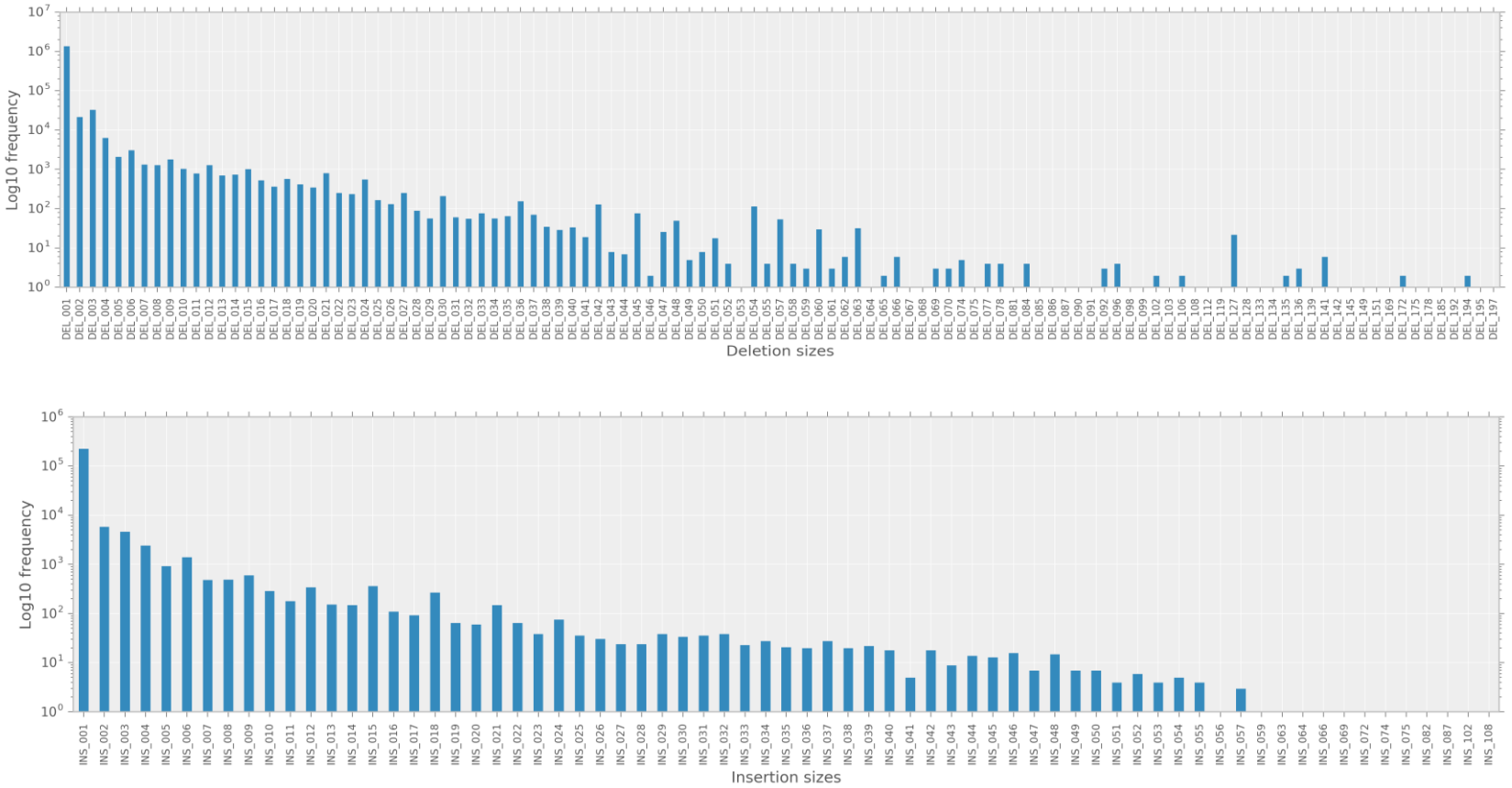

Figure 9: Distribution of the deletion and insertion sizes

The top panel shows the frequencies of deletions sizes present in all cases, while the bottom panel shows the frequencies of the insertion sizes. The x-axis corresponds to the varying indel sizes, e.g. INS_001, corresponds to an insertion of 1 base, while the y-axis represents the frequencies of the indels (log10 scale). Indels above the size of 5 occur at relatively low frequencies compared to smaller indels and therefore were aggregated into a separate category, i.e. deletions > 5 bases and insertions > 5 bases.

Of the 3089 1Mb bins studies as part of the genomic distributions, 209 were found to have no mutations across all cases. Chromosome Y particularly was shown to have low/no mutation from approximately base 29,000,000 to the end of the chromosome (Figure 10), as this region is a gene desert. Throughout the genome, the mean number of mutations (across cases) per region was highly variable, ranging from 0 to 2.9, as was the number of cases with mutations in each region.

The five most consistently mutated regions are shown in Figure 10. The region spanning chr17:7,000,001-8,000,000 was mutated in 4,817 cases within 60 different genes, inclusive of 25 phosphoproteins and 22 membrane proteins and also contains *TP53.* The region spanning chr5:140,000,001-141,000,000 is a region very rich in membrane proteins (50 genes mutated), specifically cadherin (45 mutated genes) a class of transmembrane proteins that serve as the major adhesion molecules located within cellular junctions and is mutated in 4714 cases. Overall, the two most highly mutated regions were chr1:152,000,001-153,000,000, with 25,298 mutations across all cases, a region with 25 mutated genes associated with keratinization and chr5:140,000,001-141,000,000, with 24,659 mutations, as described above. The most mutated region in a single case was also chr1:152,000,001-153,000,000 with 4738 variants altogether in a single case of liver hepatocellular carcinoma (LIHC).

Protein-coding mutations were found in 24,118 genes, as annotated by the TCGA database. Only 2214 (~9%) of the gene were found to be mutated only once, i.e. the remaining genes were recurrently mutated. Figure 11 shows the number of genes that fall into different frequency ranges (number of cases with that gene mutated). There are many genes which are recurrently mutated, however, genes which are mutated in greater than 500 cases are relatively rare (only 81 genes, 0.3% of all genes). By far the most mutated gene was *TP53* (3270 cases), followed by *TTN* (2836 case) and two mucin

genes, MUC16 and MUC4, mutated in 1839 and 1261 cases respectively. The most frequently mutated genes, along with the percentages of cases with mutations in the various cancers is shown in Figure 12. As has been discussed in the introduction, the rates of mutation in different genes vary greatly among the different cancers, specifically, in highly mutated cancer-related genes. For example, TP53 is mutated in only 1% of pheochromocytoma and paraganglioma (PCPG) and thyroid carcinoma (THCA) cases, but in as high as 91% of uterine corpus endometrial carcinoma (UCEC) cases, and is overall the most frequently mutated gene. PIK3CA is another gene highly mutated in UCECs at approximately 55% in both the MSI-high and MSS cases, but this gene has either no mutations or a very low mutation rate in most other cancers (Figure 12). As a whole it can be seen that even the most frequently mutated genes are not ubiquitously mutated in all cancers, and may in fact be a means to distinguish cancer types.

For the variant level analysis, it was computationally prohibitive to include all existing variants in the TCGA dataset as separate categories for comparison, additionally, the low frequencies are unlikely to substantively contribute to data when attempting to understand the interrelation of the different cancers, thus only variants with a frequency of four or more were analysed, resulting in the inclusion of 42030 unique variants. Figure 13 shows the number of variants that fall into different frequency ranges (number of cases with that variant) in a similar manner to Figure 11 for the mutated genes. Unlike the genes, most variants occur in the lowest frequency category (4 cases), the distinction being that the mutated genes actually represent the aggregation of several distinct variants into a single gene. Variants occurring at very high frequencies, i.e. in greater than 50 cases are rare, with only 107 variants (0.25% of variants) with such frequencies. Figure 14 shows the 10 most frequent variants found

across all cases. The *BRAF* V600E is the single most common cancer variant (5.3 % of cancers) and is known to be associated almost exclusively with melanoma (Ascierto et al. 2012), i.e. skin cutaneous melanoma (SKCM) in this dataset, and also colon (COAD) and thyroid carcinoma (THCA). The *IDH1* R132H was the second most frequent variant (4.4% of cases) and occurred in 70% of brain lower grade glioma (LGG) and is associated with a decrease in proliferation, decreased Akt phosphorylation, altered morphology, and a more contact-dependent cell migration (Bralten et al. 2011). Despite its exceedingly high frequencies in LGG, it is not seen in other cancer types. Although KRAS codon 12 mutations occur in many cancers at a very low frequency, the G12D and G12V are especially high in pancreatic cancers.

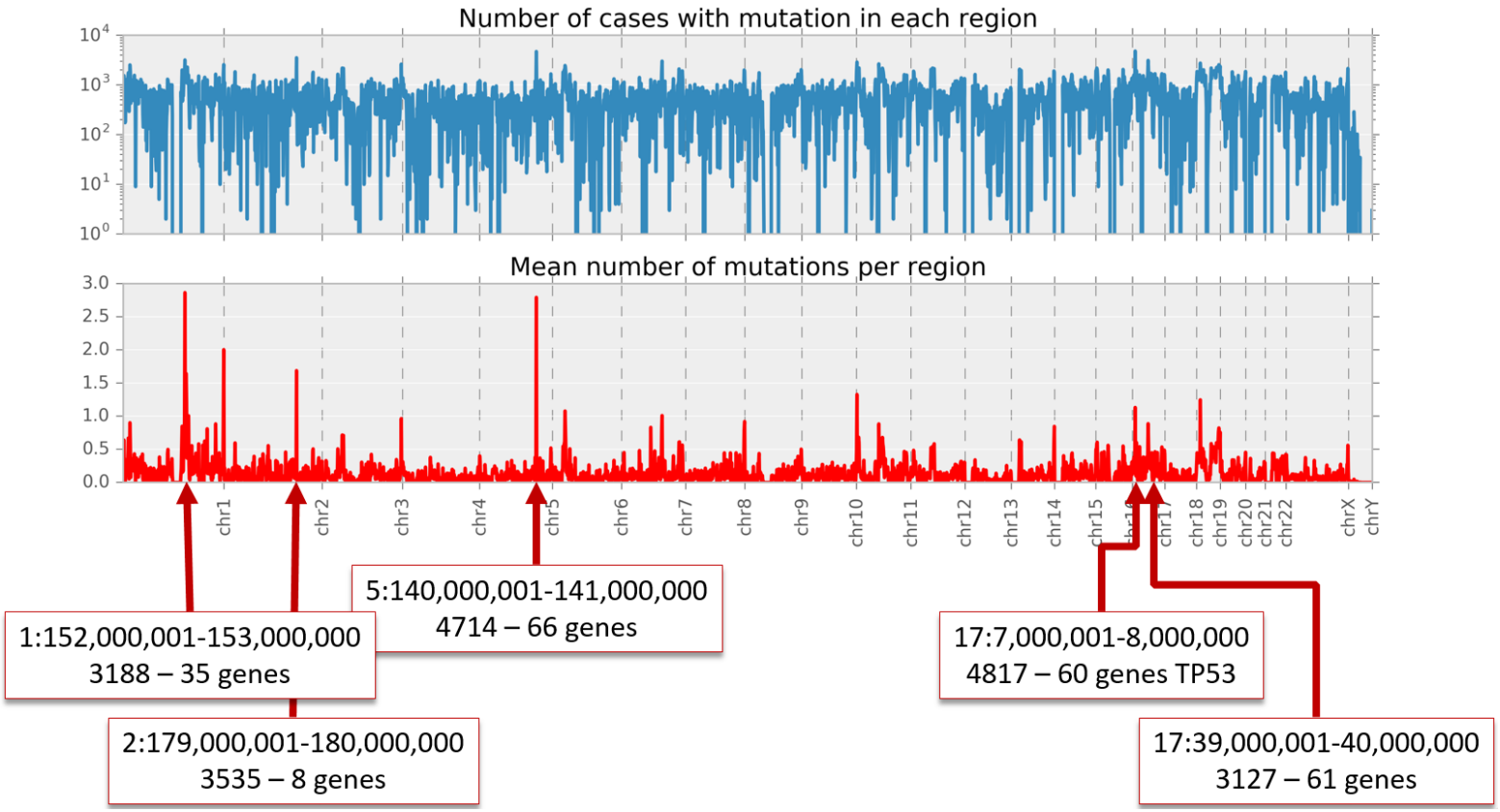

Figure 10: Distribution of the mutations across all samples

The top panel shows the number of cases with mutations in each of the 1 Mb regions of the genome, while the bottom panel shows the mean number of mutations per region across all samples. There is a great amount of variation in both the number of cases with mutations and also the mutation rates among the regions. Chromosome Y particularly is shown to have low/no mutation in many regions, a consequence of being gene deserts and therefore not covered by WES. The five most frequently mutated region as are indicated with text boxes. The leftmost number below the coordinates indicated the number of cases in the TCGA dataset that have a mutation in that region. Also indicated are the number of genes in that region with at least one mutation.

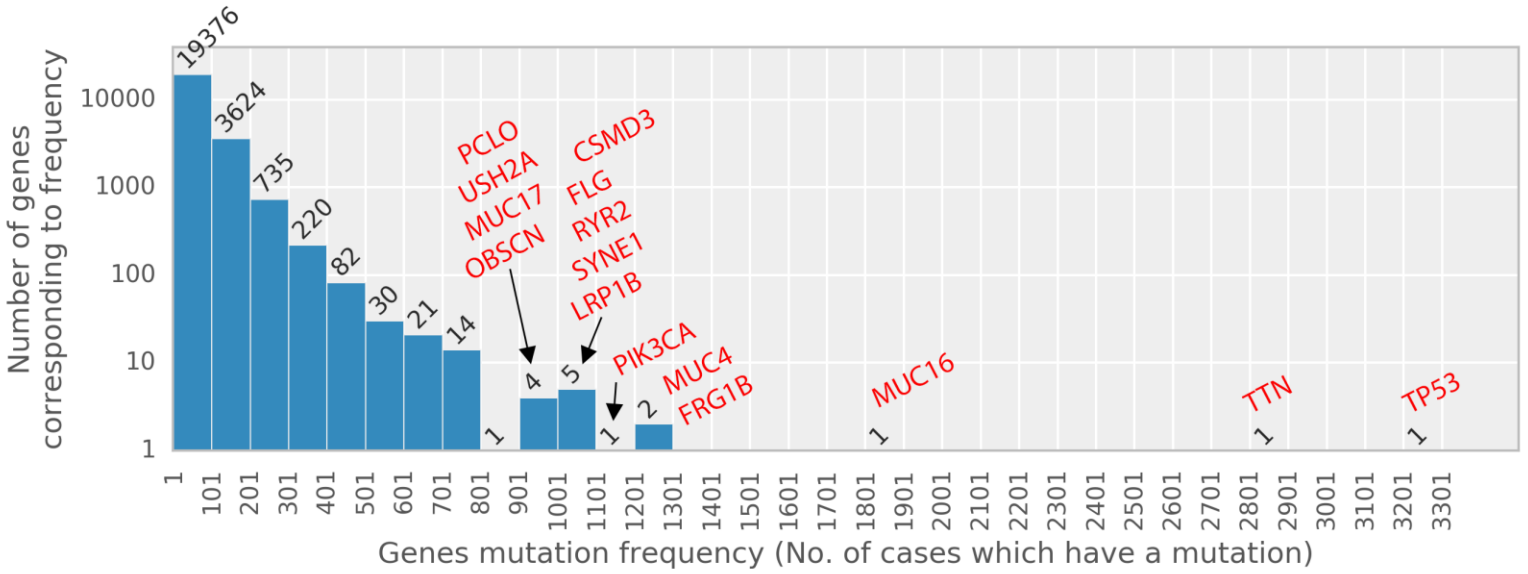

Figure 11: Distribution of the mutated gene frequencies

This histogram represents the number of genes that occur in mutation frequency ranges of 100. The numbers over the bar indicate the number of genes that fall into each range. As can be seen, there are approximately 19,000 genes that are mutated in anywhere from 1 to 100 cases, this falls to 3624 genes in 101 to 200 case. Mutation rates for genes above a frequency of 900 are very rare with the gene names indicated in red font.

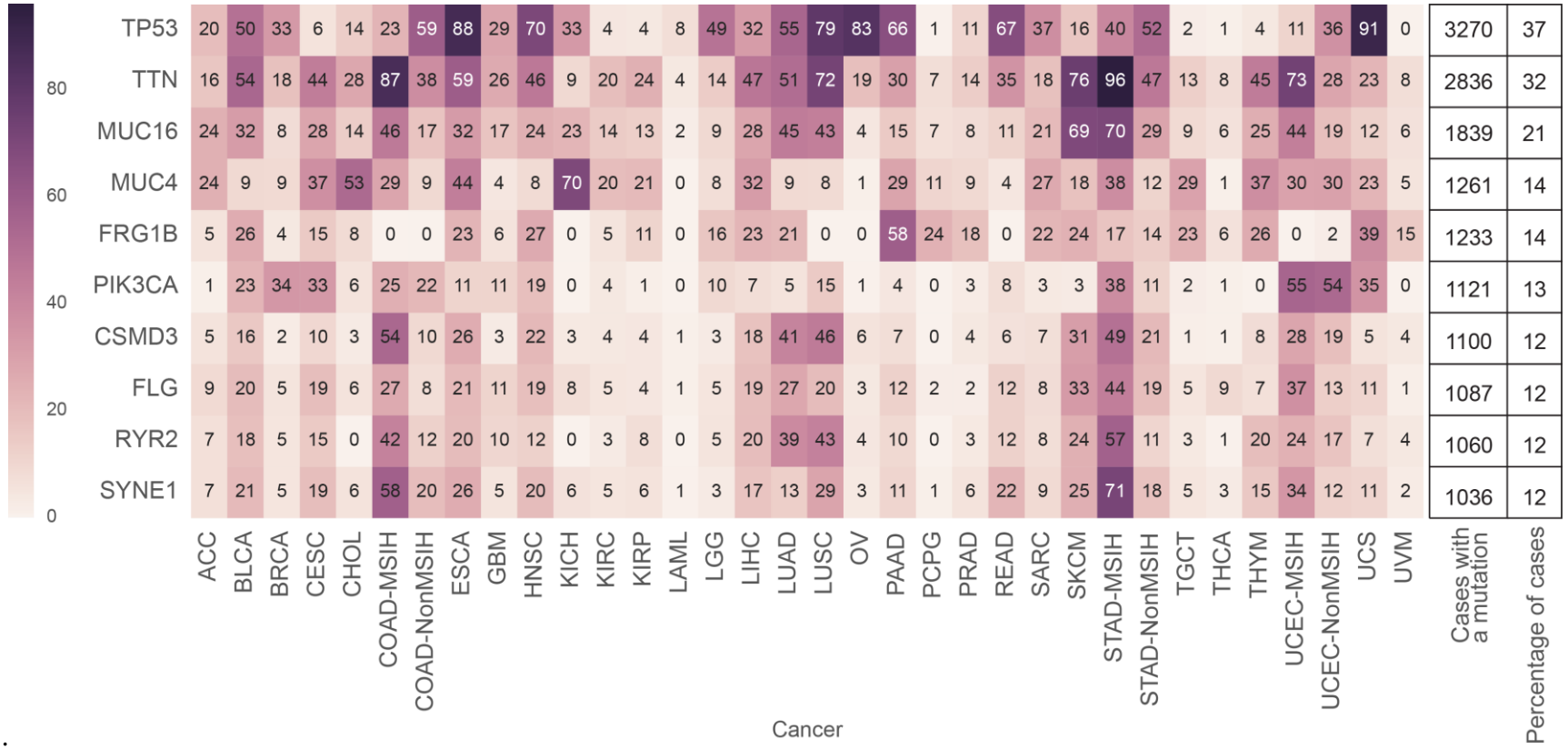

Figure 12: Percentage of cases with mutations in the 10 most mutated genes

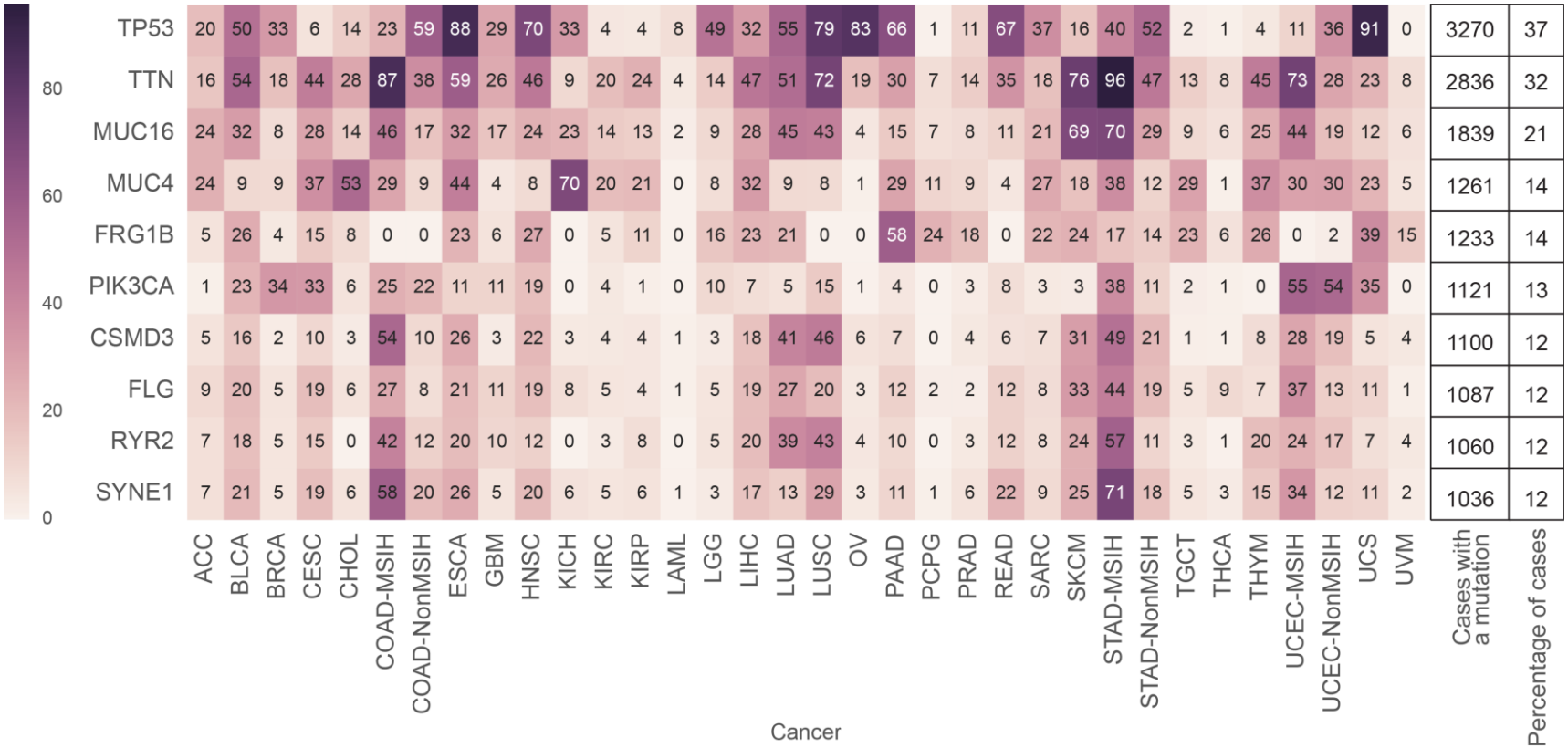

Figure 12: Percentage of cases with mutations in the 10 most mutated genes

This heatmap shows the 10 most consistently mutated genes (mutated in the highest number of cases) along with the percentages of cases in each cancer with a mutation in that genes (as indicated with the cancer abbreviation). The total number of cases with a mutation (across all cancers) is shown to the right of the heatmap ("**Cases with a mutation**"), as is the percentage of all cases ("**Percentage of cases**").

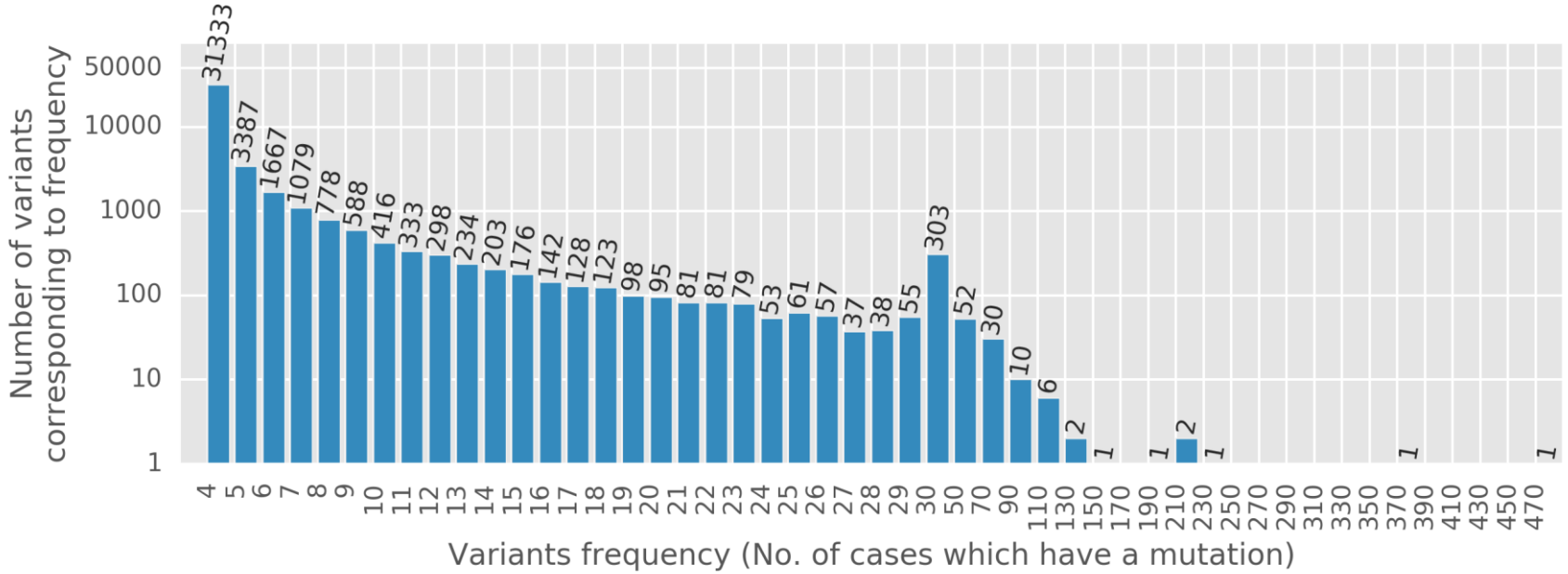

Figure 13: The size distribution of the variants across all samples

This histogram shows the number of variants at several intervals of frequencies. The first 30 size ranges are of 1 base, while variants at a frequency of greater than 30 are shown in ranges of 20 bases. Only variants of a frequency of four or more from the TCGA dataset were included in this analysis. The vast majority of variants (31333) occur in four different cases and the number of cases associated with variants diminishes with the increasing recurrence. Variants occurring in greater than 50 cases are rare.

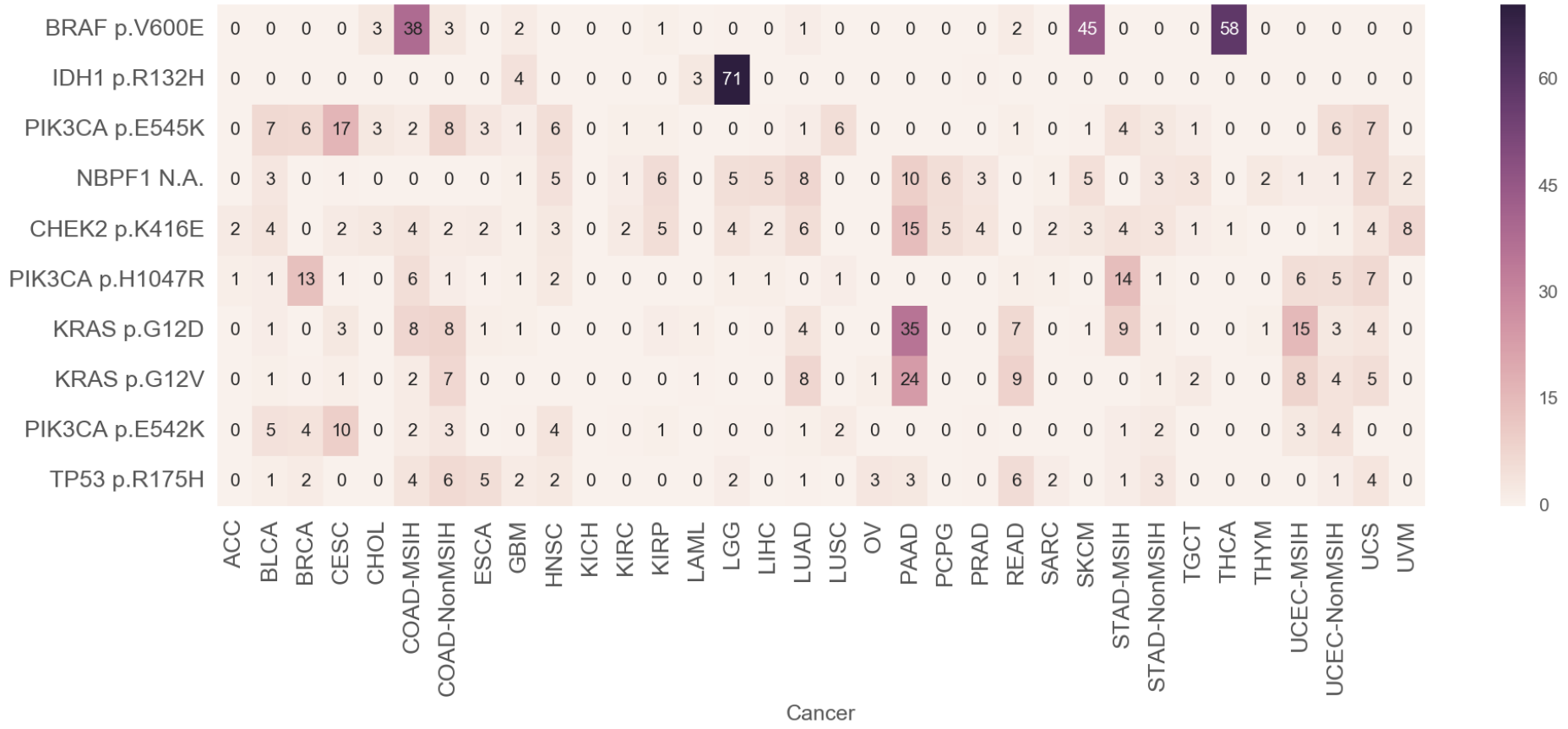

Figure 14: The 10 most frequent variants across all samples with annotations and associated cancers

This heatmap shows the 10 most frequent variants across the entire dataset. The gene (Hugo) symbol and protein changes are shown. The percentage of cases from each cancer with the corresponding mutation are indicated in the heatmap.

2.3.2. Creation of the data matrix for mutational signature analysis

Table 5 shows a representation of the data used in the mutation signature analyses. The data matrix is contained in a python pandas array comprising 8820 cases and 72542 categories (639,820,440 data points) made up of 192 trinucleotide change categories (96 as counts and 96 as proportions), 24 indel categories (6 insertion and 6 deletions categories as counts and proportions), 6178 categories of the genomic distribution (3089 each as counts and proportions), 24118 genes and 42030 variants of frequency 4 or more. The proportional data was stored as NumPy double-precision (64-bit) floating point values, the counts data was stored as unsigned integers (32-bit) while the gene and variants were stored as boolean entities.

Table 5: Representation of the matrix array created for the analysis of mutation signatures

case	cancer	TCG>TTG	Pro_TCG>TTG	Del1	Pro_Del1	Ins1	Pro_Ins1	TP53	TTN	7_140453136_140453136_A_T	3_178936091_178936091_G_A	17_7000001_8000000	Pro_17_7000001_8000000
TCGA-OR-A5J7	ACC	6	0.05	10	0.56	4	0.57	0	1	0	0	0	0
TCGA-VD-AA8T	UVM	0	0	2	0.67	1	1.00	0	0	0	0	0	0
TCGA-DK-AA74	BLCA	19	0.06	2	0.50	1	1.00	1	0	0	0	2	0.007
TCGA-BH-A1ET	BRCA	1	0.04	0	0	0	0	0	0	0	0	0	0
TCGA-D8-A1JF	BRCA	3	0.05	2	1.00	2	0.33	1	0	0	1	1	0.016
TCGA-DR-A0ZM	CESC	138	0.07	12	0.63	6	0.50	0	1	0	0	7	0.003
TCGA-AA-3712	COAD-NonMSIH	13	0.02	3	0.27	3	0.18	1	0	0	1	2	0.003
TCGA-06-2569	GBM	2	0.05	0	0	1	1.00	1	0	0	0	1	0.02
TCGA-GN-A267	SKCM	51	0.09	6	0.46	0	0	0	1	0	0	0	0
TCGA-BR-4280	STAD-MSIH	19	0.02	197	0.89	67	0.92	0	0	0	0	1	0.001
TCGA-AP-A0LT	UCEC-MSIH	38	0.05	120	0.80	26	0.87	0	0	0	0	3	0.003
TCGA-B5-A11S	UCEC-NonMSIH	7	0.07	1	0.20	2	0.67	0	1	0	0	0	0
TCGA-NG-A4VU	UCS	5	0.07	3	0.60	0	0	1	1	0	0	2	0.027

The data matrix showing examples of the data types used in this study. Every case is annotated with the TCGA case ID and the corresponding cancer type. The trinucleotide mutations are represented as counts (3rd column) and proportions (4th column). The deletions and insertions, collectively known as indel, are also represented as counts (5th and 7th columns) and proportions (6th and 8th columns). The genes (columns 9 and 10) and variants (columns 11 and 12) are represented as binary values (1 indicating present and 0 indicating absent). Lastly, the genomic distribution of the mutations is also represented as counts in each category (column 13), as well as proportions (column 14). In total, the matrix contains 8820 cases with 72542 mutation categories creating 639,820,440 data points.

Table 6: Case counts by tissue of origin

Site of Cancer	Cases
Breast	988
Brain	806
Lung	723
Kidney	686
Head and neck	526
Ovary	463
Thyroid	441
Prostate	425
Stomach	421
Bladder	412
Skin	370
Uterus	305
Adrenal gland	270
Colon	269
Soft tissue or bone	258
Liver	202
Cervix	198
Blood	192
Oesophagus	183
Pancreas	178
Testicle	150
Thymus	123
Rectum	115
Eye	80
Bile ducts	36

Table 7: Case counts by cell type of origin

Cell Type	Cases
Epithelial cells	6728
Glial cells	694
Melanocytes	450
Non-epithelial cells	315
Blasts cells	291
Myeloid cells	192
Germ cells	150

2.3.3. Multidimensional consensus cancer analysis: Overall cancer relatedness differs according to the different data type

Figure 15 to Figure 24 show the clustering results from the consensus representation of each cancer according to the different dimensions of data analysis, where clustering of the cases is performed along the x-axis (columns) as shown by the dendrogram at the top of the figures. The trinucleotide figures (Figure 15 and Figure 16) and indel figures (Figure 17 and Figure 18) also have the data categories clustered along the y-axis (rows) with a dendrogram showing the clustering results and a heatmap with the colour scheme relating to the data values. The heatmap representation has been inverse hyperbolic sine transformed so as to allow the values to be discerned over large dynamic ranges. It should be noted that this transform was only performed for visualisation but not for the actual clustering analysis.

Along with the clustering results, several annotations corresponding to the cancers have also been added, so as to observe for possible associations of these factors to the clustering groups. Of these, the MSI status annotation indicates the consensus representation of cases that were determined to exhibit microsatellite instability (MSI-high) according to the TCGA database and therefore segregated from microsatellite stable cases (non-MSI-high). The primary embryological origin, also know are germ layer, corresponding the primary tissue site of each cancer is also indicated, to study the supposition that processes related to early development may play a role in mutational processes. Organ (biological) systems (indicated in Table 3) comprise an interacting network of biologically interrelated anatomical structures that perform a specific function or task, an effort was also made to link the clustering patterns to these systems as seen in the figures. Fourteen organ systems are associated with the primary tissues of the cancers used in this study, including ‘female’ and ‘male’ to indicate

cancers that are associated with either gender. For clarity, breast cancers are listed as female due to high incidence rates in female and low rates observed in males (Ferlay et al. 2014). Several of the cancers correspond to more than one organ system due to the fact that organs may have multiple bodily functions. There is also a large comparative difference in the frequencies of occurrences within the different body, e.g. the digestive system is represented by 9 organs, while the integumentary system is represented by a single tissue (skin cutaneous melanoma (SKCM)). The sarcoma cases (SARC) are annotated as "unknown", as these cancers encompasses a broad family of rare cancers that can affect soft tissue or bone throughout the body and within the TCGA dataset, is represented by liposarcoma (fat cells in deep soft tissue), desmoid sarcoma (tendons and ligaments), nerve sheath tumour, synovial sarcoma (joints) among others. The organ of origin of each cancer ("organ") and the cancer cell type ("cell type"), as shown in Table 6, were also indicated to study possible associations. Table 7 summarises the seven cancer cell types found in the TCGA cancers. By far the most common cancer cell type is epithelial (6728 cases, 76 % of cancers), comparatively fewer germ cell cancers were present with only 150 cases, and all derived from the testicular germ cell tumours (TGCT). The mean overall survival (log10 transformed) was derived as a median of the overall survival (OS) available via the TCGA case annotations.

.

2.3.3.1. Consensus trinucleotide analysis

Figure 15 and Figure 16, show the clustering from the consensus versions of the trinucleotide data by two analysis approaches, namely as proportions (PR) and counts (CN) respectively. Clustering by the counts takes into account the total mutational load, specifically by using the city block (Manhattan) linkage metric, while the clustering by proportions arranges the results based on the interrelations of cancers by the mutation pattern alone, emphasised the use of Pearson's correlation as the metric.

Figure 47 to Figure 53 in the appendix are bar graphs showing the proportional distributions of the trinucleotide mutations from each cancer arranged with five plots per figure (for suitable visualisation) arranged according to the clustering of the proportions (Figure 15). Through these figures, the trends in trinucleotide changes from cancer to cancer can be seen quite easily. Overall C>T mutations are the most common variants, as seen in the SNV panel (top right), and Figure 47 to Figure 53, and therefore are the main determinants of the clustering. The skin cutaneous melanomas (SKCM) are highly distinct from all other cancers characterised by the over-representation of the four possible TC > TT mutations (TCA>TTA, TCC>TTC, TCG>TTG, TCT>TTT) that represent 46 % of all mutations. This distinction, however, is more apparent when clustered by PR versus CN. The MSI cases were found to have similar mutation profiles in both CN and PR having an enrichment of specifically CG > TG mutations (TCG>TTG, ACG>ATG, CCG>CTG, GCG>GTG) representing 42% of all observed mutations. Among the MSI cases, the cancers of both ectodermal and digestive origin, COAD and STAD were particularly similar (as opposed to the mesodermal and reproductive uterine corpus endometrial carcinoma (UCEC)) with a linkage distance (LD) of only 0.16 (Figure 15) when studied by proportions and 350 (Figure 16) by counts with the maximum linkage distance of 1.05 for PR and 1000 for CN. Both lung

cancer types, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), also showed great similarity by both analysis methods with distances of 0.14 (PR) and 44 (CN), however unlike the previous cancers, the lungs did not display enrichment of specific trinucleotide mutations, instead have an enrichment of a broad category if mutations, i.e. C>A/T mutations (57 % of mutations). By proportion, the most similar cancers were non-MSI colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ), two types of non-MSI colorectal cancer (LD 0.08). Similar to the MSI cases they are characterised by CG > TG mutations also representing 42 % of all mutations. Nervous system and non-MSI reproductive cancers were heterogeneous within themselves, however, all cancers clustered within a single node of maximum LD of 0.43 (PR) and 44 (CN). Two of the three squamous carcinomas, cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) and head and neck squamous cell carcinoma (HNSC) clustered closely (0.38 (PR) and 64 (CN)) but not the third (lung squamous cell carcinoma (LUSC)). The kidney cancers (kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP)) clustered closely (0.4 (PR) and 50 (CN)), however, the remaining urinary cancer, bladder urothelial carcinoma (BLCA) exhibit a very different profile. Based on trinucleotide proportions, endocrine and digestive cancers greatly differ in the overall profiles, perhaps eluding to different molecular mechanisms involved in these diseases. Prostate adenocarcinoma (PRAD), breast invasive carcinoma (BRCA), uveal melanoma (UVM), pheochromocytoma and paraganglioma (PCPG), thyroid carcinoma (THCA), acute myeloid leukaemia (LAML) and brain lower grade glioma (LGG) formed a cluster of very similar trinucleotide mutations (LD 20). The proportions seemed to discern specific cancers with greater detail than counts, however, clustering by both techniques provided comparable results. With the

exception of MSI, no association was seen between the clustering results and the additional annotations (cell type, organ etc.).

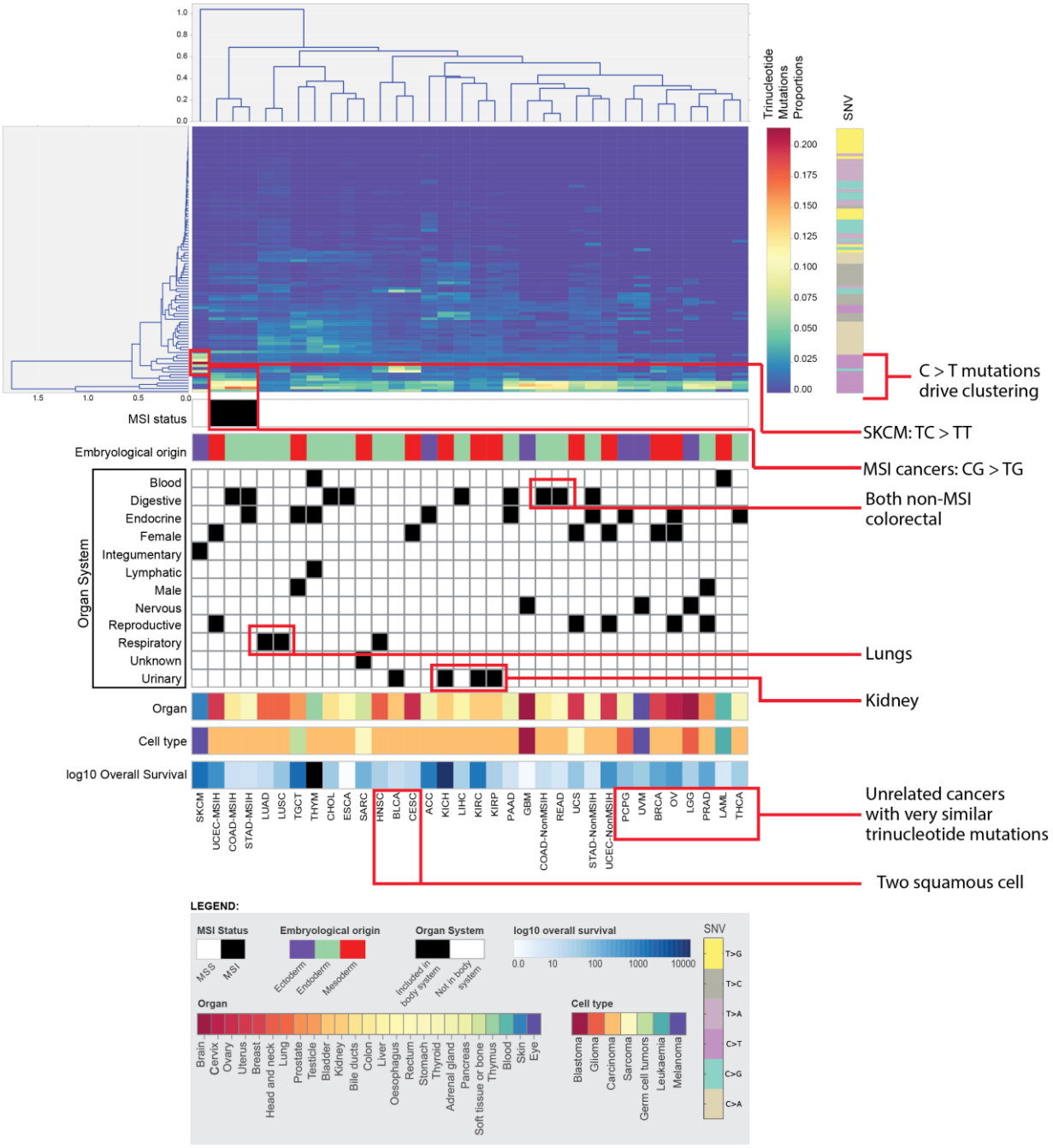

Figure 15: Clustering of the consensus trinucleotide mutation proportions by using average metric and city block linkage.

The cancers that cluster together are indicated.

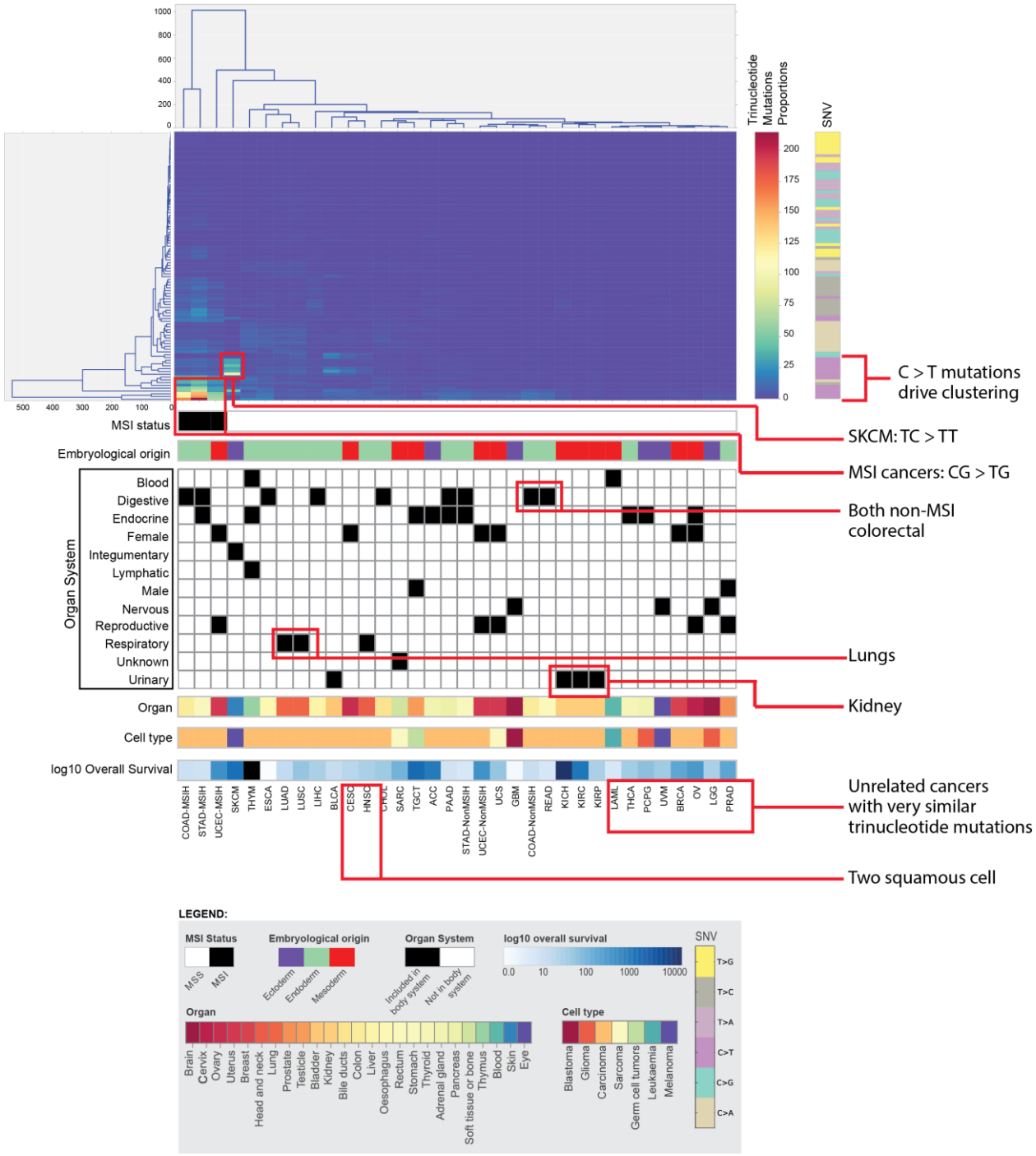

Figure 16: Clustering of the consensus trinucleotide mutation counts by using average metric and city block linkage.

The cancers that cluster together are indicated.

2.3.3.2. Consensus indels analysis

Figure 17 and Figure 18 show the consensus cancers clustering results for the indel PR and CN analysis respectively (a similar clustering approach was taken to the trinucleotide analysis). MSI cases show very similar profiles by both PR and CN and are distinguished by having mainly 1 base insertions and deletions. Figure 54 to Figure 57 in the appendix are bar graphs showing the proportional distributions of the indel mutation sizes from each cancer arranged with 10 plots per figure arranged according to the clustering of the indel proportions (Figure 17). Through these figures, the trends in indel changes from cancer to cancer can be seen quite easily.

The PR clustering seemed to segregate the MSI cases into a distinct cluster, however in the CN clustering, the ectodermal and digestive origin MSI cancers, colon adenocarcinoma (COAD) and stomach adenocarcinoma (STAD) were distinctly similar, while the COAD was more similar to the pancreatic adenocarcinoma (PAAD) due to the much higher mutational loads of the former cancers. Acute myeloid leukaemia (LAML), thyroid carcinoma (THCA) and ovarian serous cystadenocarcinoma (OV) consensus profiles showed no indels, i.e. these cancers have very low indel rates.

Cancers found to have similar profiles by the trinucleotide clustering e.g. lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), were not aggregated via indels clustering. Overall there were no discernible clustering associations within the organ systems or any association with the annotations other than MSI status. Although the urinary cancers appear adjacent to each other in the organ system heatmap, the cases do in fact occur in differing clusters, noticeable when the dendrogram is expanded. Apparent when analyzing Figure 54 to Figure 57, the cancers can be divided into four groups, 1) those with no mutations as mentioned above, 2)

cancers with only 1 base deletions (uveal melanoma (UVM), glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC)), 3) cancers with 1 base and greater than 5 base deletions and insertion (non-MSI colon adenocarcinoma (COAD), rectum adenocarcinoma (READ), adrenocortical carcinoma (ACC), sarcoma (SARC), pancreatic adenocarcinoma (PAAD)) 4) and all other cancers with 1 to >5 base insertions and 1 base insertions. By the PR analysis, brain lower grade glioma (LGG) and prostate adenocarcinoma (PRAD) had almost identical profiles, as did the esophageal carcinoma (ESCA) and head and neck squamous cell carcinoma (HNSC) both from group 4. By CN, lung adenocarcinoma (LUAD) and thymoma (THYM) were almost identical, as were bladder urothelial carcinoma (BLCA) and kidney renal clear cell carcinoma (KIRC) and prostate adenocarcinoma (PRAD), brain lower grade glioma (LGG) and pheochromocytoma and paraganglioma (PCPG).

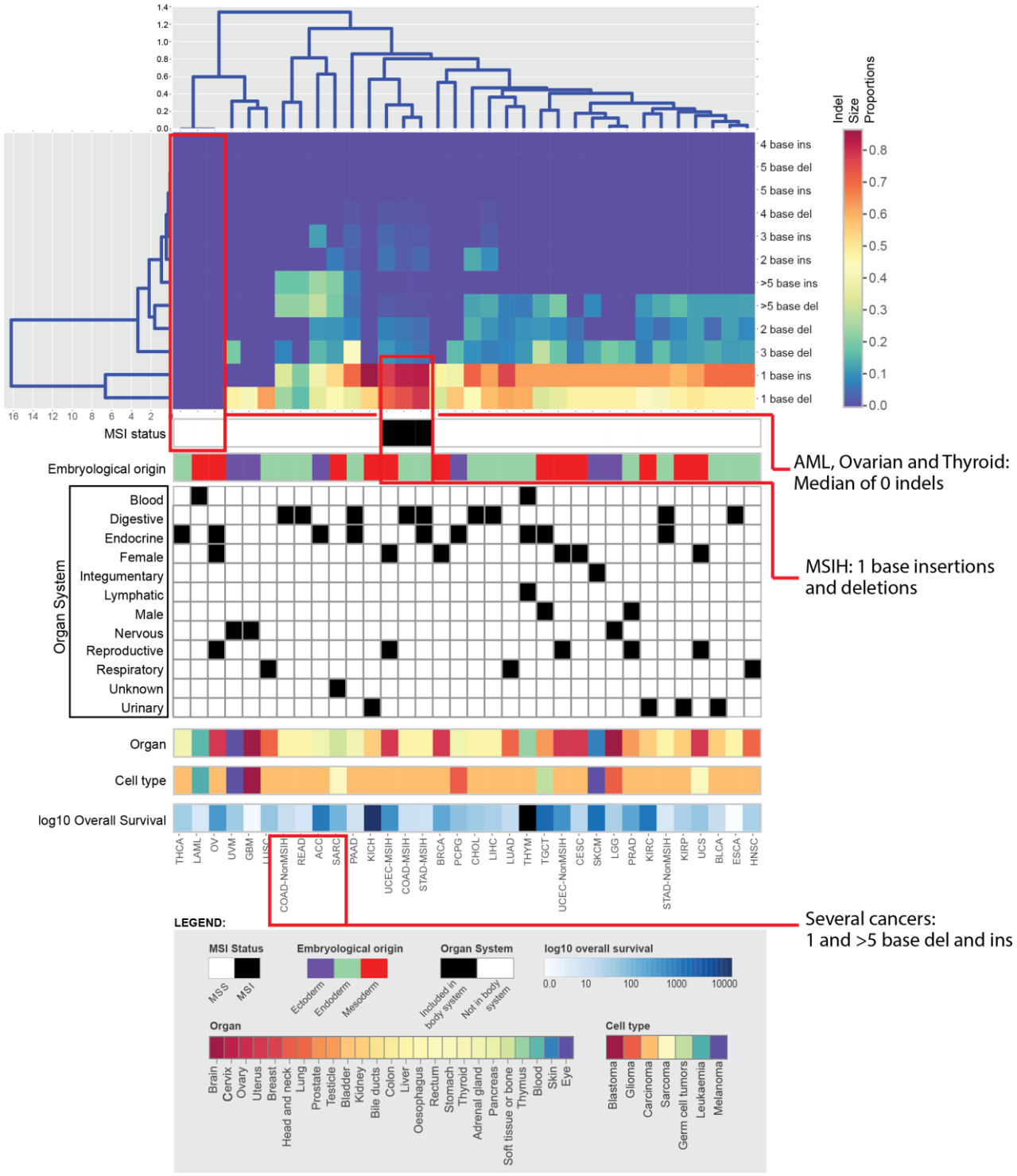

Figure 17: Clustering of the consensus indel proportions by using average metric and city block linkage.

The cancers that cluster together are indicated. del: deletion, ins: insertion

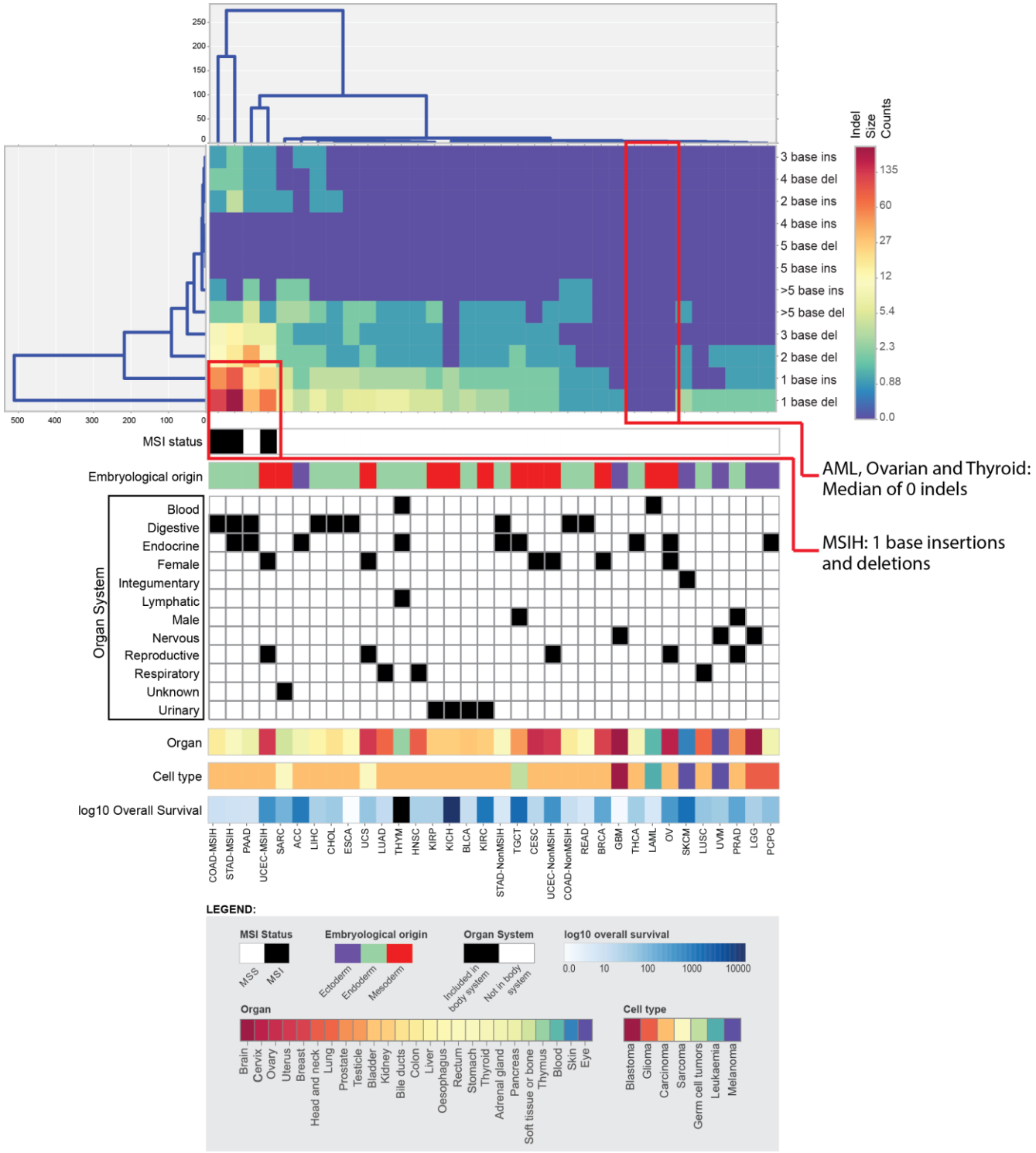

Figure 18: Clustering of the consensus indel counts by using average metric and city block linkage.

The cancers that cluster together are indicated. del: deletion, ins: insertion

2.3.3.3. Consensus genomic distribution of mutations

Figure 19 and Figure 20 show results from the PR and CN clustering on the consensus versions of the per Mb genomic distribution of the mutations, in a similar clustering approach to the previous two dimensions.

Unlike the indels and trinucleotide clustering, where the clustering patterns seemed to differ somewhat between the PR and CN data, as evidenced by the different ordering of cases along the x-axis, in this genomic distribution analysis, both the PR and CN clustering resulted in almost identical clustering order, only differing in the relative linkage distances, where the PR analysis generated a smaller relative range of linkage distance values and therefore more discernible branches in the dendrogram.

Interestingly, clustering appeared to be associated with embryological origin, i.e. cancers with primary tissues arising in the endoderm clustered to the left of the dendrogram, while mesodermal derived cancers to the right, as seen with the "embryological origin" annotations in Figure 19 and Figure 20, however, this tendency was more obvious in the PR analysis. This may suggest that the embryological origin, i.e. characteristics related to early developmental processes may play a role in the mutation patterns seen. There also appeared to be a tendency for the non-carcinoma cancers to cluster to the right, with only the skin cutaneous melanoma (SKCM) bucking the trend. The respiratory cancers clustered closely next to each other, indicating the lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) do in fact have similar genomic distribution profiles, in addition, these cancer were previously seen to be similar in the trinucleotide clustering, but not the indels. Based on linkage distance in both the PR and CN clustering, prostate adenocarcinoma (PRAD), pheochromocytoma and paraganglioma (PCPG), acute myeloid leukemia (LAML), breast invasive carcinoma (BRCA), and kidney renal papillary cell carcinoma (KIRP)

were found to have almost identical genomic distribution profiles. Much like the indel analyses, there did not seem to be an association with organ systems overall.

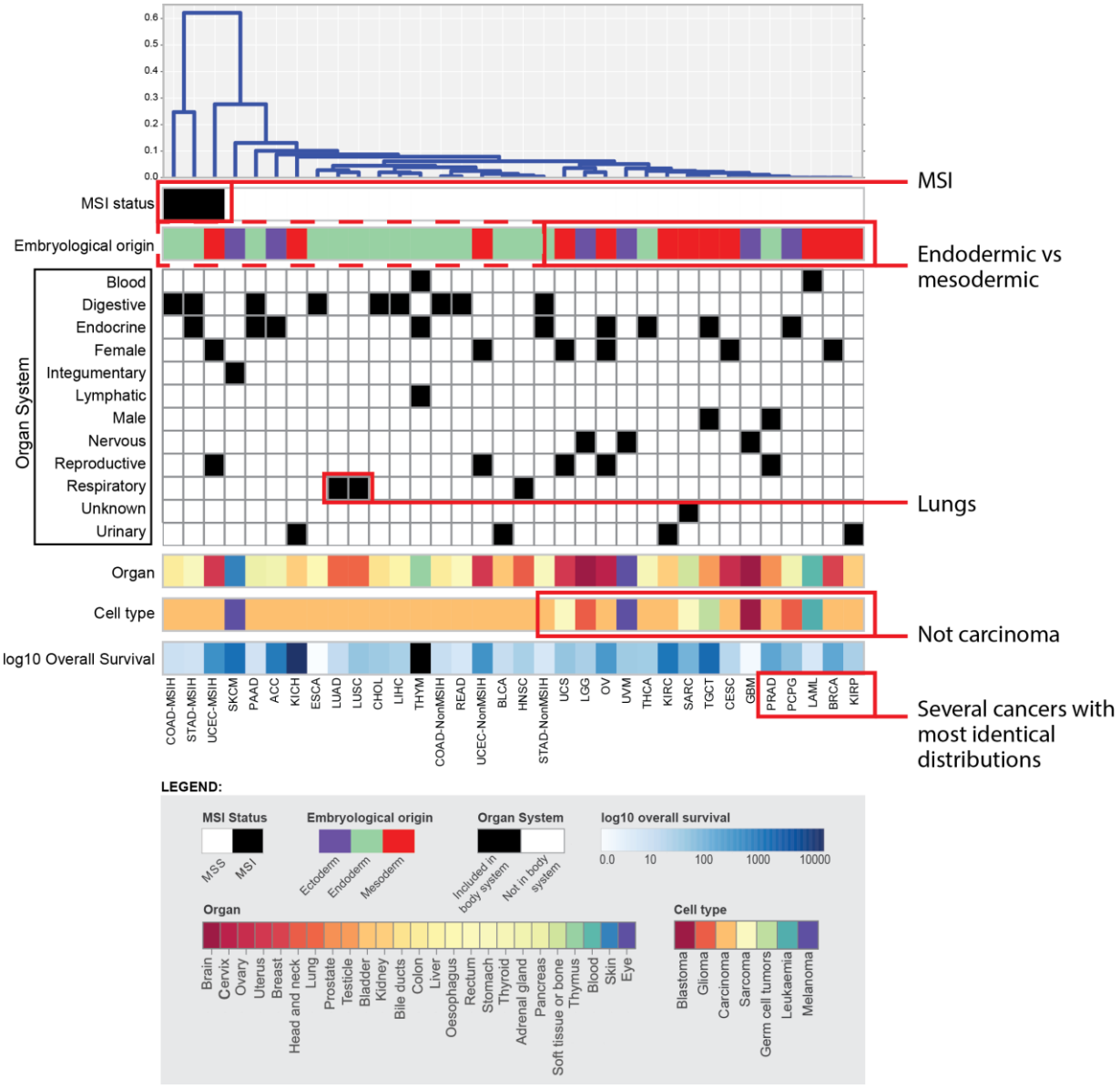

Figure 19: Clustering of the consensus mutational distribution proportions by using complete metric and city block linkage

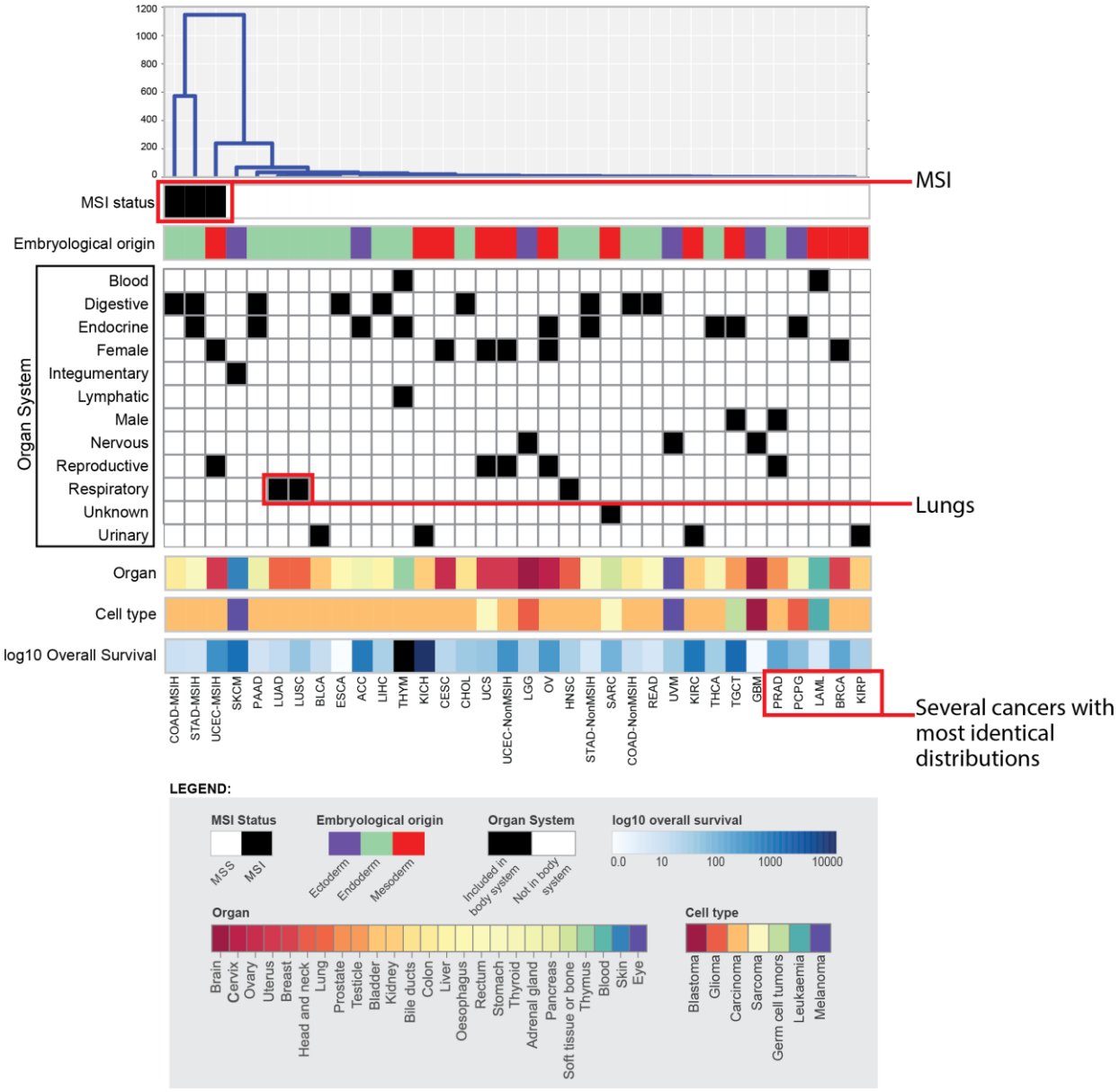

Figure 20: Clustering of the consensus mutational distribution counts using complete metric, city block linkage

2.3.3.4. Consensus genes and consensus variants analysis

Figure 21 and Figure 22 show results from the clustering of the consensus versions of the gene and variant data respectively, using the complete metric and city block linkage, as this data represents the mean values of each gene or variant category, unlike the non-consensus data (individual cases) which is binary.

2.3.3.4.1. Analysis of consensus gene profile

The most closely related cancers, based on linkage distance, clustered to the right side of the dendrogram in the genes based analysis (Figure 21) i.e. adrenocortical carcinoma (ACC) to thyroid carcinoma (THCA). These cancers tend to be cancers with low mutational loads as indicated in Figure 6. This would indicate that overall, there are fewer specific genes that are consistently mutated in these cancers. Acute myeloid leukemia (LAML) and THCA seemed to be particularly similar in their mutated genes profiles, as well as being the two cancers with the lowest mutational loads and share mutations in 781 genes. As with the trinucleotide and genomic distribution analyses, the two lung cancer associated via clustering, indicating not only similar mutational loads but also mutations in similar genes. The most distinct cancer appears to be the MSI stomach adenocarcinoma (STAD), with the other MSI cases also being distinct from all other cases, again, most likely due the large number of mutated genes, as related to the overall mutational load. In a similar way to the genomic distribution analysis, the cancers with an endodermic origin tended to cluster to the left, and these cancers tended to have much higher mutational loads Figure 6.

2.3.3.4.2. Analysis of consensus variants profile

Unlike the clustering pattern of the consensus genes profiles, the equivalent analysis using the consensus variants profiles did not mirror the mutational load. The

rationale behind clustering using this dataset is to identify similarities based on the recurrences of specific variants. Liver hepatocellular carcinoma (LIHC) seemed to have a vastly different mutation profiles compared to all other cases, although this does not seem to be linked specifically to a mutation load abnormality, as liver hepatocellular carcinoma (LIHC) is centrally located among all cancers according to mutation load (Figure 6). To a lesser extent pancreatic adenocarcinoma (PAAD) also exhibited a very different profile from all other cancers, another cancer with a moderate mutation load. As seen in Figure 14, PAAD is associated with many recurrent mutations, and is in fact the cancer type that seems to have the largest number of recurrent mutations e.g. chr22:29091840-29091840 (CHEK2 K416E) in 14 % of cases, chr3:178952085-178952085 A>G (PIK3CA H1047R) in 35 % of cases, and chr12:25398284-25398284 C>T and chr12:25398284-25398284 C>A (KRAS G12D and G12V) in 35% and 23 % of cases, all of which are cancer related variants, but with especially high rates in PAAD (variants shown with protein change annotation) (Wellcome Trust Sanger Institute 2016a). Both the non-MSI large intestinal cancers, colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) were closely related, as were two kidney cancers (kidney chromophobe (KICH) and kidney renal papillary cell carcinoma (KIRP)), but not the kidney renal clear cell carcinoma (KIRC). The two brain related cancers (brain lower grade glioma (LGG) and glioblastoma multiforme (GBM)) also showed great similarity in their variant profiles.

As described in this section, the cancers ordering after clustering of mutated genes closely resembles the overall mutational load of the individual cancers, as seen in section 2.3.1, but this is not seen with the clustering of variants. This observation is likely due to the fact that the gene level analysis involves the aggregation of all variants into genes and therefore normalizes for the variations seen when looking at just

individual variants, as such it is likely that most variants are in fact non-specific passenger events as opposed to driver events in cancers, as have been speculated previously (Stephens et al. 2012), however there is a tendency towards mutation in specific genes.

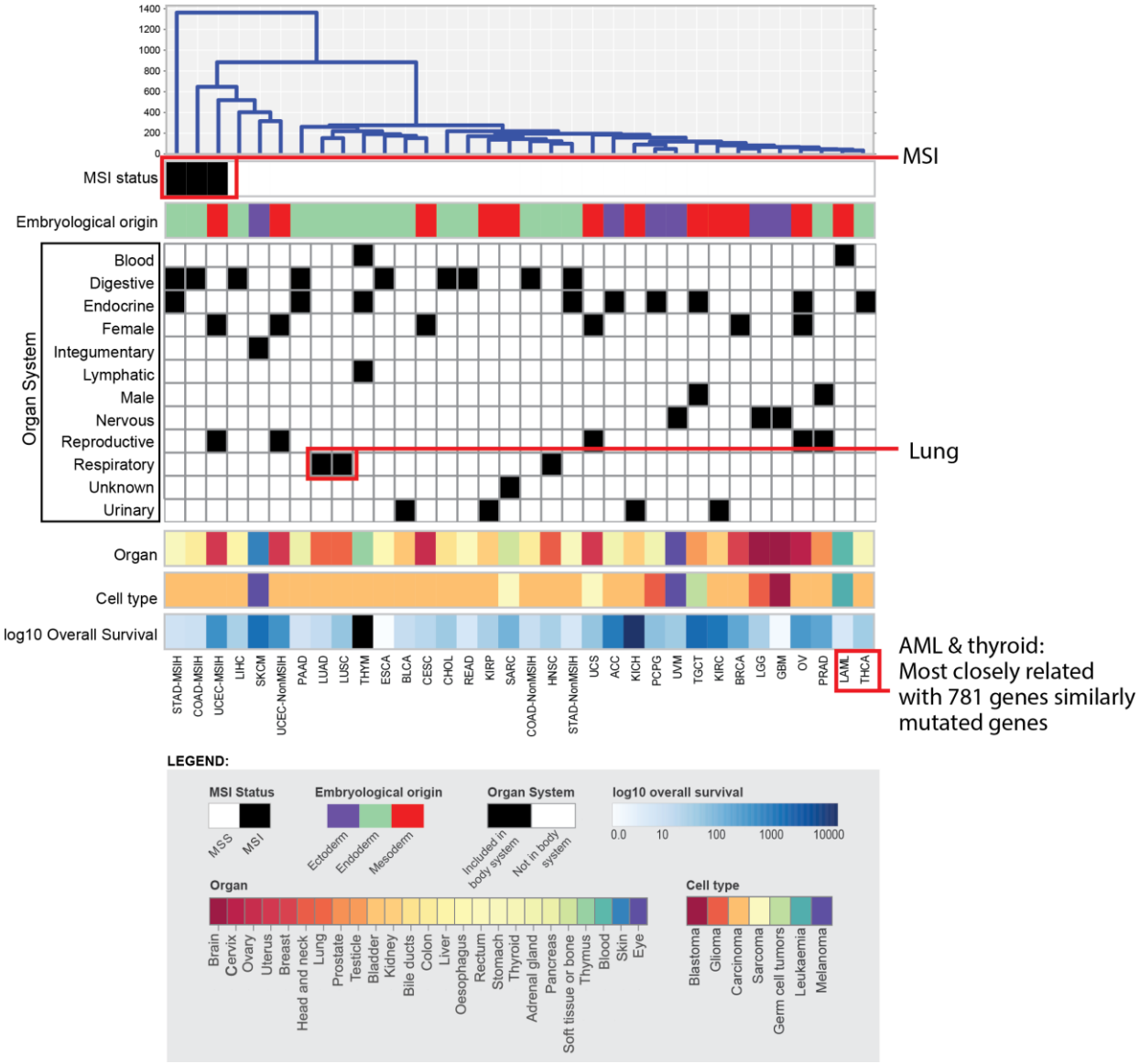

Figure 21: Clustering of the consensus mutated genes by using complete metric and city block linkage

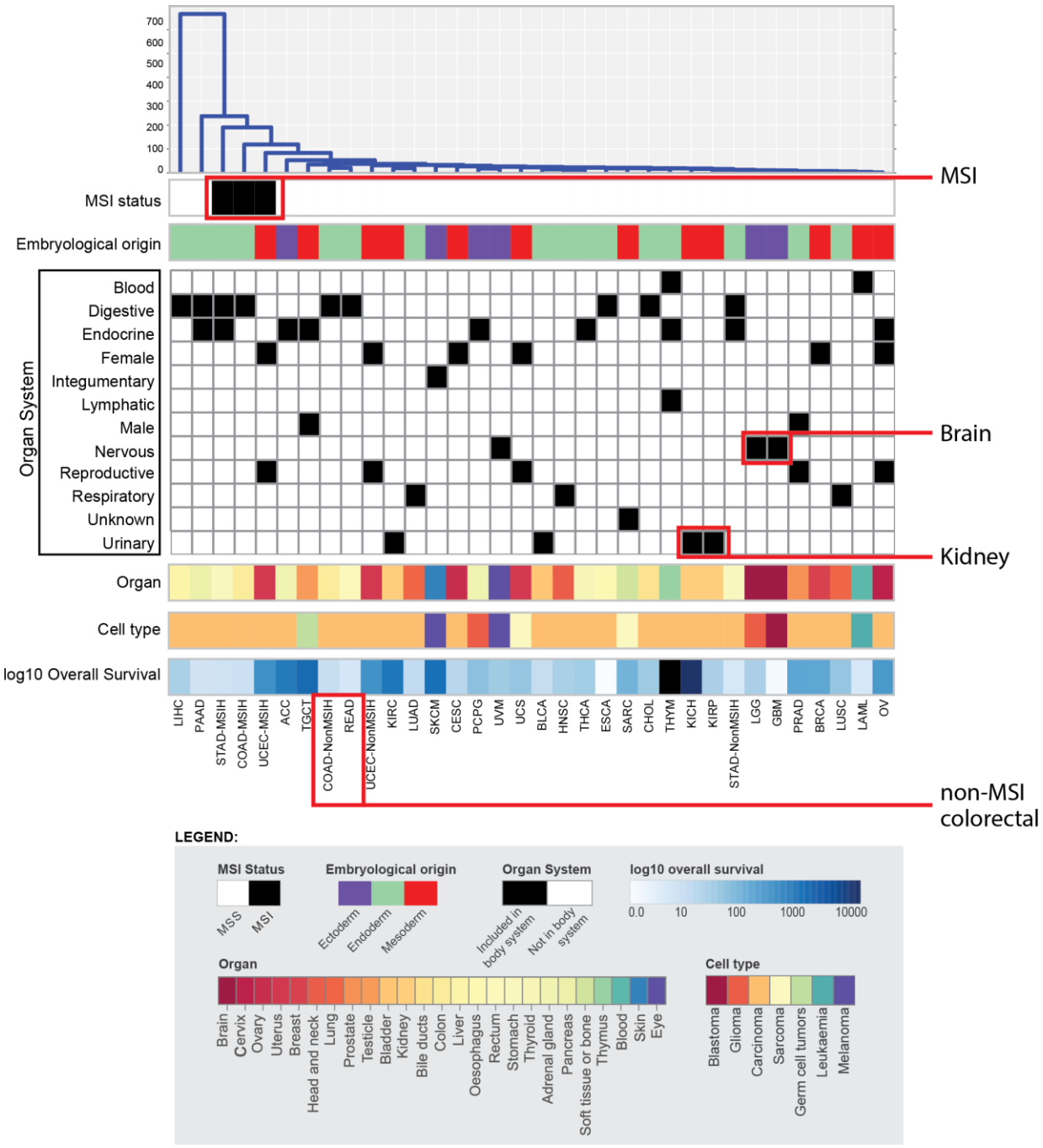

Figure 22: Clustering of the consensus variants by using complete metric and city block linkage

2.3.3.5. Consensus multidimensional analysis

Figure 23 and Figure 24 show results from the PR and CN clustering on the consensus versions of the multidimensional data respectively. By combing data from the many dimensions into a single analysis, fewer associations were seen between the different cancers, suggesting that rather than helping to discern the different cancers, the combination many actually mask the effect of the individual dimensions, or at least masks it from the relatively simple analysis method that was used, unsupervised hierarchical clustering. Potentially, complex relationships may be determined by utilizing more advanced clustering methodologies e.g. neural networks or support vector machines, however this was not attempted in this chapter.

Despite this, two sets of cancers systems did seem to cluster with each other, namely the two lung cancers (lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC)) and the two kidney cancers (kidney chromophobe (KICH) and kidney renal clear cell carcinoma (KIRC)). The lung cancers specifically have been shown to associate with each other in all dimensions of analysis except with indel clustering.

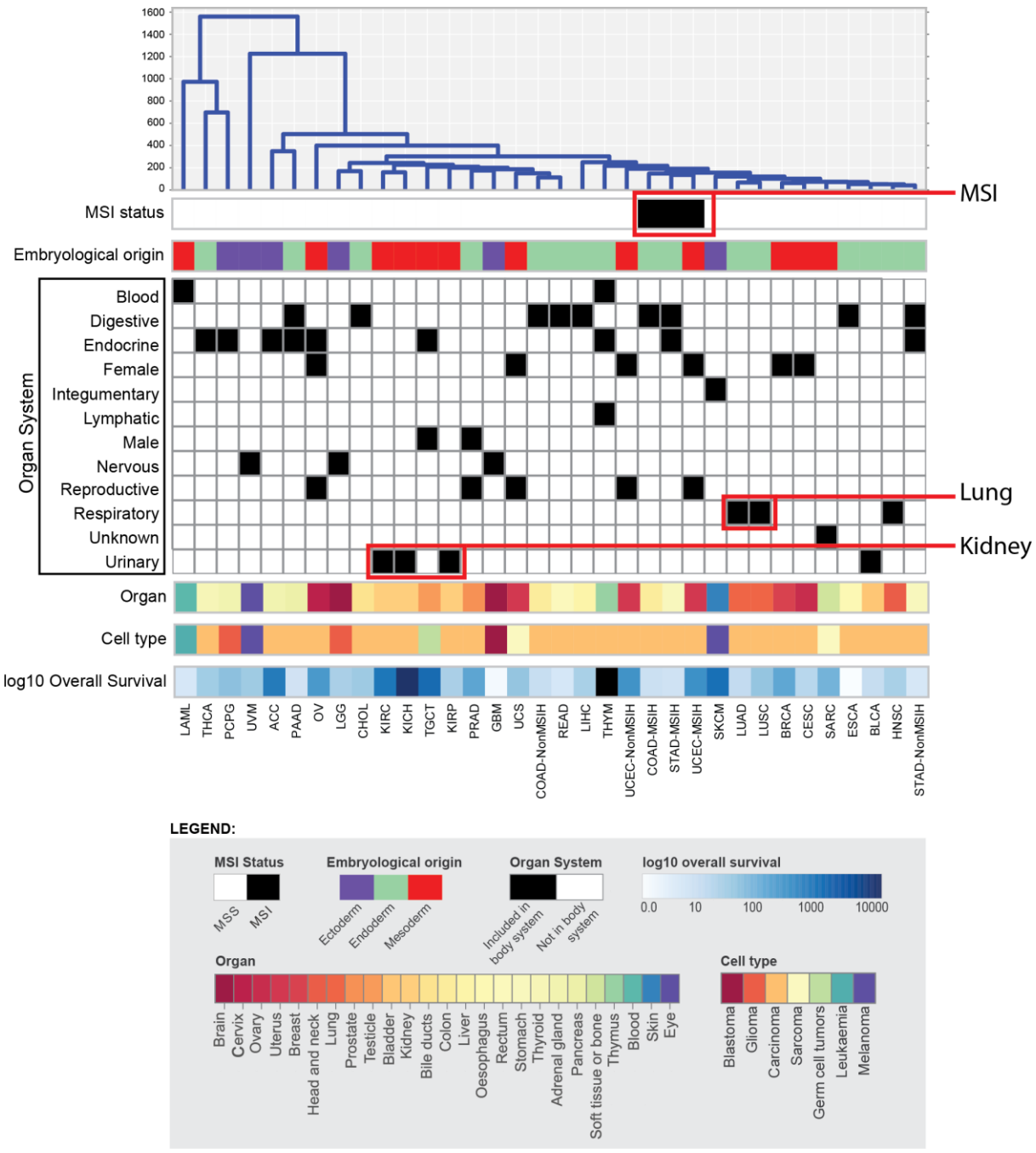

Figure 23: Clustering of the consensus multidimensional mutations with proportions by using average metric and city block linkage

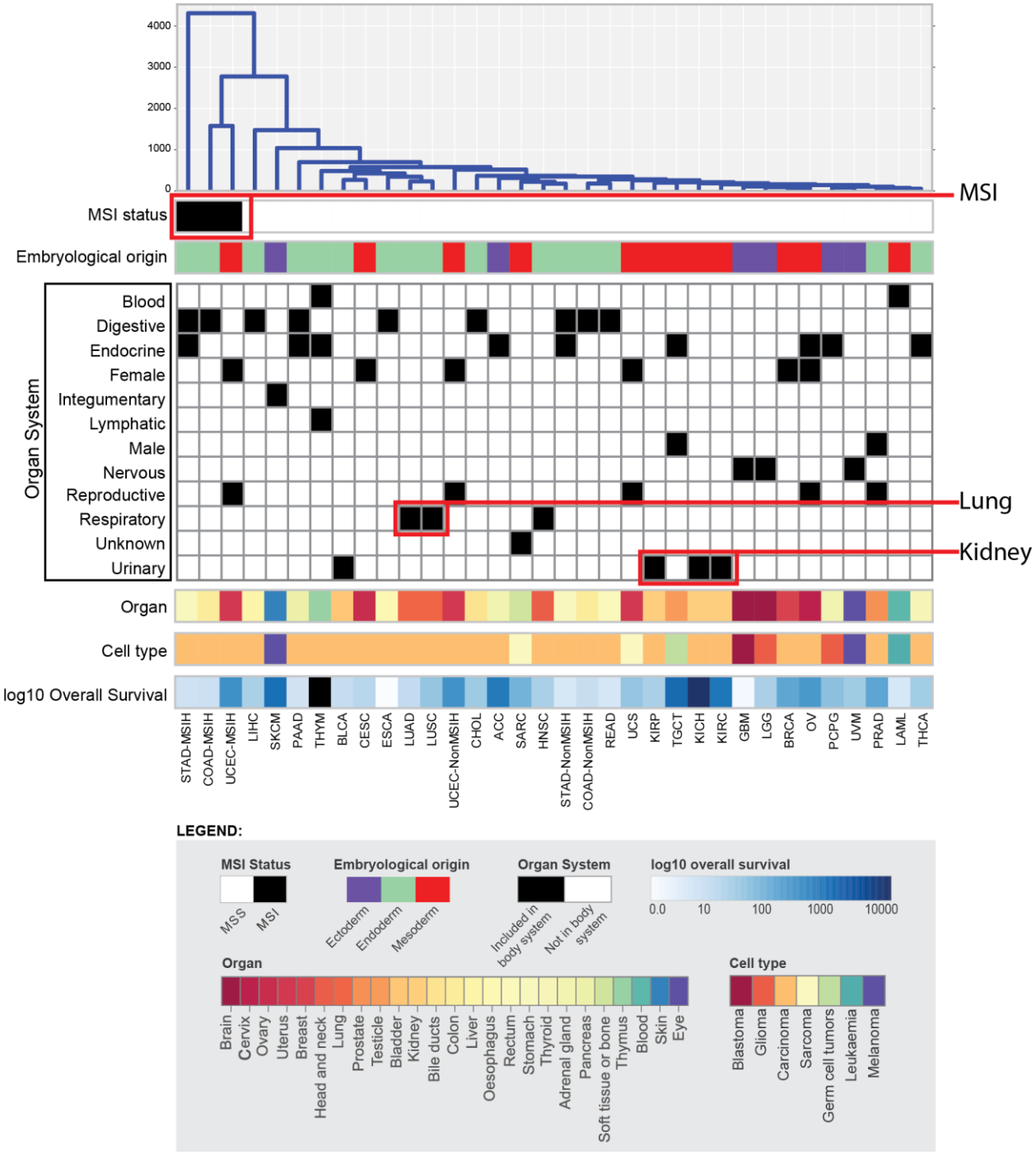

Figure 24: Clustering of the consensus multidimensional mutations with counts by using average metric and city block linkage

2.3.4. Analysis at multiple dimensions of all cases reveals several profiles specific to cancer type or subsets of cases

In the previous section 2.3.3, the relatedness of cancers was studied by the clustering of consensus versions of each cancer to understand the overall relatedness of the different cancers. In this section, a similar clustering analysis is performed on the 8820 individual cases used in this study, and as before, a comparison to several annotations was made. In addition, age at diagnosis and gender were also added to the set of annotations that were used in section 2.3.3.

The results can be seen in Figure 25 to Figure 42, and due to space constraints, the legend for these is shown in Figure 27. A difference when compared to the consensus diagrams is the inclusion of a second heatmap below the dendrogram in all figures, to indicate the clustering position for cases from each cancer. Four different clustering types were observed which distinguish how the cases within each cancer were ordered after the hierarchical clustering. Unique cancer signature (UC) cancers are homogeneous across cases, i.e. most cases from these cancers cluster into a single group, and share a mutation profile unique to the cancer, i.e. little overlap in clustering with other cancers, these are indicated with a red bounding box in the figures. Unique subtype signature (US) cancers that are overall heterogeneous however there is at least one homogeneous subgroup with a profile unique to that subgroup, these are indicated with an orange bounding box in the figures. Shared Cancer signature (SCS) cancers are homogeneous across cases, however profiles are not unique across cancers i.e. the clustering of cases from this cancer overlap with cases of other cancers and are indicated with a green coloured bounding box. Non-definable type (NDT) cancers are heterogeneous across cases, i.e. no unique cancer profile can be discerned from this cancer. This classification system is standardized across all dimensions of the mutation

analysis. Congregations (clustering together) of annotations are indicated with a purple bounding box and these include the congregation of microsatellite instability and embryological origin. Cases that fall within the UC, US or SCS cluster are further studied to elucidate specific characteristics of these cases that set them apart from other cases, i.e. their unique mutational profiles.

2.3.4.1. Trinucleotide mutations by cases

Figure 25 and Figure 26 show the agglomerative unsupervised hierarchical clustering of the trinucleotide mutation according to all cases by PR and CN respectively. Clustering of cases is done along the x-axis and trinucleotide mutations clustered along the y-axis. A heat map showing the proportions of the various trinucleotide mutations with intensities corresponding to the colour bar to the right of the heatmap. As can be seen, the distribution of most cancers are highly heterogeneous, that is, they do not tend to cluster according to the cancer type, except for a few cancers. When clustered by PR of the trinucleotide mutations, four cancers, lung adenocarcinoma (LUAD), skin cutaneous melanoma (SKCM) and testicular germ cell tumours (TGCT) and thymoma (THYM) formed UC clusters, i.e. unique mutational profiles across all cases. Three cancer types, bladder urothelial carcinoma (BLCA), liver hepatocellular carcinoma (LIHC) and stomach adenocarcinoma (STAD) MSI-high formed SCS clusters, i.e. a subset of cases from these cancers seemed to have unique profiles. When clustered by CN, only two cancers had UC clusters, i.e. TGCT and thymoma (THYM), both clustered as in PR as well. By CN, SKCM had two US clusters and acute myeloid leukaemia (LAML) was an SCS cluster. Unlike PR, the CN analysis revealed a tendency for the endodermal origin cancers to cluster to the left and the mesodermal cancers to the right of the dendrogram, a phenomenon which was also seen in the genomic distributions consensus clustering (section 2.3.3.3, Figure 19 and

Figure 20). By both analysis approaches (PR and CN), distinct subgroupings of cancers were observed, however generally there appeared to be a greater number of cancers discernible by using proportions.

The most comprehensive work on mutation profiles thus far was done by Alexandrov et al. (Alexandrov and Stratton 2014), where the authors established 21 mutational signatures of mutations in the TCGA dataset. Additional mutational data and further analysis have revealed 7 additional signatures (Wellcome Trust Sanger Institute 2016b). The Alexandrov study used the proportions of nucleotide changes, i.e. PR analysis, to elucidate trinucleotide mutation profiles in an unsupervised manner, in an approach which was agnostic to cancer type. The work presented in this thesis has specifically attempted to elucidated signatures that are unique to cancers or cancer subsets, where the signature may not be shared by more than one cancer. As described in the previous paragraph, several of these unique signatures have been identified. Figure 28 to Figure 30 shows comparisons of the unique cancers signatures found via the PR analysis, to those established by the Alexandrov approach. This approach assumes that the unique cluster is a component of the overall Alexandrov mutational signature analysis. The figures are arranged with the closest matching Alexandrov signature shown above the cancer with the unique profile with the cosine similarity score (css) shown. The bladder urothelial carcinoma (BLCA) closely resembles signature 2 (css = 0.81) (Figure 28), a signature that is linked to activity of the AID/APOBEC family of cytidine deaminases, and most likely APOBEC1, APOBEC3A and/or APOBEC3B (S. a Roberts et al. 2013), and it has been suggested that activation these proteins may be caused by tissue inflammation, viral infection, or even retrotransposon jumping. The liver hepatocellular carcinoma (LIHC) PR signature is most similar to signature 12 (css = 0.76), a signature with unknown aetiology, but

known to be associated with these cancers. The low r value, however, does suggest that the signatures are in fact different. In Figure 29 it can be seen that the lung adenocarcinoma (LUAD) is closely associated with Alexandrov signature 4 (css = 0.93), associated with tobacco carcinogens (e.g., benzo[a]pyrene) showing strand bias for C>A mutations(Pfeifer et al. 2002). Skin cutaneous melanoma (SKCM) identical to signature 7 (css = 0.99), a signature linked to large numbers of CC>TT dinucleotide mutations at dipyrimidines as specifically associated with melanomas. Testicular germ cell tumours (TGCT), thymoma (THYM) and the MSI cases were most closely related to signature 6 (css = 0.75, 0.79 and 0.97), as mutational signature found in 17 cancer types and linked to defective DNA mismatch repair, i.e., as would explain the high cosine similarity score in the MSI cases. As noted the liver hepatocellular carcinoma (LIHC), TGCT and THYM had poor cosine similarity scores, i.e. weak associations with the closest Alexandrov signature. Most interestingly the TGCT has not been analysed within the original Alexandrov dataset nor in the expanded dataset with 30 signature, and this may indicate that the unique profiles seen in the TGCT is in effect a newly revealed mutational signature that is not currently included in the 30 established signatures.

Unlike the study of PR mutational signatures, CN signatures have not been studied, i.e. all previous studies (Greenman et al. 2007; Lawrence et al. 2013; Alexandrov and Stratton 2014) have only look at the problem using proportions of the trinucleotides. To understand the unique features of the cancer-specific count signatures, these signatures were compared to all other cancers. Table 8 shows the five most statistically altered trinucleotide changes from the UC, US and SCS cancers, while the entire profile (all 96 trinucleotide changes) compared to a consensus representation of all cancers are shown in Figure 31 and Figure 32. TCGT does not show any specific trinucleotide

mutation rates from the general rates seen in all cancers, however, does seem to have fewer mutations in certain trinucleotide categories without a consistent pattern. THYM has a statistically much higher number of mutations overall when compared to the other cancers, but specifically increased A<u>C</u>>A<u>A</u> mutations. The two SKCM clusters show much higher mutation rates in C<u>C</u>>C<u>T</u> despite having much lower mutation loads overall when compared to all other cancers. Acute myeloid leukaemia (LAML) showed much lower mutation rates in most of the mutations categories, specifically <u>C</u>G>T<u>G</u> mutations. The endodermal cancers had higher mutation rates in most categories, except lower numbers of C>G mutations. The most statistically differing mutation rates across all cancers in Table 8 were A<u>C</u>>A<u>A</u> mutations.

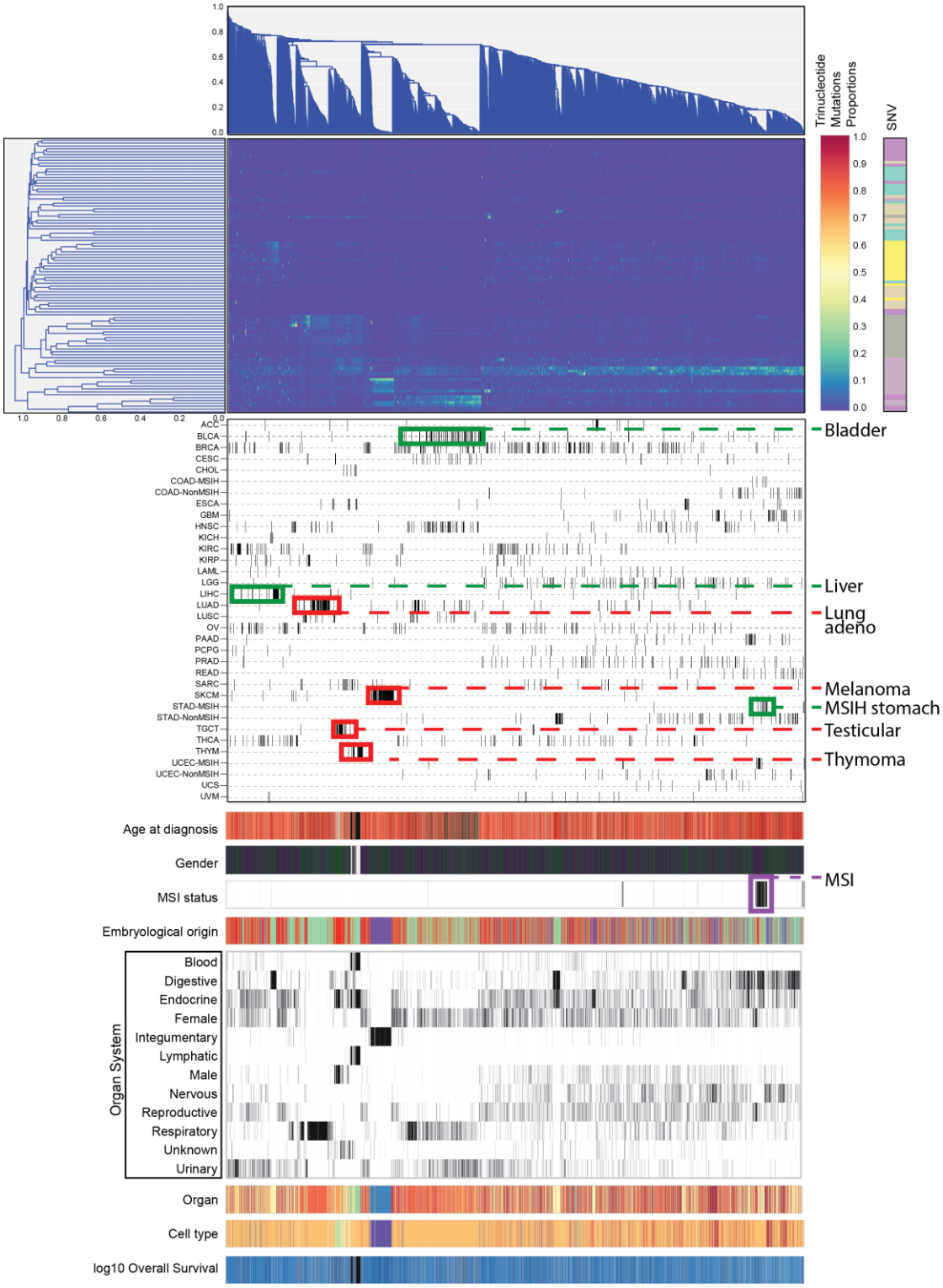

Figure 25: Clustering of trinucleotide mutation proportions in all cases by using average metric and correlation linkage.

The legend for this figure is inserted as Figure 27. Clustering has been performed on all cases in the TCGA according to the proportions of trinucleotide changes. The x-axis dendrogram shows clustering of the data by case, while the y-axis dendrogram shows clustering by the mutation. The top heat map shows the proportion of mutation by case (each column totals to 1). UC clusters are shown with a red boundary, SCS clusters are shown with a green boundary and the MSI cases cluster is shown with a purple boundary.

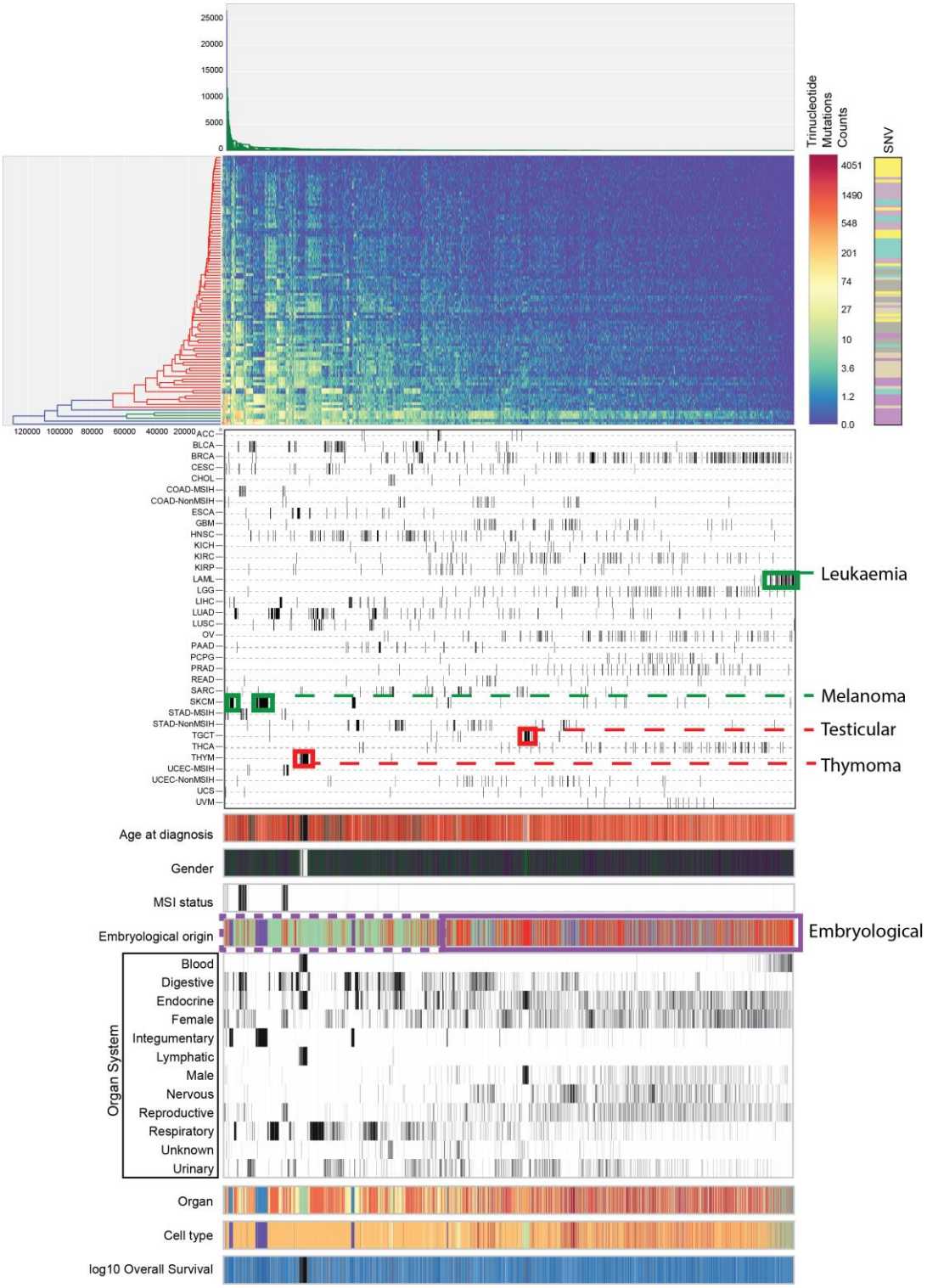

Figure 26: Clustering of trinucleotide mutation counts in all cases by using average metric and city block linkage.

The legend for this figure is inserted as Figure 27. Clustering has been performed on all cases in the TCGA according to the counts of trinucleotide changes. The x-axis dendrogram shows clustering of the data by case, while the y-axis dendrogram shows clustering by the mutation. The top heat map shows the counts of mutation by case. UC clusters are shown with a red boundary, US clusters with an orange boundary, SCS cancers are shown with a green boundary and the MSI cases cluster and endodermal case clustering are shown with a purple boundary.

LEGEND:

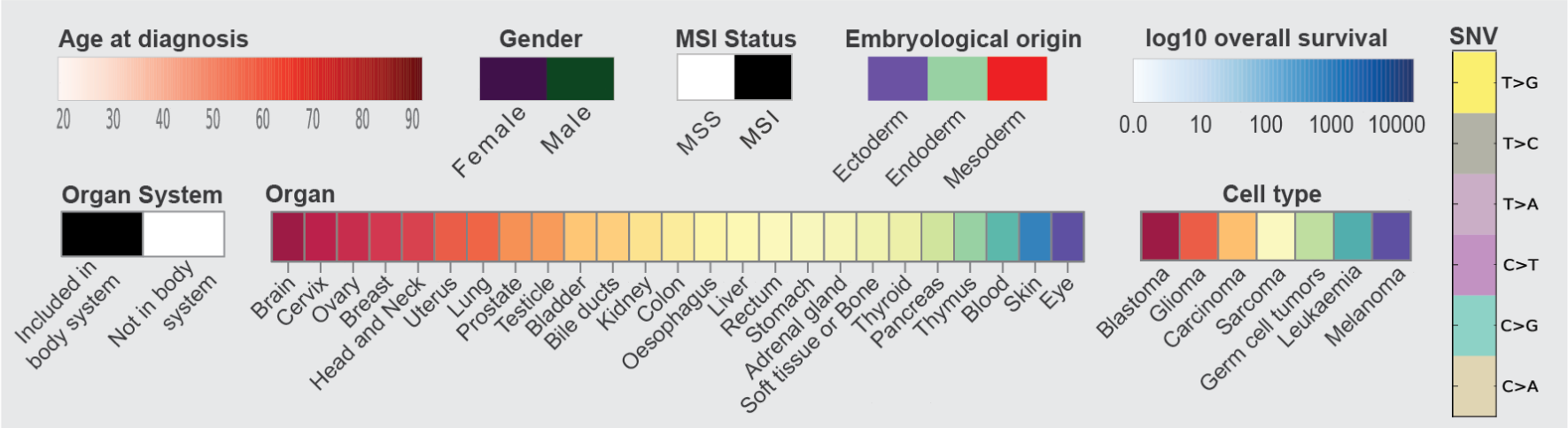

Figure 27: Legend for clustering of all cases.

This legend is used for Figure 25 and Figure 26,

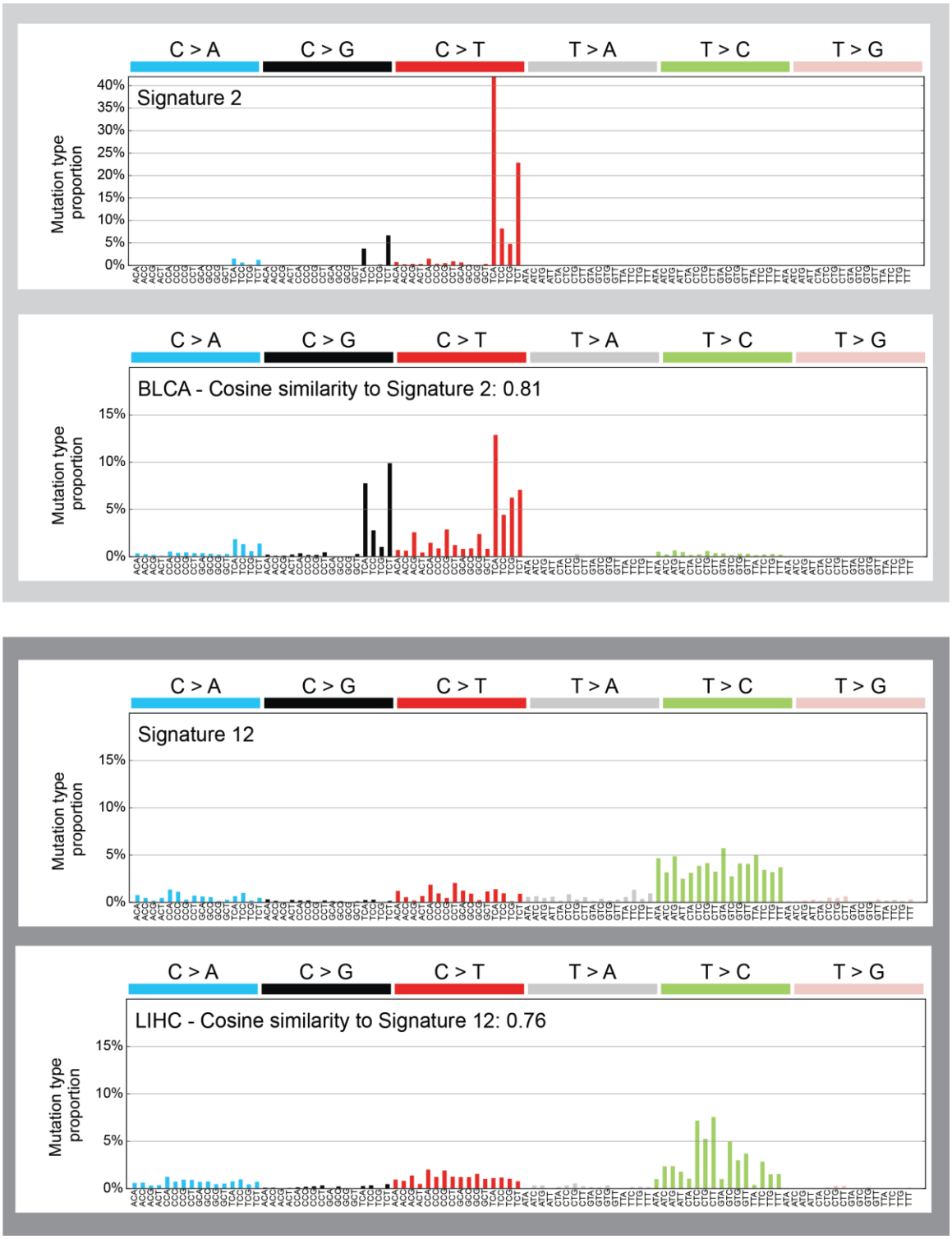

Figure 28: Alexandrov signatures compared to the derived unique cancer signatures (BLCA and LIHC)

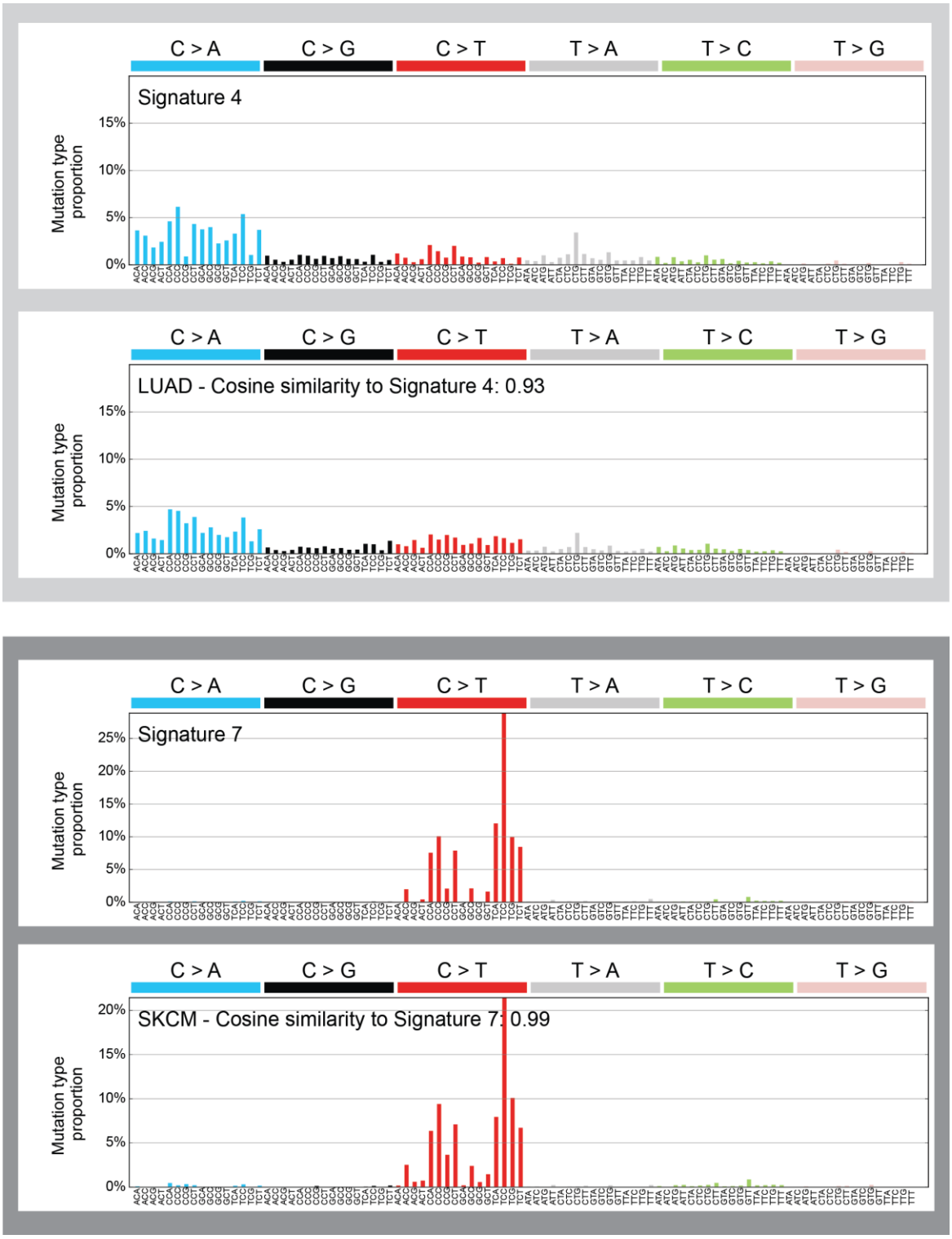

Figure 29: Alexandrov signatures compared to the derived unique cancer signatures (LUAD and SKCM)

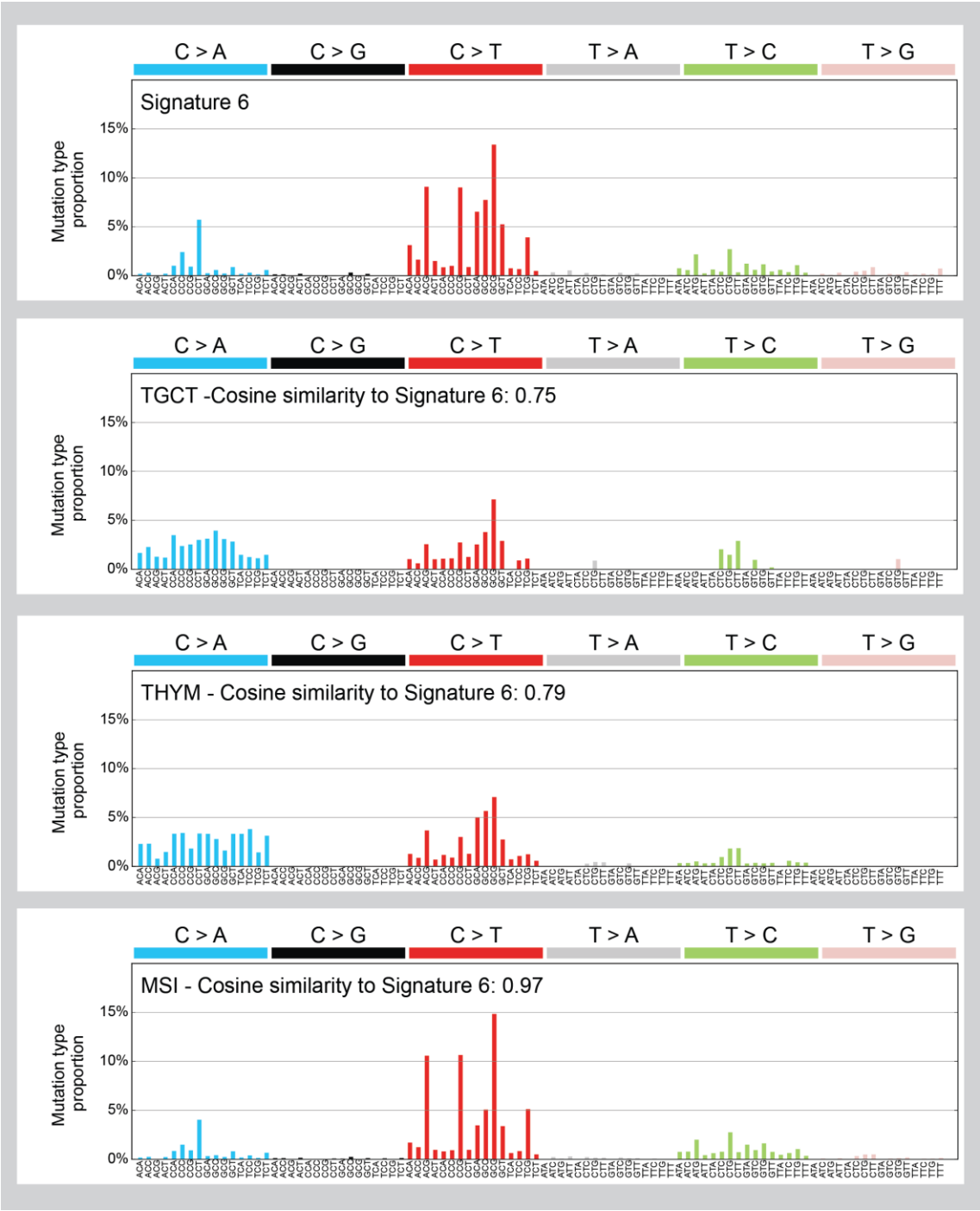

Figure 30: Alexandrov signatures compared to the derived unique cancer signatures (TCGT, THYM and MSI-high)

Table 8: The five most variable trinucleotides observed in the UC, US and SCS cancers by CN clustering

	Trinucleotide change	Mean value of cancer cluster	Mean value of all other cases	P value	fdr_bh
TGCT	**CCT>CGT**	0.32	1.01	0.0017	0.07
	CTT>CAT	0.19	0.91	0.0017	0.07
	CTG>CGG	0.13	0.90	0.0049	0.07
	TTG>TAG	0.10	0.47	0.0052	0.07
	GTG>GAG	0.22	0.73	0.0055	0.07
THYM	**ACA>AAA**	6.88	1.90	3E-40	3E-38
	ACC>AAC	7.04	1.87	3E-39	2E-37
	TCC>TAC	10.97	3.30	1E-18	4E-17
	GCA>GTA	14.18	3.49	5E-11	1E-09
	CCC>CAC	9.93	3.35	9E-11	2E-09
SKCM(1)	**TCC>TTC**	322.00	12.37	9E-50	8E-48
	CCA>CTA	85.35	5.10	2E-35	1E-33
	TTT>TAT	5.76	0.58	7E-29	2E-27
	CCC>CTC	128.65	6.14	3E-27	8E-26
	CCT>CTT	92.71	5.72	3E-21	6E-20
SKCM(2)	**TCC>TTC**	132.33	10.48	1E-79	1E-77
	CCA>CTA	39.17	4.54	3E-67	1E-65
	CCC>CTC	57.43	5.31	6E-50	2E-48
	CCT>CTT	43.08	5.11	1E-40	3E-39
	ACC>ATC	16.12	2.86	1E-38	2E-37
LAML	**ACC>AGC**	0.04	0.68	3E-09	1E-07
	CCT>CGT	0.03	1.02	4E-09	1E-07
	ACA>AAA	0.09	2.00	4E-09	1E-07
	ACT>AGT	0.04	0.74	3E-08	7E-07
	ATG>AAG	0.03	0.73	3E-08	7E-07
Embryological origin	**ACA>AAA**	3.91	0.72	0	0
	ACC>AAC	3.88	0.70	6E-286	3E-284
	TCC>TAC	7.50	0.79	1E-263	4E-262
	CTT>CAT	1.85	0.30	3E-260	8E-259
	ATG>AAG	1.41	0.29	7E-258	1E-256

Shown are the five most statistically differentially mutated trinucleotide for each of the unique cancer profiles derived from the counts clustering of all cases. The mean value for the unique cancer profiles and all other cases is shown, along with the t-test p-value and the benjamini hochberg false discovery rate value. There are two categories of SKCM corresponding to the two US clusters.

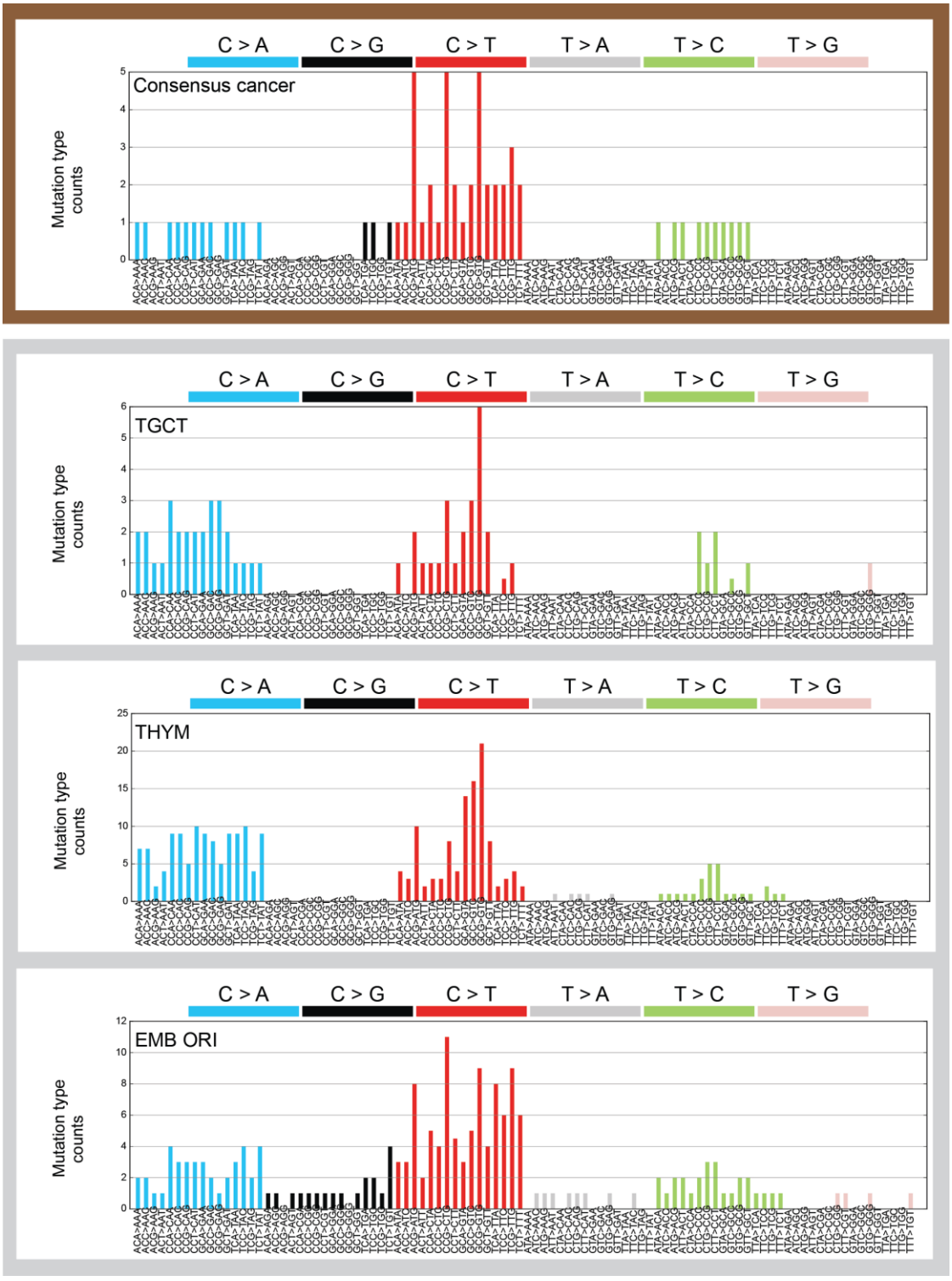

Figure 31: The unique cancer profiles identified by trinucleotide CN analysis in TCGT, THYM and EMB ORI

The top panel shows the consensus trinucleotide count patterns of all 8820 cases in the dataset by deriving the median value of each trinucleotide category. TCGT and THYM represent the signature seen in the UC clusters from these cancers, while EMB ORI represents the signature seen in the endodermal cancers.

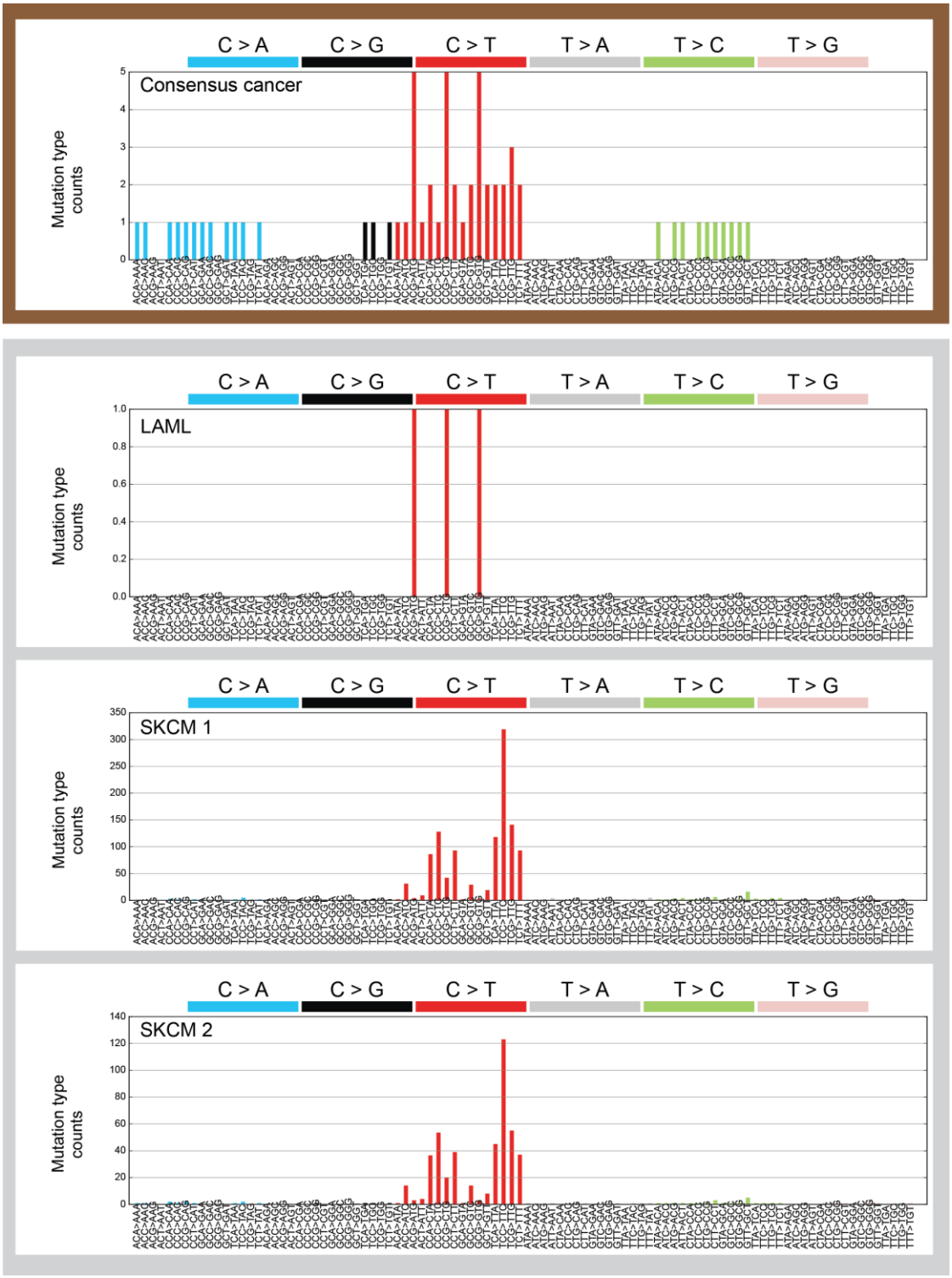

Figure 32: The unique cancer profiles identified by trinucleotide CN analysis in LAML, SKCM 1 and SKCM 2

The top panel shows the consensus trinucleotide count patterns of all 8820 cases in the dataset by deriving the median value of each trinucleotide category. LAML represents the signature seen in the SCS cluster, while SKCM 1 and 2 represent the signature seen in the two separate US clusters from that cancer.

2.3.4.2. Indels sizes by cases

Figure 33 and Figure 34 show the clustering results from the PR and CN indel analysis of all cases, in a similar matter to the trinucleotides in the previous section 2.3.4.1. Both analyses reveal that the vast majority of mutations are 1 base deletions and insertions, with subsets of cancers distinguished by larger mutations. As shown by the green bounding box, several cases lack indels entirely, specifically a large number of thyroid carcinoma (THCA), ovarian serous cystadenocarcinoma (OV) and acute myeloid leukaemia (LAML) cases. The PR analysis revealed three UC cancers, namely pancreatic adenocarcinoma (PAAD), stomach adenocarcinoma MSI (STAD-MSIH) and THCA. While the CN revealed only two UC clusters, from PAAD, also observed by PR analysis and testicular germ cell tumours (TGCT).

As with the proportions CN analysis, there is are comprehensive comparisons that can be made concerning indel size signatures in existing literature, as such these signatures were separately compared to all other cancers in order to understand the unique features these unique signatures. Table 9 shows the five most statistically altered indels from the unique cancers by both PR and CN analysis, while the entire profiles (all 12 possible indel changes) compared to a consensus representation of all cancers are shown in Figure 35 and Figure 36. As seen in Table 9, STAD-MSI cases have much higher proportions of 1 base insertions (mean: 0.85 vs 0.44) and deletions (mean: 0.90 vs 0.53) and much lower rates of 5 base indels and 3 base deletions. PAAD has higher proportions of 3 base deletions and 1 base insertions. This cancer type also has much higher counts of 3 base deletions (mean: 84.8 vs 1.84) and greater than 5 based insertions (mean: 15.22 vs 0.36). THCA is characterized by an high proportion of 1 base deletions (mean:0.98 vs 0.44) but few 1 base deletions (mean:0.00 vs 0.55), and

low relative counts of greater than 5 base insertions and 3 base insertions (mean:0.00 vs 0.42). Within the indel analysis, there were fewer discernible groups when analysing by CN vs PR, a similar observation to the trinucleotide analysis. When compared to the trinucleotide analysis, indels were not as capable of identifying unique cancer groups, but still capable provide an interesting dimension of unique cancer differences.

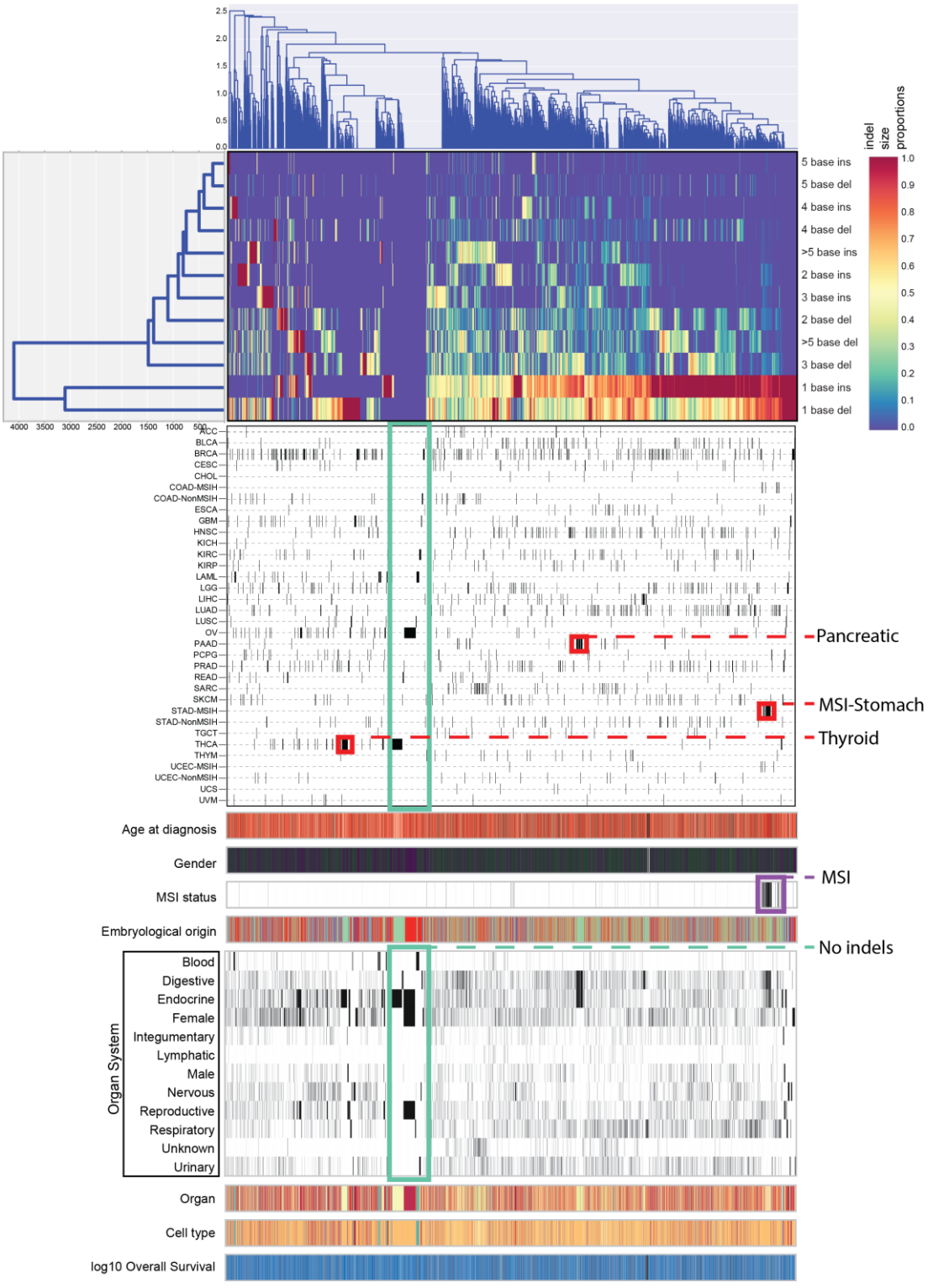

Figure 33: Clustering of indel size proportions in all cases by using average metric and city block linkage

The legend for this figure is inserted as Figure 27. Clustering has been performed on all cases in the TCGA according to the proportions of indel sizes changes. The x-axis dendrogram shows clustering of the data by case, while the y-axis dendrogram shows clustering by the mutation. The top heat map shows the proportion of mutations by case (each column is a total of 1). UC clusters are shown with a red boundary and the MSI cases cluster are shown with a purple boundary. The subset of cases without indels is shown with a green boundary.

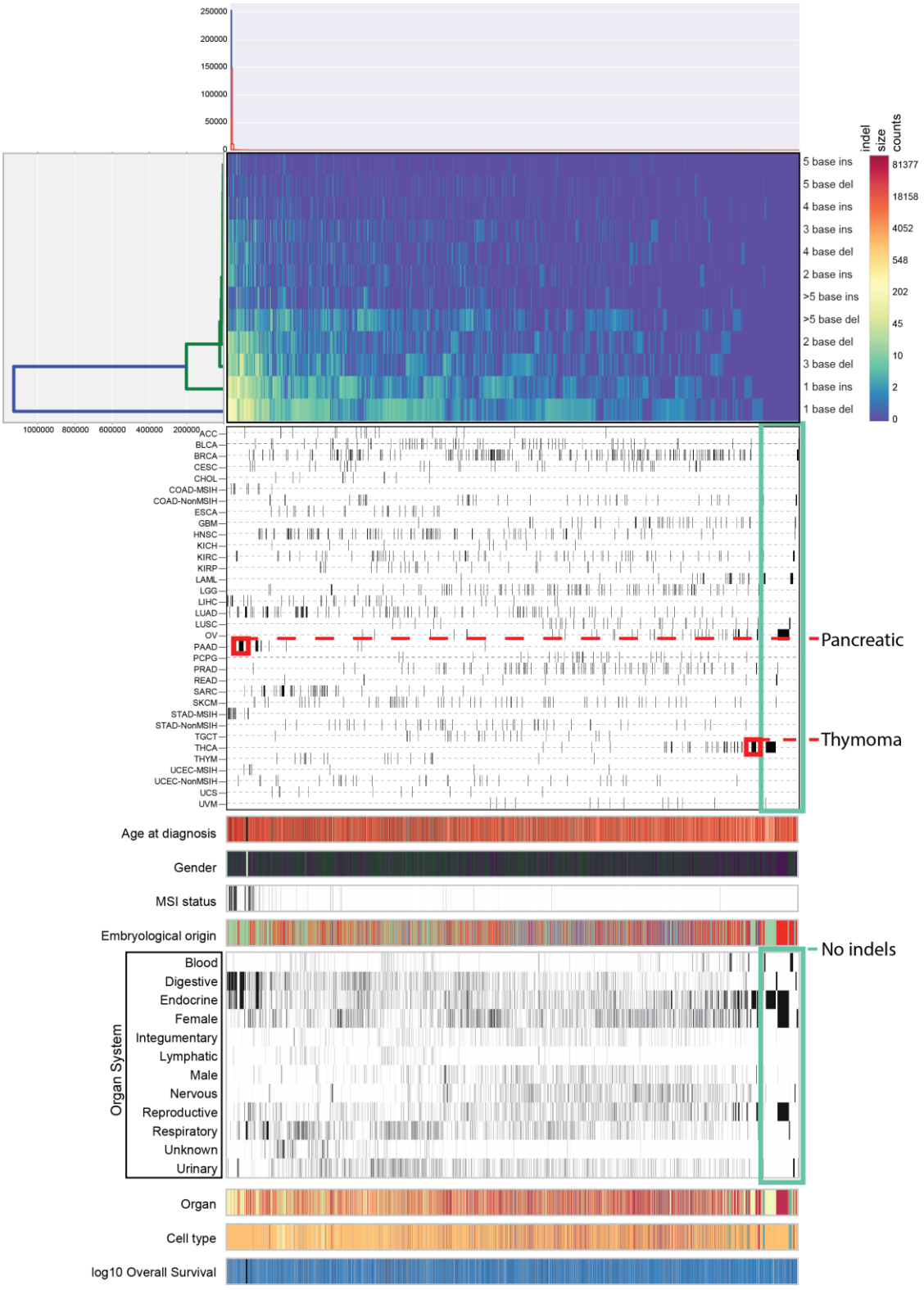

Figure 34: Clustering of indel size counts in all cases by using average metric and city block linkage.

The legend for this figure is inserted as Figure 27. Clustering has been performed on all cases in the TCGA according to the counts of indel sizes changes. The x-axis dendrogram shows clustering of the data by case, while the y-axis dendrogram shows clustering by the mutation. The top heat map shows the counts of mutations by case. UC clusters are shown with a red boundary. The subset of cases without indels is shown with a green boundary.

Table 9: The five most statistically different indel sizes observed in the uniquely clustered cancers by PR and CN analysis

Analysis type	Cancer	Indel change	Mean value of cancer cluster	Mean value of all other cases	P value	fdr_bh
Proportions	STAD-MSIH	Del1	0.85	0.44	4.9E-52	5.94E-51
		Ins1	0.90	0.53	2.8E-25	1.67E-24
		Del>5	0.01	0.15	2.3E-13	9.28E-13
		Del3	0.07	0.16	1.2E-06	3.58E-06
		Ins>5	0.01	0.07	0.00023	0.000563
	PAAD	Del3	0.46	0.15	1.6E-51	1.91E-50
		Ins1	0.82	0.54	5E-12	3E-11
		Del>5	0.04	0.15	1.3E-06	5.13E-06
		Del4	0.01	0.04	0.00594	0.017806
		Del5	0.00	0.01	0.08006	0.192148
	THCA	Del1	0.98	0.44	2.1E-69	2.53E-68
		Ins1	0.00	0.55	7.4E-42	4.45E-41
		Del>5	0.00	0.15	2E-12	6.33E-12
		Del3	0.01	0.16	2.1E-12	6.33E-12
		Del2	0.00	0.11	2.4E-10	5.85E-10
Counts	PAAD	Del3	84.82	1.84	0	0
		Ins>5	15.22	0.36	3E-237	1.7E-236
		Ins5	1.32	0.07	3E-133	1.3E-132
		Del>5	7.07	1.32	1E-105	4.4E-105
		Del2	14.10	1.57	6.6E-88	1.58E-87
	THCA	Del>5	0.00	1.39	6.5E-09	7.83E-08
		Ins3	0.00	0.42	9.8E-05	0.000587
		Del4	0.00	0.46	0.00022	0.000896
		Ins2	0.00	0.49	0.00159	0.004625
		Del5	0.00	0.15	0.00193	0.004625

Shown are the five most statistically differentially mutated indels for each of the unique cancer profiles derived from the indel PR and CN clustering of all cases. The mean value for the unique cancer profiles and all other cases is shown, along with the t-test p-value and the benjamini hochberg false discovery rate value.

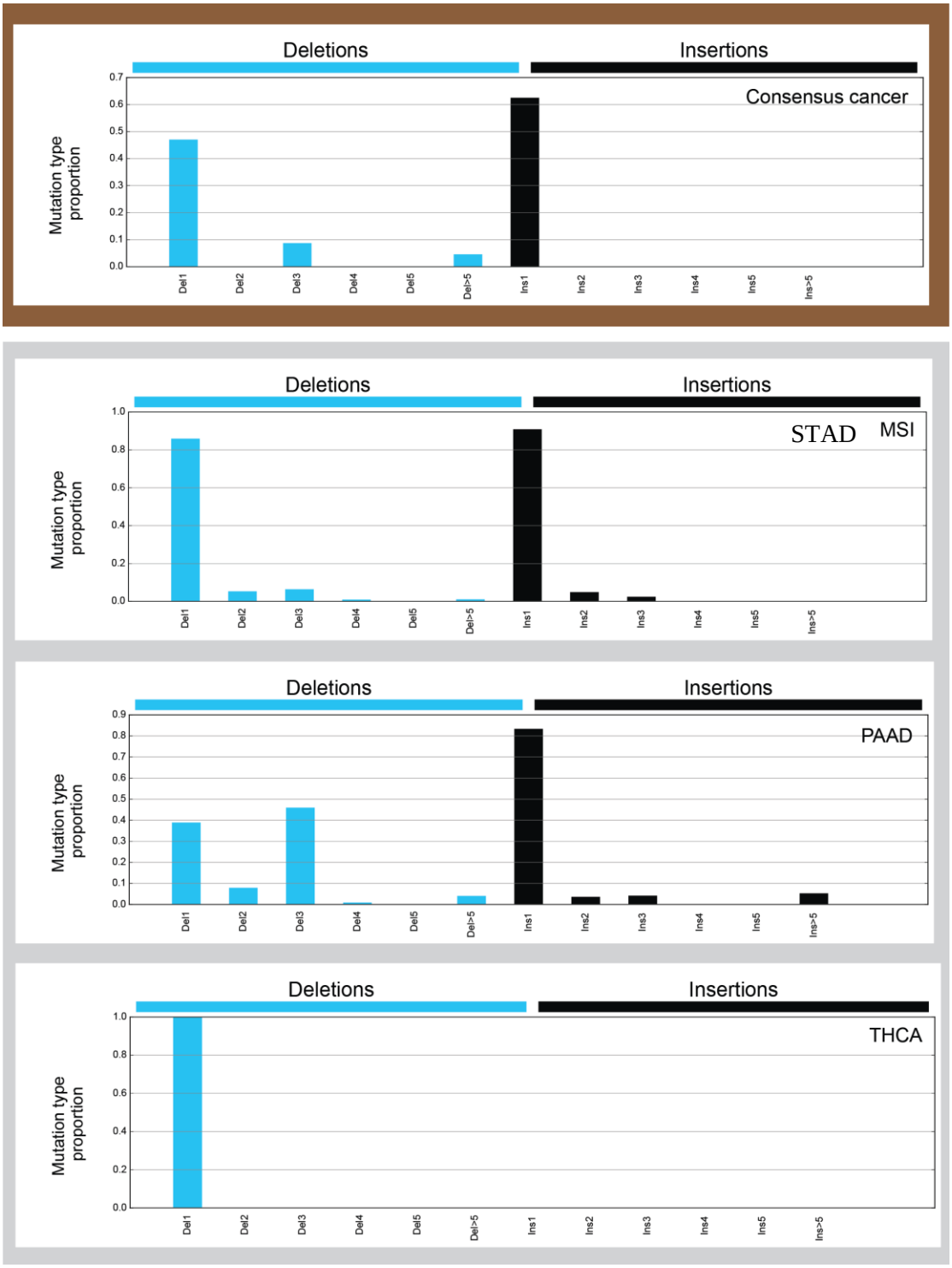

Figure 35: The unique cancer profiles identified by indel PR analysis in STAD MSI-high, PAAD and THCA

The top panel shows the consensus indel proportions pattern of all 8820 cases in the dataset by deriving the median value of each indel category. MSI-high, PAAD and THCA represent the signatures seen in the UC clusters from these cancers.

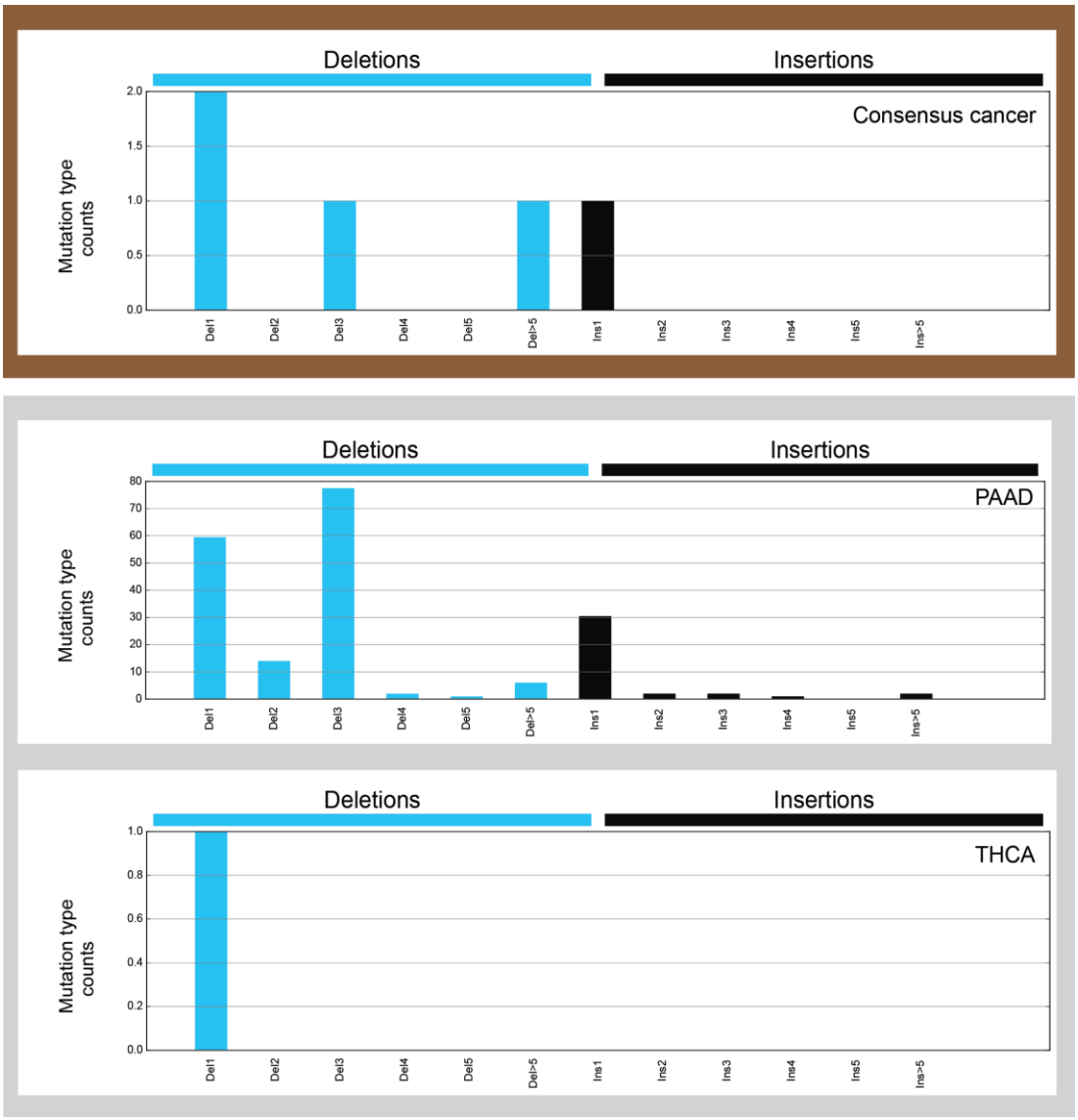

Figure 36: The unique cancer profiles identified by indel CN analysis in PAAD and THCA

The top panel shows the consensus indel count pattern of all 8820 cases in the dataset by deriving the median value of each indel category. PAAD and THCA represent the signatures seen in the UC clusters from these cancers.

2.3.4.3. Genomic distribution of mutations by cases

Figure 37 and Figure 38 show the clustering results from the PR and CN genomic distribution analysis of all cases. The PR analysis revealed that acute myeloid leukaemia (LAML) formed a UC cluster, pancreatic adenocarcinoma (PAAD) formed a US cluster and colon adenocarcinoma (COAD-MSIH), skin cutaneous melanoma (SKMC) and stomach adenocarcinoma (STAD-MSIH) formed SCS clusters as did the MSI cases as a whole. As with the consensus genomic distribution analysis (2.3.3.3) and indel counts analysis (2.3.4.2), a segregation of the endodermal and mesodermal cancers (EMB_ORI) was also observed. By the CR analysis, LAML, STAD-MSI, thymoma (THYM) and uterine corpus endometrial carcinoma (UCES-MSIH) formed SCS clusters and also segregated the endodermal and mesodermal cancers. A consistent observation seen here and in the other dimensions, is that PRs tend to discern more cancer specific mutational signatures, however, CNs do provide a slight degree of additional detail, e.g. THYM and UCES-MSIH SCS clusters are only seen with counts.

The identified cancer-specific signatures were separately compared to all other cancers so as to elucidate the unique mutational features of these cases, i.e. identify the specific region with differential mutation rates. This was done by performing an independent samples t-test for all the chromosomal bins between the cases of the cancers of interest versus all other cases and then performing multiple testing correction. Table 10 and Table 11 show the five most differentially mutated bins for each of the unique cancer clusters based on the PR and CN analyses, ranked by Holm-Šídák (HS) multiple correction. The mean bin value from the clustered genes (column heading ‘mean 1 Mb mutation rate of cancer of interest’) and all other cases (column heading ‘mean 1 Mb mutation rate of all other cancers’) are shown, as are the p-values of the t-test (column heading ‘1Mb mutation T-test p-value’) and the HS values (column

heading ‘1Mb mutation holm sidak multiple correction’). The column with the heading “Cancer genes” shows the occurrence of mutated cancer-related genes, derived from cross-referencing mutated genes from the mutated genes analysis (section 2.3.1) within each 1 Mb bin against the COSMIC curated cancer genes (Wellcome Trust Sanger Institute 2016a).

Mutational distributions have been speculated to be related to chromatin organization (Polak et al. 2015) likely due to the accessibility of the DNA or “openness” of the chromatin, as such, an effort was made to see if DNase-seq called peaks, which correspond to “open” chromatin states (accessible DNA) (Meyer and Liu 2014) could be linked the genomic mutation distributions rates. To do this DNase-seq peaks were obtained from the Roadmap Epigenomics Project (Kundaje et al. 2015) and a representative cell line from the ROADMAP project was chosen as a surrogate for each of the cancers of interest. The selected cell line was of the same tissue type or a similar developmental process as the cancers from the clustered mutational profiles (COAD-MSIH, LAML, SKCM, STAD-MSI, PAAD, THYM and UCES-MSIH), done because chromatin modifications are believed to be tissue specific (Bonn et al. 2012; Cotney et al. 2012; Yen and Kellis 2015) and therefore cell lines of similarity biological origin may provide similar chromatin organization patterns. As can be seen in Table 12 column 1, DNase-seq was available for 39 cell lines in the ROADMAP project database. The cancers of interest are shown in the second column to the right if the cell lines that most closely matched its tissue type. The rationale for the choice is shown in the third column. The density of accessible bases per Mb region was determined for each cell line in the ROADMAP project by aggregating all bases within DNase peaks which occur within the same bin coordinates as the mutational distributions. To elucidate the potential for differences in chromatin state, the number of accessible DNA

bases in each 1 Mb bin from the surrogate cell line was compared to all other ROADMAP DNase-seq cell lines via a one-samples T-test and then corrected for by fdr_bh, following which a rank order was then assigned corresponding to the fdr_bh values, in ascending order. The fdr_bh and rank orders for the five most differentially mutated bins for each cancer are shown in Table 10 and Table 11 with the column headings 'Accessible bases t-test fdr_bh' and 'Accessible bases t-test fdr_bh rank' respectively.

Statistically enriched regions were seen in all the clustered cancers as evidenced by differing means and associated low HS score. When specifically looking at the association with cancer-related genes, only two of the cancers showed any association in the most significantly different regions. *RUNX1* was found to be associated with LAML in the PR and CN analysis, while *DNMT3A*, *NPM1*, *FLT3* and *TET2* associated in only the PR and as would be expected, these five genes are associated with oncogenic events in leukemias (Osato 2004; Gaidzik et al. 2011; Ley et al. 2010; He 2013; Levis 2011; Weissmann et al. 2012). The PR analysis of the EMB_ORI revealed regions with mutations in *IDH1* and *PIK3CA*, where these genes were found to be more mutated in the endodermal set of cancers than the mesodermal. Interestingly, both *IDH1* (Borodovsky, Seltzer, and Riggins 2012) and *PIK3CA* (Bhattacharya, Mohd Omar, and Soong 2016; Hao et al. 2016) are cancer-related genes that have been linked to the dysregulation of metabolic pathways in cancer and *IDH1* is specifically known to be dysregulated by epigenetic modifiers (Roy, Walsh, and Chan 2014).

The rank column ('Accessible bases t-test fdr_bh rank'), indicates the ranking order of the statistical differences in chromatin state (DNas-seq) between the surrogate cells line versus all other DNAase-seq cell lines. As can be seen, there does not seem to be an association between the genomic distributions of mutations (per Mb mutations

rate) (column heading “1Mb mutation holm sidak multiple correction”) and this ranking. Although the presented tables merely display the five most statistically significant regions, this lack of concordance is seen throughout the distribution of both corrected p-values. On its surface, this lack of association suggests that mutation rate and chromatin state may not be related, in contrast to previous work of Polak et al (Polak et al. 2015). However, there may be other explanations for this. For example, this observation could be indicative of the inappropriateness of the use of the surrogate cell lines to represent the patient-derived data used in the mutation analysis. However, if the assumption is made that the surrogate model is correct, then the lack of concordance could be due to the fact that the accessibility of DNA alone, as determined by DNase-seq, is not an adequate determinant of DNA exposure to mutational effects. Histone marks or other epigenetic markers may instead play a role in mutational processes and may be further areas of investigation. RNA expression of the various genes may also be used as an indirect measure of DNA accessibility (Blatti et al. 2015). Taking this approach with the TCGA data would have the advantage of being able not only to use patient clinical data but to also match, case for case, the expression data with the mutation data used in this study and can be other avenues of investigation in future work.

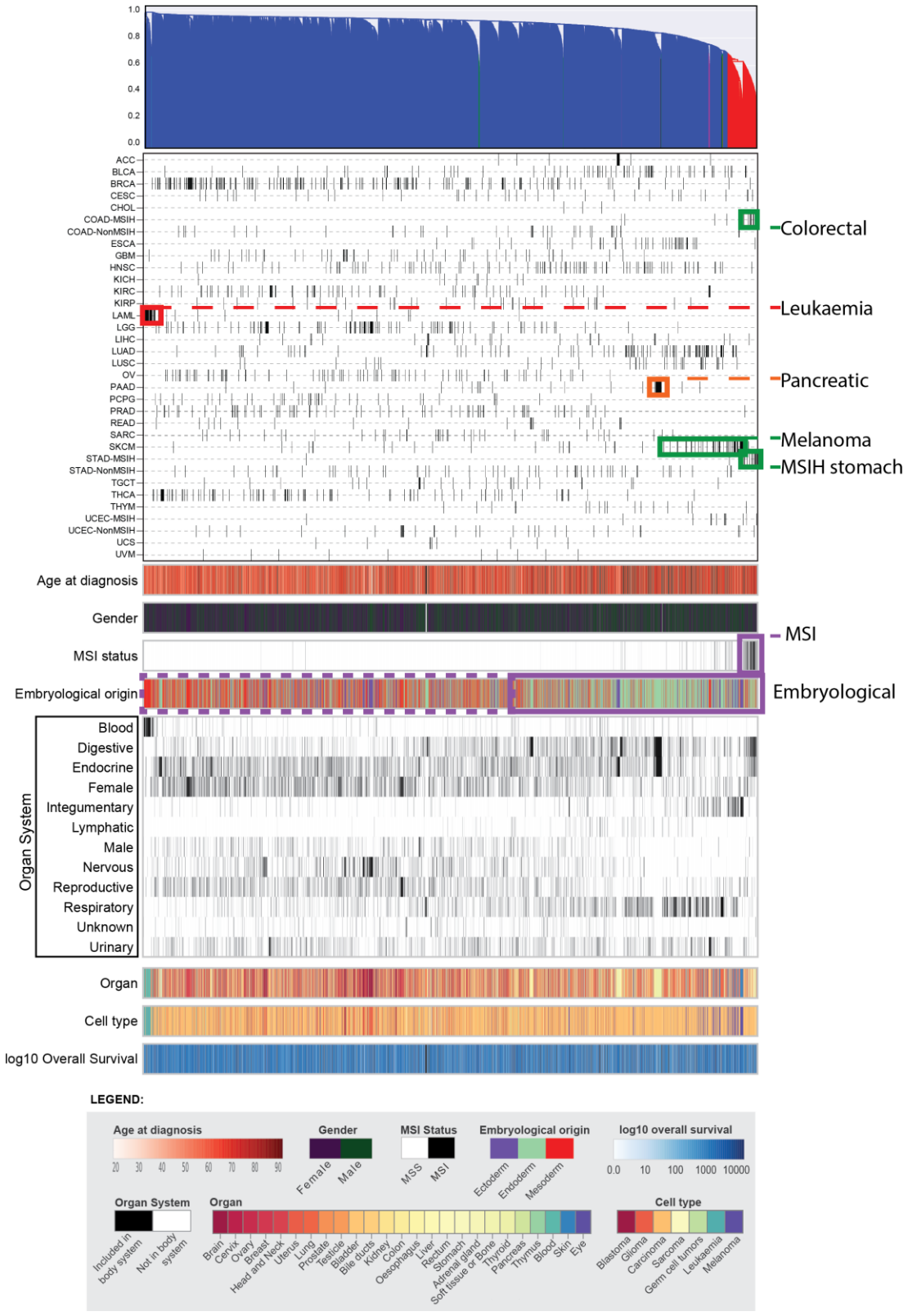

Figure 37: Clustering of genomic distribution proportions in all cases by using average metric and city block linkage.

Clustering has been performed on all cases in the TCGA according to the proportions of the distribution of mutations per Mb bin. Cases are clustered along the x-axis. UC clusters are shown with a red boundary, US clusters with an orange boundary, SCS clusters are shown with a green boundary and the MSI cases cluster and endodermal case clustering is shown with a purple boundary.

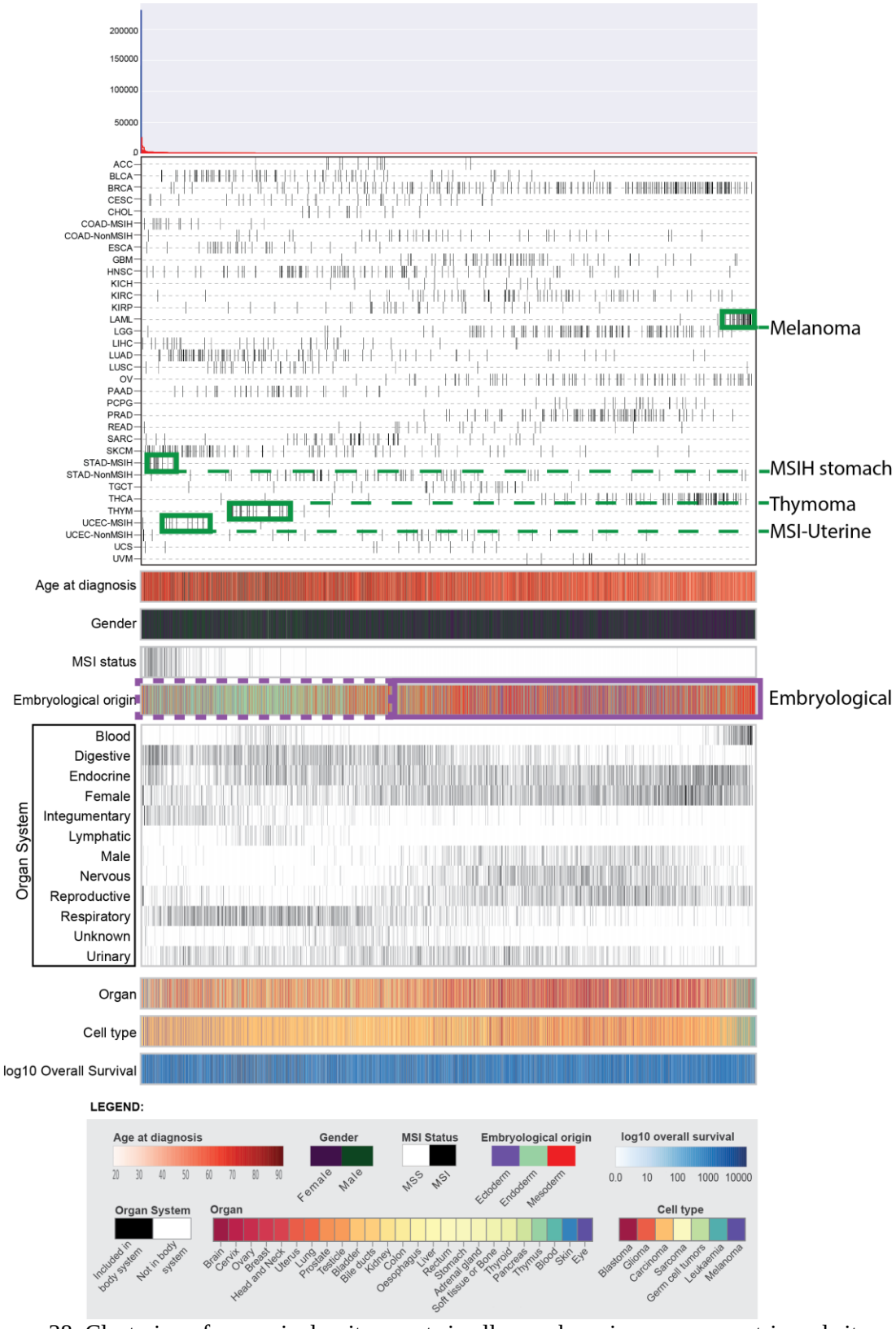

Figure 38: Clustering of genomic density counts in all cases by using average metric and city block linkage.

Clustering has been performed on all cases in the TCGA according to the counts of the distribution of mutations per Mb bin. Cases are clustered along the x-axis. SCS clusters are shown with a green boundary and the MSI cases cluster and endodermal case clustering are shown with a purple boundary.

Table 10: Unique mutational features revealed by genomic distribution analysis (1)

		1 Mb region	mean 1 Mb mutation rate of cancer of interest	mean 1 Mb mutation rate of all other cancers	1Mb mutation T-test p-value	1Mb mutation holm sidak multiple correction	Cancer genes	Accessible bases t-test fdr_bh	Accessible bases t-test fdr_bh rank
LAML	**CN**	chr14:84000001-85000000	0.01	0.00	0	0.76		0	727
		chr21:36000001-37000000	0.11	0.03	0	0.99	RUNX1	0.24	2615
		chrX:73000001-74000000	0.01	0.21	0.00	1.00		0	1186
		chr9:68000001-69000000	0	0.11	0.01	1.00		0	1024
		chr20:29000001-30000000	0	0.35	0.01	1.00		0	2152
	PR	chr2:25000001-26000000	0.05	0	0	0	DNMT3A	0.03	2347
		chr5:170000001-171000000	0.06	0	0	0	NPM1	0	354
		chr13:28000001-29000000	0.07	0	0	0	FLT3	0	1638
		chr4:106000001-107000000	0.02	0	0	0	TET2	0.01	2228
		chr21:36000001-37000000	0.01	0	0	0	RUNX1	0.24	2615
STAD _MSIH	**CN**	chr8:22000001-23000000	1.72	0.18	0	0		0	695
		chr4:134000001-135000000	0.53	0.06	0	0		0	699
		chr8:21000001-22000000	0.86	0.10	0	0		0	1045
		chr8:23000001-24000000	0.65	0.09	0	0		0	600
		chr8:1-1000000	0.46	0.05	0	0		0	981
	PR	chrY:27000001-28000000	0	0	0	0		1.00	2845
		chr20:29000001-30000000	0	0	0	0.98		0	12
		chr3:195000001-196000000	0	0	0	1.00		0	717
		chr21:11000001-12000000	0	0	0.01	1.00		0	415
		chr8:22000001-23000000	0	0	0.02	1.00		0	695
EMB_ ORI	**CN**	chrX:73000001-74000000	0.45	0.04	0	0			
		chr8:10000001-11000000	0.38	0.04	0	0			
		chr8:2000001-3000000	0.28	0.02	0	0			
		chr15:25000001-26000000	0.58	0.05	0	0			
		chr8:24000001-25000000	0.29	0.03	0	0			
	PR	chr5:140000001-141000000	0.01	0.01	0	0			
		chr2:209000001-210000000	0	0	0	0	IDH1		
		chr3:178000001-179000000	8E-04	0.002	3E-26	0	PIK3CA		
		chr2:179000001-180000000	0.005	0.003	2E-23	0			
		1_152000001_153000000	0.005	0.003	4E-21	0			

Table 11: Unique mutational features revealed by genomic distribution analysis (2)

		1 Mb region	mean 1 Mb mutation rate of cancer of interest	mean 1 Mb mutation rate of all other cancers	1Mb mutation T-test p-value	1Mb mutation holm sidak multiple correction	Cancer genes	Accessible bases t-test fdr_bh	Accessible bases t-test fdr_bh rank
UCEU _MSIH	**CN**	chr4:134000001-135000000	0.35	0.06	0	0		0.90	2827
		chr8:22000001-23000000	0.99	0.19	0	0		0	1385
		chrX:39000001-40000000	0.27	0.05	0	0	BCOR	0	397
		chr5:67000001-68000000	0.41	0.05	0	0	PIK3R1	0	1589
		chr8:23000001-24000000	0.40	0.09	0	0		0	1719
THYM	**CN**	chr4:45000001-46000000	0.01	0	0	0		0	1305
		chr12:84000001-85000000	0.01	0	0	0		0	881
		chrY:26000001-27000000	0.01	0	0	0		1.00	2858
		chr14:83000001-84000000	0.02	0	0	0		0	1187
		chr1:142000001-143000000	0.35	0.11	0	0.05		0	1542
PAAD	**PR**	chr1:121000001-122000000	0	0	0	0		0	714
		chr9:20000001-21000000	0	0	0	0		0.03	1692
		chr7:142000001-143000000	0.01	0	0	0		0.11	2012
		chr20:29000001-30000000	0.01	0	0	0		0	12
		chr7:114000001-115000000	0	0	0	0		0.23	2204
SKCM	**PR**	chr2:179000001-180000000	0.01	0	0	0		0.44	2484
		chr19:9000001-10000000	0.01	0	0	0		0.67	2650
		chr14:22000001-23000000	0	0	0	0		0.13	2183
		chr17:10000001-11000000	0	0	0	0		0	1249
		chr18:28000001-29000000	0	0	0	0		0.27	2356
COAD _MSIH	**PR**	chrY:27000001-28000000	0	0	0	0		1.00	2842
		chr20:29000001-30000000	0	0	0	1.00		0.01	1645
		chr3:195000001-196000000	0	0	0.01	1.00		0.01	1562
		chr21:11000001-12000000	0	0	0.01	1.00		0.01	1652
		chr8:22000001-23000000	0	0	0.02	1.00		0	853

Table 12: Cell lines from the ROADMAP project that correspond to cancers with unique mutational distribution profiles

ROADMAP_Cell_name	Cancer_type	Rationale of choice of this cell line
H1 Cells (human embryonic stem cells)		
H1 BMP4 Derived Mesendoderm Cultured Cells	Meso	Mesendoderm cell
H1 BMP4 Derived Trophoblast Cultured Cells		
H1 Derived Mesenchymal Stem Cells	Ecto	mesenchyme derived from ectoderm
H1 Derived Neuronal Progenitor Cultured Cells		
H9 Cells (human embryonic stem cells)		
IMR90 fetal lung fibroblasts Cell Line		
iPS DF 6.9 Cells		
iPS DF 19.11 Cells		
Breast variant Human Mammary Epithelial Cells (vHMEC)		
Primary monocytes from peripheral blood	LAML	Myeloid cell
Primary B cells from peripheral blood		
Primary T cells from cord blood		
Primary T cells from peripheral blood		
Primary Natural Killer cells from peripheral blood		
Primary hematopoietic stem cells G-CSF-mobilized Female		
Primary hematopoietic stem cells G-CSF-mobilized Male		
Foreskin Fibroblast Primary Cells skin01		
Foreskin Fibroblast Primary Cells skin02		
Foreskin Keratinocyte Primary Cells skin02		
Foreskin Melanocyte Primary Cells skin01	SKCM	Melanocyte
Fetal Adrenal Gland		
Fetal Brain Male		
Fetal Brain Female		
Fetal Heart		
Fetal Intestine Large	COAD_MSIH	Intestine
Fetal Intestine Small		
Fetal Kidney		
Fetal Lung		
Fetal Muscle Trunk		
Fetal Muscle Leg		
Placenta		
Fetal Stomach		
Fetal Thymus	THYM	Thymus
Gastric	STAD_MSIH	Adult stomach (not fetal)
Ovary	UCEC-MSIH	Closest matching tissue type in terms of developments progression and body system
Pancreas	PAAD	
Psoas Muscle		
Small Intestine		

Cell line available via the ROADMAP database. The leftmost column shows all available cell lines the second column show the cancer type with the unique clustering profile next to the ROADMAP cell line selected to represent it. The rationale for the choice may not be immediately obvious and is therefore stated is the third column.

2.3.4.4. Genes and variants by cases

As with the other analyses, the clustering of the mutated genes, shown in Figure 39, exhibited heterogeneity in the mutation patterns among the cases within each cancer type, however, this approach did seem to reveal the greatest number of unique sub-cancer clusters. Pancreatic adenocarcinoma (PAAD) cases formed into a UC cluster, while kidney renal clear cell carcinoma (KIRC), acute myeloid leukaemia (LAML), skin cutaneous melanoma (SKCM) and thyroid carcinoma (THCA) formed US clusters, as did all MSI cases. Breast invasive carcinoma (BRCA) and brain lower grade glioma (LGG) formed two US clusters each. Table 13 and Table 14 lists the 5 most variable mutated genes between each unique cancer cluster and all other cancers. The mean number of cases with mutation(s) for each gene is shown for the clustered cancers in the column 'mean mutation rate - clustered' and all other cases in the column 'mean mutation rate – others'. A fisher's exact test was used to compare the mutation rates in the clustered cancers vs all other cases, with the p-values and odds ratio shown in the appropriately named columns. Fdr_bh correction was performed and was then used to rank the genes according to statistical significance. The novelty of the approach taken in this study is that other cancers mutation studies typically compare mutation rates in just a cancer alone, i.e. tumour vs normal. The findings here identify genes that are specific to cancers types, i.e. not seen in other cancers or seen at much lower rates.

BRCA had twelve genes that had significantly different rates of mutation in each cluster, based on an fdr_bh threshold of less than 0.05 (Table 13). Specifically, differences in the mutation rates in *GATA3* in both clusters, along with *PIK3CA* and *MAP3K1* among other genes in either one of the clusters. Seven genes were significantly different in KIRC, but specifically higher rates of *VHL* and *PBRM1* mutations were seen. 10 significantly mutated genes were identified in the LAML,

including *RUNX1*, *FLT3*, *NPM1* and *IDH2* of which the former two were identified by the genomic distribution analysis in the section 2.3.4.3. One cluster of LGG had sixty-eight significantly different genes while the other just four (Table 14). In both clusters, *IDH1* was the most significant gene. Over 16 thousand genes were significantly different in the MSI cases compared to all other cases, obviously due to the exceedingly high mutation rates of these cancers (Figure 6) of which most are random passenger mutations, however, *RNF43*, *MLL2* and *ARID1A* have particularly high rates of mutations in the MSI cases. PAAD, the only UC clustered cancer, has 1085 significantly different genes, most notably, *KRAS*, *RIOK1* and *JMY*. 6994 genes were differentially mutated in SKCM, while that figure was 213 was in THCA. The most significant genes in SKCM are the glycoprotein genes *THSD7B* and *MUC16* and structural genes *PCLO*, *DNAH5* and *TTN*. THCA is especially distinct in the number of *BRAF* mutations.

The clustering of the variants with a frequency of four or more in the TCGA dataset is as shown in Figure 40. ACC, LAML and PAAD grouped as a UC clusters, while BRCA, LGG, PCPG THCA and all the MSI cases formed US clusters. Table 15 and Table 16 show the 5 most statically different variants that distinguish each cluster, ranked according to the fdr_bh. Interesting the most statistically different variants in each cancer cluster were unique to that cluster. Annotation of the 5 most statistically different variants showed that the mutated genes were also unique to each cluster.

Overall, the clustering and subsequent identification of unique features of these clusters show that, at least in part, certain cancers or subsets of cancers can be identified based on the mutated genes or variants alone.

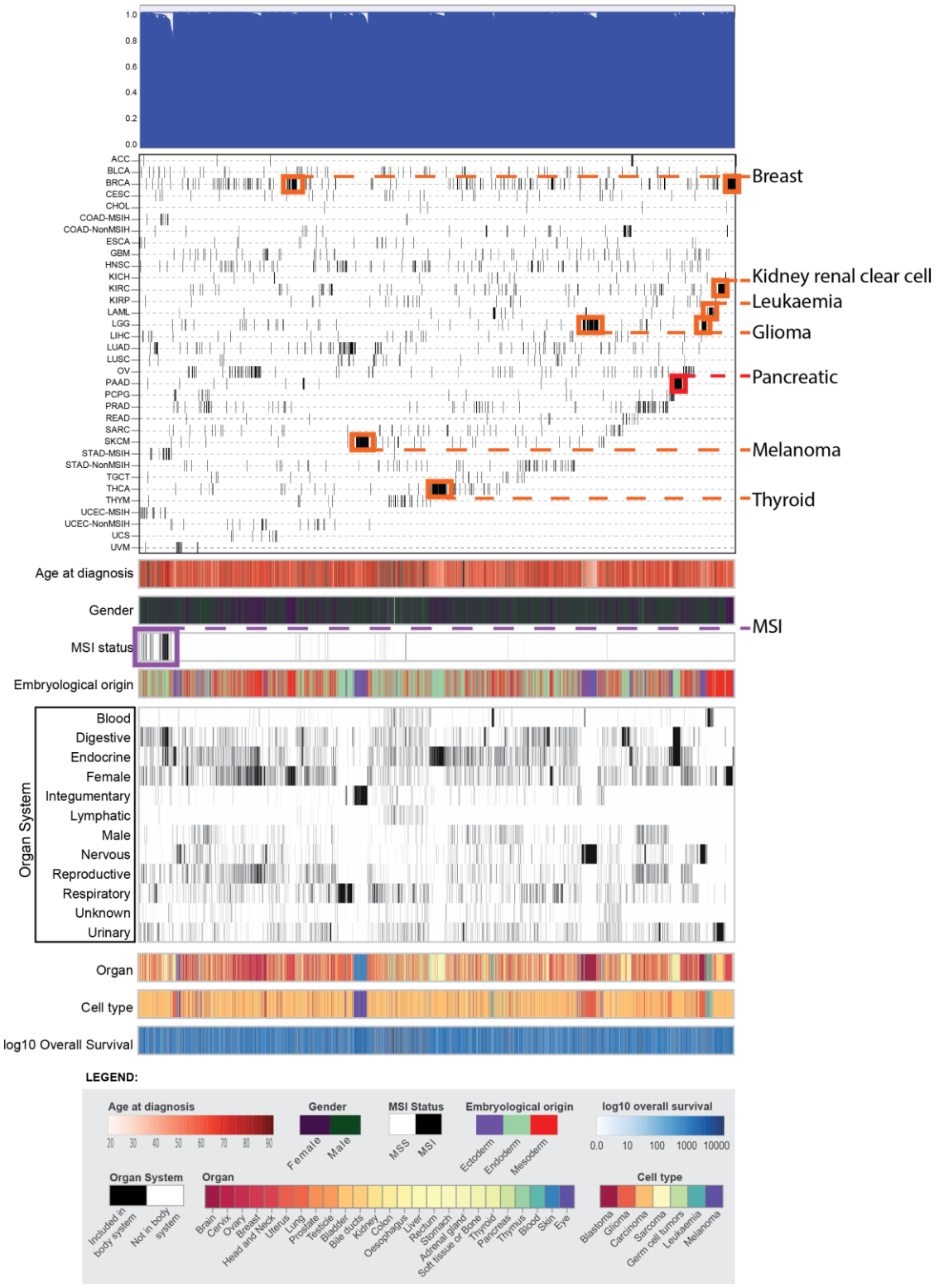

Figure 39: Clustering of mutated genes in all cases by using complete metric and Jaccard linkage.

Clustering has been performed on all cases in the TCGA according to the 24118 genes with mutations. The x-axis dendrogram shows clustering of the data by case. UC clusters are shown with a red boundary, SCS clusters are shown with a green boundary and the MSI cases cluster and endodermal case clustering are shown with a purple boundary.

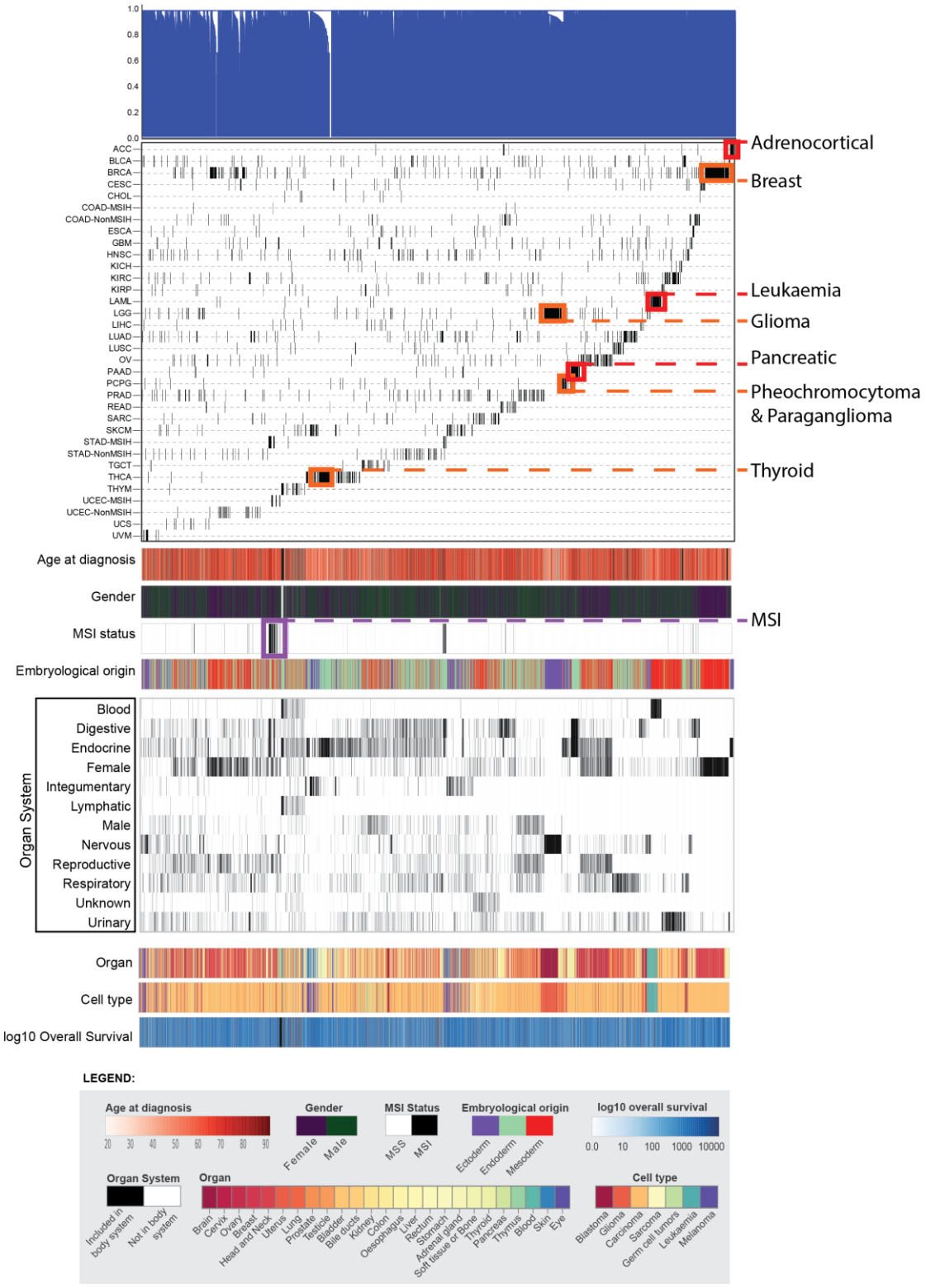

Figure 40: Clustering of mutated variants in all cases by using complete metric and Jaccard linkage.

Clustering has been performed on all cases in the TCGA according to 42,030 recurrent variants. The x-axis dendrogram shows clustering of the data by case. UC clusters are shown with a red boundary, SCS clusters are shown with a green boundary and the MSI cases cluster and endodermal case clustering are shown with a purple boundary.

Table 13: Most statically different genes as determined by the clustering of mutated genes frequencies (1)

Cancer	Gene symbol	Name	Odds_Ratio	p_value	fdr_bh	mean mutation rate - clustered	mean mutation rate – others	cluster_pos	cluster_neg	non_cluster_pos	non_cluster_neg
BRCA_1	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	117.41	8E-101	2E-96	0.94	0.01	122	8	999	7691
	MAP3K1	mitogen-activated protein kinase kinase kinase 1	11.959	8E-19	1E-14	0.22	0.34	28	102	195	8495
	TTN	titin	0.251	1E-08	0.0001	0.11	0.04	14	116	2822	5868
	TP53	tumor protein p53	0.3043	6E-08	0.0004	0.15	0.03	20	110	3250	5440
	GATA3	GATA binding protein 3	6.0823	8E-08	0.0004	0.12	0.37	16	114	196	8494
BRCA_2	GATA3	GATA binding protein 3	37.88	8E-49	2E-44	0.42	0.29	48	66	164	8542
	TP53	tumor protein p53	0.0926	1E-15	2E-11	0.05	0.03	6	108	3264	5442
	TTN	titin	0.0952	3E-13	2E-09	0.04	0.04	5	109	2831	5875
	CBFB	core-binding factor beta subunit	18.115	1E-08	6E-05	0.08	0.72	9	105	41	8665
	MUC16	mucin 16, cell surface associated	0.136160515	1.52293E-07	0.000524716	0.04	0.06	4	110	1835	6871
KIRC	VHL	von Hippel-Lindau tumor suppressor	400.35	3E-146	7E-142	0.89	0.06	102	12	181	8525
	PBRM1	polybromo 1	23.587	6E-45	7E-41	0.48	0.15	55	59	331	8375
	TP53	tumor protein p53	0.0297	2E-20	2E-16	0.02	0.03	2	112	3268	5438
	MUC16	mucin 16, cell surface associated	0.0668	2E-09	1E-05	0.02	0.06	2	112	1837	6869
	BAP1	BRCA1 associated protein 1	6.6692	2E-07	0.0008	0.12	0.36	14	100	179	8527
LAML	FLT3	fms related tyrosine kinase 3	31.503	2E-31	5E-27	0.42	0.19	32	45	193	8550
	NPM1	nucleophosmin	42.856	2E-23	2E-19	0.26	0.45	20	57	71	8672
	IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial	25.342	2E-14	2E-10	0.18	0.45	14	63	76	8667
	RUNX1	runt related transcription factor 1	18.826	9E-13	5E-09	0.18	0.38	14	63	102	8641
	CEBPA	CCAAT/enhancer binding protein alpha	60.771	1E-12	5E-09	0.12	0.78	9	68	19	8724
LGG_1	IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	1745.1	8E-265	2E-260	0.99	0.01	206	3	326	8285
	ATRX	ATRX, chromatin remodeler	45.785	3E-132	3E-128	0.72	0.12	150	59	453	8158
	TP53	tumor protein p53	6.8491	1E-36	9E-33	0.79	0.01	166	43	3104	5507
	TTN	titin	0.1937	5E-16	3E-12	0.09	0.06	18	191	2818	5793
	CSMD1	CUB and Sushi multiple domains 1	0	1E-08	5E-05	0.00	0.22	0	209	755	7856

Table 14: Most statically different genes as determined by the clustering of mutated genes frequencies (2)

Cancer	Gene symbol	Name	Odds_Ratio	p_value	fdr_bh	mean mutation rate - clustered	mean mutation rate – others	cluster_pos	cluster_neg	non_cluster_pos	non_cluster_neg
LGG_2	IDH1	isocitrate dehydrogenase (NADP(+)) 1, cytosolic	135.85	4E-82	9E-78	0.88	0.02	75	10	457	8278
	CIC	capicua transcriptional repressor	113.23	8E-80	9E-76	0.78	0.07	66	19	260	8475
	FUBP1	far upstream element binding protein 1	29.416	4E-25	3E-21	0.29	0.33	25	60	122	8613
	TP53	tumor protein p53	0.1047	3E-11	2E-07	0.06	0.02	5	80	3265	5470
MSIH	RNF43	ring finger protein 43	85.899	2E-118	4E-114	0.60	0.31	100	66	150	8504
	MLL2	myeloid/lymphoid or mixed-lineage leukemia protein 2	55.308	5E-113	6E-109	0.69	0.13	115	51	339	8315
	ARID1A	AT-rich interaction domain 1A	38.886	2E-100	2E-96	0.73	0.07	122	44	576	8078
	RPL22	ribosomal protein L22	77.572	3E-76	2E-72	0.37	0.61	62	104	66	8588
	ZFHX3	zinc finger homeobox 3	24.035	1E-68	5E-65	0.53	0.17	88	78	388	8266
PAAD	KRAS	KRAS proto-oncogene, GTPase	43.093	3E-73	8E-69	0.75	0.05	85	28	573	8134
	RBM14-RBM4		222.24	4E-68	4E-64	0.38	0.74	43	70	24	8683
	RIOK1	RIO kinase 1	70.546	7E-55	5E-51	0.39	0.47	44	69	78	8629
	FRG1B		23.247	1E-53	6E-50	0.78	0.02	88	25	1145	7562
	JMY	junction mediating and regulatory protein, p53 cofactor	51.823	3E-49	1E-45	0.38	0.41	43	70	102	8605
SKCM	MUC16	mucin 16, cell surface associated	40.426	7E-99	2E-94	0.91	0.01	174	18	1665	6963
	DNAH5	dynein axonemal heavy chain 5	21.794	3E-85	3E-81	0.67	0.08	129	63	741	7887
	TTN	titin	37.018	5E-76	4E-72	0.94	0.00	181	11	2655	5973
	PCLO	piccolo presynaptic cytomatrix protein	17.677	2E-75	1E-71	0.66	0.07	126	66	841	7787
	MGAM	maltase-glucoamylase	21.981	6E-73	3E-69	0.51	0.20	97	95	383	8245
THCA	BRAF	B-Raf proto-oncogene, serine/threonine kinase	87.137	3E-196	8E-192	0.83	0.08	202	41	459	8118
	TP53	tumor protein p53	0.0067	4E-48	4E-44	0.00	0.07	1	242	3269	5308
	TTN	titin	0.1835	3E-19	2E-15	0.08	0.07	20	223	2816	5761
	MUC16	mucin 16, cell surface associated	0.1418	1E-14	6E-11	0.04	0.11	9	234	1830	6747
	MUC4	mucin 4, cell surface associated	0.0482	5E-14	2E-10	0.01	0.16	2	241	1259	7318

Table 15: Five most statically different variants as determined by the clustering of mutated variant frequencies (1)

Cancer	Variant	Gene symbol	mean mutation rate - clustered	mean mutation rate – others	Odds_Ratio	p_value	fdr_bh	Gene description
ACC	**chr16:88599697-88599705 AGCCTCTGG>-**	ZFPM1	0.49	0.00	381.198864	3E-43	1E-38	zinc finger protein, FOG family member 1
	chr15:63414083-63414083 A>C	LACTB	0.40	0.00	849.765306	6.1E-40	1E-35	lactamase, beta
	chr8:146033347-146033347 T>C	ZNF517	0.38	0.00	246.893417	3E-32	4E-28	zinc finger protein 517
	chr16:57562804-57562804 G>A	CCDC102A	0.34	0.00	411.120235	3E-31	3E-27	coiled-coil domain containing 102A
	chr7:30634661-30634661 C>G	GARS	0.36	0.00	275.62037	3.6E-31	3E-27	glycyl-tRNA synthetase
BRCA	**chr10:8111433-8111434 CA>-**	GATA3	0.05	0.00	139.717201	1.7E-21	7E-17	GATA binding protein 3
	chr20:46279837-46279839 CAG>-	NCOA3	0.04	0.00	158.435159	6.9E-17	1E-12	nuclear receptor coactivator 3
	chr7:36552787-36552788 ->G	AOAH	0.04	0.00	39.5806916	1.1E-13	1E-09	acyloxyacyl hydrolase (neutrophil)
	chr12:124887058-124887059 ->GCT	NCOR2	0.03	0.00	48.3142857	2.8E-11	3E-07	nuclear receptor corepressor 2
	chr16:68772218-68772218 C>T	CDH1	0.02	0.00	inf	1.8E-10	1E-06	cadherin 1, type 1, E-cadherin (epithelial)
LAML	**chr2:25457242-25457242 C>T**	DNMT3A	0.14	0.00	688.253968	1.4E-34	6E-30	DNA (cytosine-5-)-methyltransferase 3 alpha
	chr5:170837543-170837544 ->TCTG	NPM1	0.13	0.00	1297.53543	8.5E-34	2E-29	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
	chr15:90631934-90631934 C>T	IDH2	0.11	0.00	533.661538	2E-27	3E-23	isocitrate dehydrogenase 2 (NADP+), mitochondrial
	chr5:170837547-170837548 ->TCTG	NPM1	0.10	0.00	248.187023	3.4E-24	4E-20	nucleophosmin (nucleolar phosphoprotein B23, numatrin)
	chr13:28592642-28592642 C>A	FLT3	0.07	0.00	637.720588	1.2E-17	1E-13	fms-related tyrosine kinase 3
LGG	**chr2:209113112-209113112 C>T**	IDH1	1.00	0.02	inf	0	0	isocitrate dehydrogenase 1 (NADP+), soluble
	chr17:7577121-7577121 G>A	TP53	0.15	0.01	19.2685662	5.3E-30	1E-25	tumor protein p53
	chr9:139413070-139413072 AGA>-	NOTCH1	0.03	0.00	147.244635	1.2E-11	2E-07	notch 1
	chrX:76909629-76909629 G>A	ATRX	0.02	0.00	109.493617	1E-08	0.0001	alpha thalassemia/mental retardation syndrome X-linked
	chr19:42791715-42791715 C>T	CIC	0.02	0.00	181.737288	8.6E-08	0.0007	capicua transcriptional repressor

Table 16: Five most statically different variants as determined by the clustering of mutated variant frequencies (2)

Cancer	Variant	Gene symbol	mean mutation rate - clustered	mean mutation rate – others	Odds_Ratio	p_value	fdr_bh	Description
MSI	**chr1:6257785-6257785 T>-**	RPL22	0.41	0.00	153.923077	6.3E-67	3E-62	ribosomal protein L22
	chr17:56435161-56435161 C>-	RNF43	0.39	0.01	126.403664	1E-61	2E-57	ring finger protein 43
	chr2:148683686-148683686 A>-	ACVR2A	0.25	0.00	92.6006098	9.1E-39	1E-34	activin A receptor, type IIA
	chr10:890939-890939 T>-	LARP4B	0.20	0.00	197.704545	2.4E-35	3E-31	La ribonucleoprotein domain family, member 4B
	chr8:103289349-103289349 T>-	UBR5	0.21	0.00	127.660281	2.4E-34	2E-30	ubiquitin protein ligase E3 component n-recognin 5
PAAD	**chr11:66411364-66411384 (21)>-**	RBM14-RBM4	0.39	0.00	310.543379	2.7E-77	1E-72	RBM14-RBM4 readthrough
	chr6:7393450-7393452 GAC>-	RIOK1	0.35	0.00	311.769231	2.5E-69	5E-65	RIO kinase 1
	chr5:78610444-78610479 (36)>-	JMY	0.34	0.00	346.803311	1.4E-68	2E-64	junction mediating and regulatory protein, p53 cofactor
	chr1:152671515-152671556 (42)>-	LCE2A	0.29	0.00	325.256684	2.5E-58	3E-54	late cornified envelope 2A
	chr12:55615114-55615116 CTT>-	OR10A7	0.26	0.00	504.707865	4.6E-54	4E-50	olfactory receptor, family 10, subfamily A, member 7
PCPG	**chr1:248020556-248020556 G>C**	TRIM58	0.18	0.00	218.441315	3.6E-27	2E-22	tripartite motif containing 58
	chr17:59489893-59489893 T>C	C17orf82	0.13	0.00	114.763158	2.9E-17	6E-13	chromosome 17 open reading frame 82
	chr19:50881820-50881821 ->AAC	NR1H2	0.11	0.00	161.892393	9.4E-17	1E-12	nuclear receptor subfamily 1, group H, member 2
	chr17:16097825-16097825 T>G	NCOR1	0.14	0.00	49.7428571	1.7E-15	2E-11	nuclear receptor corepressor 1
	chr3:195512186-195512186 T>C	MUC4	0.08	0.00	inf	7.1E-15	6E-11	mucin 4, cell surface associated
THCA	**chr7:140453136-140453136 A>T**	BRAF	0.98	0.03	1244.67797	1E-232	4E-228	B-Raf proto-oncogene, serine/threonine kinase
	chr1:152280782-152280782 A>G	FLG	0.08	0.00	29	1.5E-15	3E-11	filaggrin
	chr1:152281479-152281479 G>T	FLG	0.04	0.00	18.2198732	1E-07	0.0014	filaggrin
	chr1:186363119-186363119 C>G	C1orf27	0.03	0.00	17.4908722	5E-06	0.0471	chromosome 1 open reading frame 27
	chr1:152281039-152281039 G>A	FLG	0.02	0.00	65.4318182	5.6E-06	0.0471	filaggrin

2.3.4.5. Multidimensional clustering analysis by cases

Several unique mutational profiles have been identified via the interrogation of the different dimensions of DNA mutations. For example, through the use of mutated genes, one UC cluster was identified and six US clusters, suggesting that the cancer type of the cases within these clusters can be determined, with a degree of certainty, based just on the mutational profiles. To improve on this, the 10 dimensions were combined into two separate datasets, one using the proportions values for trinucleotide counts, indel sizes and genomic distribution (multidimensional proportions) and the other using the counts (multidimensional counts), both of these datasets also contain the genes and variants dimensions. Figure 41 and Figure 42 represent the results from the multidimensional proportions clustering and multidimensional counts clustering respectively. The idea behind the combined analysis was to maximise the subtype differentiation using all available data perhaps overcoming the limitations of the individual dimensions of analysis. Figure 41 and Figure 42 show however that a greater discernment was not observed, but rather that cancer or sub-cancer specific clustering suffered by combining the various data dimensions, and that unique cancer subtyping with all dimensions does not appear possible via hierarchical clustering. As speculated in section 2.3.3.5, this is may be due to the fact that the unique profiles of the individual dimensions are masked when analysed by hierarchical clustering. In both the multidimensional PR and CN analyses, acute myeloid leukaemia (LAML) and stomach adenocarcinoma (STAD) from SCS clusters, while the endodermal and mesodermal cancers tend to cluster separately. No UC or US clusters were observed.

A summary of the cluster types and corresponding cancers from each dimension of analysis is shown in Table 17.

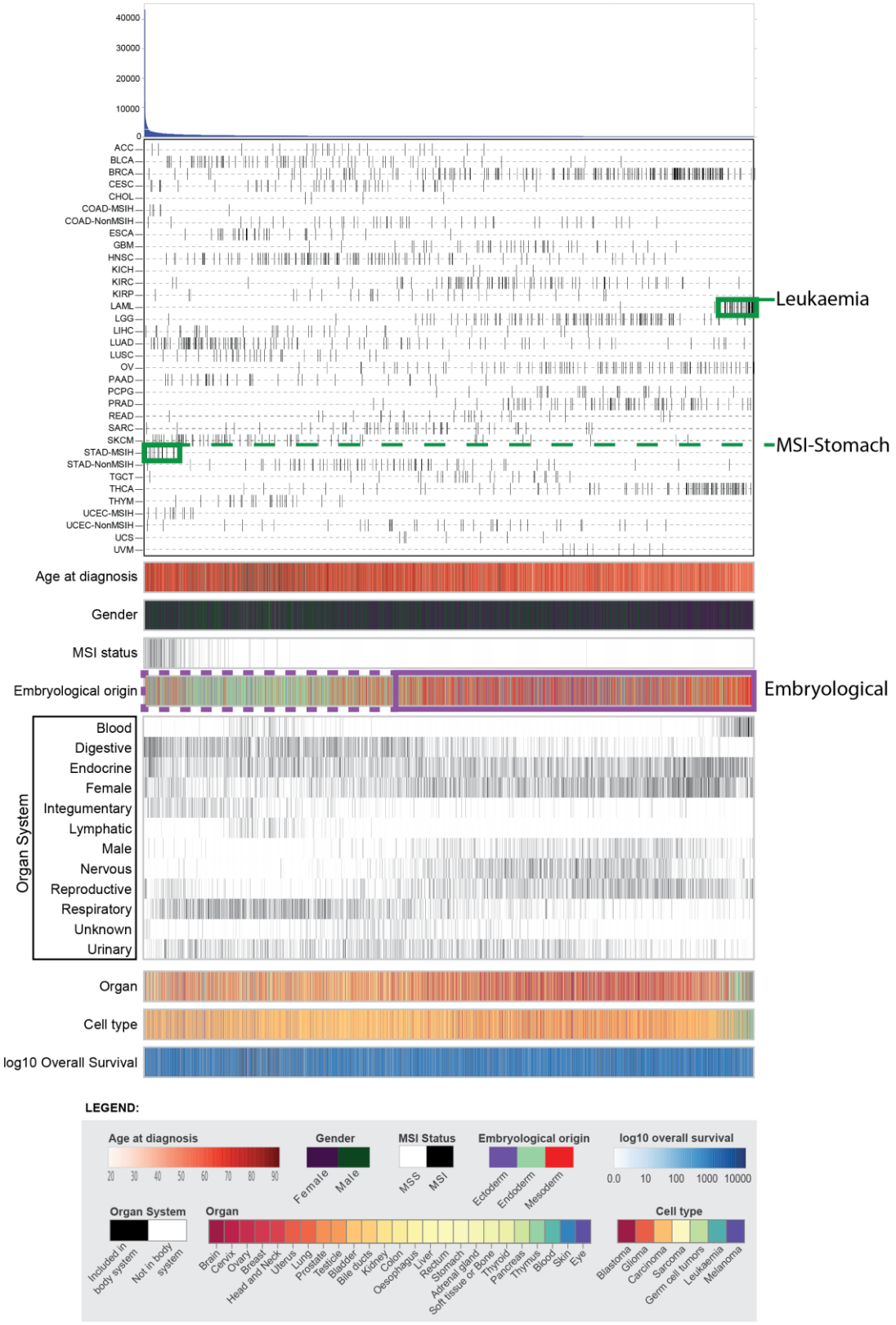

Figure 41: Clustering of all dimensions with proportions in all cases by using average metric and city block linkage.

Clustering has been performed on all cases in the TCGA according to a combination of all dimensions of data using proportions rather than counts. The x-axis dendrogram shows clustering of the data by case.

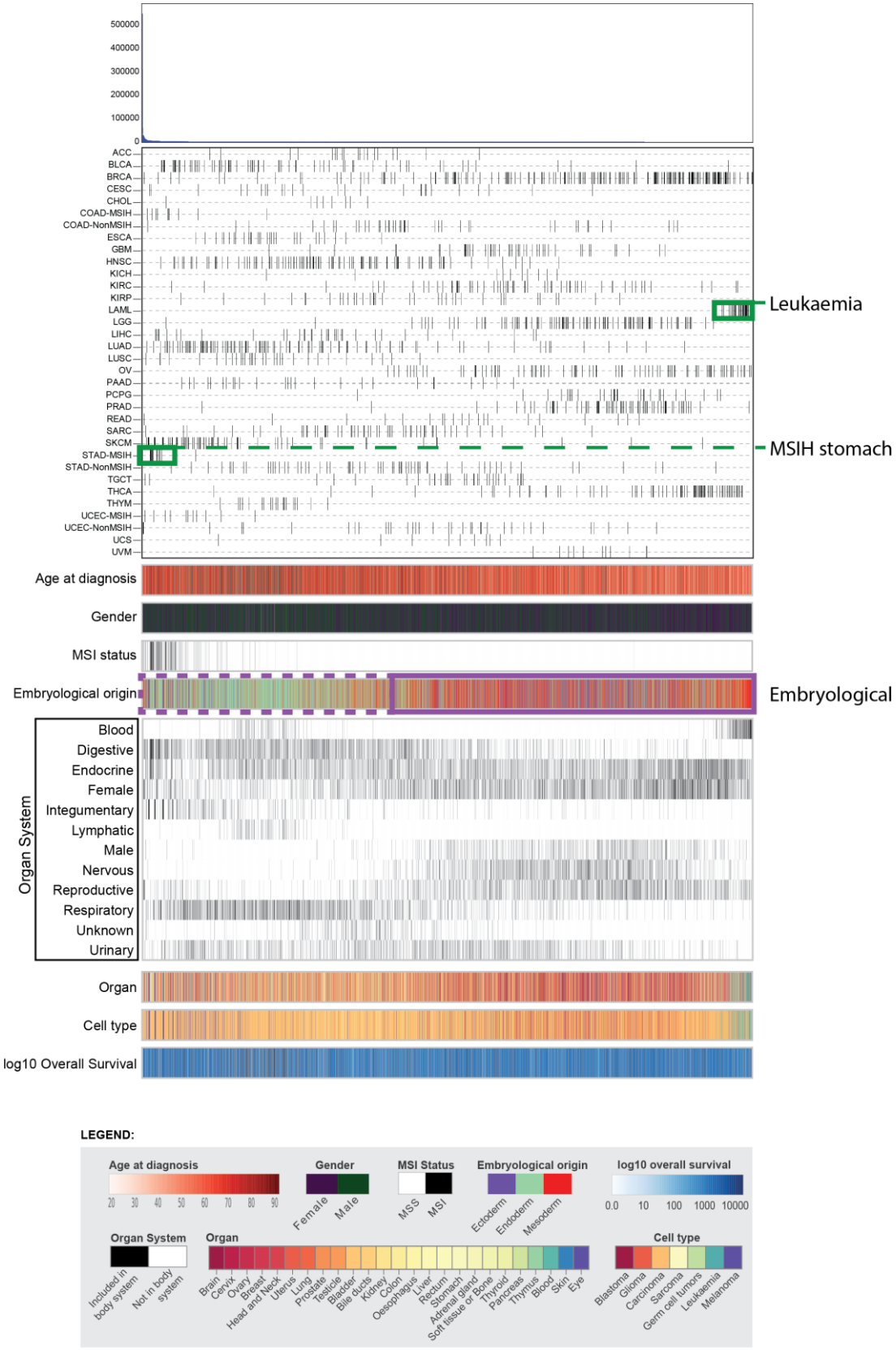

Figure 42: Clustering of all dimensions with counts in all cases by using average metric and city block linkage.

Clustering has been performed on all cases in the TCGA according to a combination of all dimensions of data using counts rather than proportions. The x-axis dendrogram shows clustering of the data by case.

Table 17: Incidences of distinct cancer mutational profiles in all 10 dimensions of analysis

Cancer code	Cancer	Trinucleotide-Pro	Trinucleotide-Counts	Indel-Pro	Indel-Counts	Recurrent variants	Genes	Distribution-Pro	Distribution-Counts	Multidimensional-Pro	Multidimensional-Counts
ACC	Adrenocortical carcinoma	-	-	-	-	UC	-	-	-	-	-
BLCA	Bladder urothelial carcinoma	SCS	-	-	-	-	-	-	-	-	-
BRCA	Breast invasive carcinoma	-	-	-	-	US	US	-	-	-	-
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	-	-	-	-	-	-	-	-	-	-
CHOL	Cholangiocarcinoma	-	-	-	-	-	-	-	-	-	-
COAD-MSIH	Colon adenocarcinoma (MSI)	-	-	-	-	-	-	SCS	-	-	-
COAD-NonMSIH	Colon adenocarcinoma	-	-	-	-	-	-	-	-	-	-
ESCA	Esophageal carcinoma	-	-	-	-	-	-	-	-	-	-
GBM	Glioblastoma multiforme	-	-	-	-	-	-	-	-	-	-
HNSC	Head and neck squamous cell carcinoma	-	-	-	-	-	-	-	-	-	-
KICH	Kidney chromophobe	-	-	-	-	-	-	-	-	-	-
KIRC	Kidney renal clear cell carcinoma	-	-	-	-	-	US	-	-	-	-
KIRP	Kidney renal papillary cell carcinoma	-	-	-	-	-	-	-	-	-	-
LAML	Acute myeloid leukemia	-	SCS	-	-	UC	US	UC	SCS	SCS	SCS
LGG	Brain lower grade glioma	-	-	-	-	US	US	-	-	-	-
LIHC	Liver hepatocellular carcinoma	SCS	-	-	-	-	-	-	-	-	-
LUAD	Lung adenocarcinoma	UC	-	-	-	-	-	-	-	-	-
LUSC	Lung squamous cell carcinoma	-	-	-	-	-	-	-	-	-	-
OV	Ovarian serous cystadenocarcinoma	-	-	-	-	-	-	-	-	-	-
PAAD	Pancreatic adenocarcinoma	-	-	US	US	UC	UC	US	-	-	-
PCPG	Pheochromocytoma and paraganglioma	-	-	-	-	US	-	-	-	-	-
PRAD	Prostate adenocarcinoma	-	-	-	-	-	-	-	-	-	-
READ	Rectum adenocarcinoma	-	-	-	-	-	-	-	-	-	-
SARC	Sarcoma	-	-	-	-	-	-	-	-	-	-
SKCM	Skin cutaneous melanoma	UC	US	-	-	-	US	SCS	-	-	-
STAD-MSIH	Stomach adenocarcinoma (MSI)	-	-	-	-	-	-	-	US	SCS	SCS
STAD-NonMSIH	Stomach adenocarcinoma	-	-	-	-	-	-	-	-	-	-
TGCT	Testicular germ cell tumours	UC	UC	-	-	-	-	-	-	-	-
THCA	Thyroid carcinoma	-	-	US	US	US	US	-	-	-	-
THYM	Thymoma	UC	UC	-	-	-	-	-	SCS	-	-
UCEC-MSIH	Uterine corpus endometrial carcinoma (MSI)	-	-	-	-	-	SCS	-	SCS	-	-
UCEC-NonMSIH	Uterine corpus endometrial carcinoma	-	-	-	-	-	-	-	-	-	-
UCS	Uterine carcinosarcoma	-	-	-	-	-	-	-	-	-	-
UVM	Uveal melanoma	-	-	-	-	-	-	-	-	-	-

UC	Unique cancer signature	SCS	Shared Cancer signature
US	Unique subtype signature	-	Non-definable type

2.4. Discussion

2.4.1. Choice of reference data

The Cancer Genome Atlas (TCGA) is a US multicentre effort to catalogue pan-cancer aberrations that was started in 2005 (The Cancer Genome Atlas 2013). This database catalogues Clinical, WES, WGS, whole transcription sequencing, miRNA-seq, methylation and CNV cancer aberrations with varying representation of this data types in the different cancers, however, the most consistent data type is WES data. Several studies have been created from this huge research effort, several of which are considered be benchmark papers in various aspects of cancer studies (http://www.nature.com/tcga/). The stated aim of the project is to provide data and genome analysis to the cancer research community with the intent that this data will support new discoveries and accelerate the pace of cancer research. The cancers types that are decided for inclusion are based on either having poor prognosis or a significant overall public health impact and also when there is the availability of human tumour and matched normal tissue of high quality. From the perspective of a researcher, the three main attractive features of the TCGA dataset are 1) it is a huge source of results from many cancers, 2) it is multidimensional in its data analysis including DNA, RNA and methylation results and 3) the data is well curated and highly standardized across all datasets. Considering all these features, this dataset was the perfect choice to study cancer subtypes as part of the work presented in this thesis as it represents all the major subtypes that a researcher would generally encounter in cancer studies.

The initial challenge of dealing with the TCGA data is that it that for any given cancer, the WES results are supplied as several MAF files of different versions of the analysis with overlapping cases. As there was no definitive version that could be

considered superior, all data was combined taking consideration to avoid multiple copies of any one case. Where any case was in more than one version, the variants were combined without duplication, the downside of this being the bias towards a higher overall mutation load.

2.4.2. Microsatellite instability consideration

Microsatellite instability (MSI) is a hypermutable phenotype caused by the loss of DNA mismatch repair activity primarily due to mutations in mismatch repair genes (Boland and Goel 2010). Cells with abnormally functioning MMR are unable to correct errors that occur during DNA replication and consequently accumulate errors. New alleles in the abnormal sample not found in the corresponding normal sample indicate the presence of MSI. Several PCR-based methods exist for the determination of MSI status including the Promega MSI kit (Promega Corporation,) and the Type-it Microsatellite PCR Kit (Qiagen). As the underlying mechanisms of MSI, unlike other cancers, may be mismatch repair gene, we postulated that MSI-high cases for a given cancer subtype should be considered a different subset of cancer and as such was separated prior to further analysis. Specifically, colon adenocarcinoma (COAD), stomach adenocarcinoma (STAD) and uterine corpus endometrial carcinoma (UCEC) were segregated according to MSI status (MSI high and not-MSI high) based on the TCGA annotation. MSI annotation was also available for rectum adenocarcinoma (READ), esophageal carcinoma (ESCA) and uterine carcinosarcoma (UCS) however there were too few MSI-high cases to be statistically meaningful. As observed, the MSI cases consistently segregated from the non-MSI cases of the same cancers and in fact, the MSI cases from all three cancers tended to cluster together, a phenomenon observed in all dimensions of the consensus analysis (section 2.3.3) and half of the individual

cases analyses (section 2.3.4). As revealed by the indel analysis, these cancers were identifiable by having higher rates of 1 base insertion and deletions, huge mutational loads that result in many more variants and consequently mutated genes, and a much higher genomic mutational distribution rate. It is, therefore, necessary to segregate MSI cases from all other cases when attempting to understanding the underlying mutational phenomenon, even though this has not been done in previous studies utilising the TCGA dataset.

2.4.3. Mutational signatures are distinct in certain cancer subsets

Perhaps the first publication to highlight that, overall, cancers have specific patterns of SNVs was by Michael Stratton's group in 2007 (Greenman et al. 2007). In this work, the authors studied the sequencing results from the putative 518 genes that form the kinome, in eight different cancer types. The kinome is made of proteins kinases, meaning that they catalyse phosphorylation reactions and are major control points in cellular behaviour (Manning et al. 2002) and are major drug targets. The specific reason for selecting these gene are twofold. Firstly, this research, although recent, pooled data from research projects performed before the mass expansion of NGS and therefore tended to be of a much smaller scale than current WES or WGS datasets and thus necessitated a more concentrated approach to DNA sequencing. Secondly, these genes were specifically chosen for study since they are the most likely set of genes to be biologically significant and therefore under selection pressure. The main weakness of these datasets is that it uses pooled data from several experiments, and no one sample was sufficient to create a mutation signature due to the relatively low number of mutations from a kinome sequencing experiment. That notwithstanding, this was the first time such an observation had been made at such a large level and was the first piece of evidence that site of origin prediction was theoretically possible. The most

recent example of a large-scale study of mutation profiles was published 6 years later and leveraged on the rich source of WES results from the TCGA database. This study introduced the notion of trinucleotide mutations and the heterogeneity of the cancers by plotting the non-negative matrix factorization (NMF) of the mutation as a radial plot, the authors, however, did not go into any detail concerning this phenomenon as the publication's main focus was the development of MutSig, the mutation ranking software package mentioned previously in this text.

In section 2.3.3 of this thesis, consensus representations of each cancer type were used to determine if there are patterns of relatedness between the 34 cancer types in this study, done by clustering the mutation results in any of 8 different methods of analysing the mutations, referred to as the 'dimensions'. In addition, the aggregation of the dimensions was also done, i.e. the all dimensions proportions approach that utilised proportional data and the all dimensions counts preferring counts when applicable to the different dimension, i.e. 10 dimensions altogether. The clustering was compared to specific cancer characteristics to study possible associations of these characteristics to the clustering patterns. Several of cancers were revealed to have similar mutation profiles, dependent on the mutation dimension being studied. It should be noted that the different cancer are represented by datasets that were derived at different times, in many situations, from different sequencing centres and analysed with different bioinformatics pipelines. Despite this differences, several related cancer were consistently similar to each other. The two lung cancers, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), were paired together in the clustering in the trinucleotide, mutated genes and genomic distribution dimensions. The three kidney cancers (kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP)) and the two non-MSI colorectal cancers

(colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ)) clustered by the trinucleotide and recurrent variants dimensions. The non-lung squamous cell cancers (cervical squamous cell carcinoma (CESC) and head and neck squamous cell carcinoma (HNSC)) also clustered by the trinucleotide profiles. The seemingly unrelated cancers acute myeloid leukaemia (LAML), thyroid carcinoma (THCA), ovarian serous cystadenocarcinoma (OV) were consistently cluster by the indel and mutated genes dimensions. The three MSI cancers clustered in all of the ten dimensions, suggesting how the MSI phenomenon really is a phenotype in itself. These consistent associations between different cancers suggest that there are truly underlying patterns that can identify at least some cancer types.

In section 2.3.4, analysis was performed on the 8820 individual cases to identify if there are cancer specific or cancer sub-group specific mutational signatures according to the 10 dimensions. The clustering pattern of each cancer was divided into four types as described in section 2.3.4, with the intention of identifying cancers with varying degrees of exclusivity in mutational signature patterns. UC clustering suggests that all cases of a cancer type have a unique mutational signature as compared to the cases in all other cancers. Theoretically, these cancers types could be identified, within limits, based on the appropriate mutational signature alone. Clustering is termed US when only a subset of the cases within a cancers type exhibit a unique mutational signature. As such, a given case from a cancer with US clustering may be definitively identified if its signature falls within the unique subset, but not if its signature deviates from this. SCS clustering indicates that all the cases from a given cancers fall within a specific mutational signature, indicated by a narrow range of clustering, however, the signature is not unique to that cancer, i.e. cases from at least one cancer shares this profile. Despite having a specific signature, these cancers cannot be definitively identified due to the

overlapping signature patterns with other cancer(s). Lastly, NDT clustering indicates cancers where there is great heterogeneity in the clustering, and thus mutational signatures and based on simple hierarchical clustering alone, it would be impossible to identify the cancer of any case in these cancers. Table 17 summarises clustering profiles seen in the 34 cancer subtypes used in this study. Each dimension of analysis is presented as a column and cancer types along the rows. The annotations in the cells indicate the cluster type that was observed in each cancer with each annotation dimension. Seventeen of the cancers studied showed NDT clustering in all analysis, suggesting that predicting the cancers type/site of origin from these cancers would be exceedingly challenging. Seven cancers showed UC clustering in at least one dimension of analysis, seven cancers without UC clustering showed US clustering in at least one dimension. Four cancers without UC or SCS clustering showed SCS clustering clustering in at least one dimension. The UC and US cancers put together suggest that at least 14 of the 34 cancers have cases that may be identified by the mutational signatures alone. Acute myeloid leukaemia was the most identifiable cancer, with UC clustering in genomic density proportions and the recurrent variants dimensions, also with US clustering with the analysis of mutated genes. Pancreatic adenocarcinoma (PAAD), thymoma and testicular germ cell tumours were also highly identifiable, have UC clusters in two dimensions each. Overall, the trinucleotide proportions seem to be the best classifier for specific cancer signatures, with four cancers having UC clusters. The variants analysis (three cancers with UC, and four with US clusters) and the mutated genes analysis (one cancer with a UC cluster (PAAD) and six cancers with US clusters) were also particularly informative.

2.4.4. Computational considerations

The python programming language was chosen for this analysis due to its power and flexibility, specifically its strength with its array data structure via the use of numpy and pandas. Much thought was put into the computational framework for this project, due to the very large dataset being used. A comparison was performed against R 3.1.0, which has a very similar array interface to pandas, for loading of the large dataset which had over 584 x 10^6 elements, and data manipulation and python appeared to be significantly quicker and more reliable (data not shown). Added to this, python added the functionality of several powerful libraries, such as the scikit-learn used in chapter 3. The analysis presented here must be performed on a high-performance machine due to the large computational and memory requirements, specifically a server with four 64-bit 15-core Intel Xeon processors (3.2 GHz each) and 512 GB of RAM.

2.4.5. Limitations of studying small mutations with hierarchical clustering

Perhaps the most obvious shortcoming of the analysis done so far is that only the small mutations have been studied, albeit in great detail, specifically data generated from SNV and indels using a somatic variant caller. Potentially a more detail analysis of cancer subtypes may have been possible with the inclusion of the other aspect of the TCGA data, inclusive of copy number variation data, RNA expression, miRNA, methylation and others. Avoiding these analyses was, however, a conscious decision, as the intention was to carry out a summary that would be part of a framework to predict cancer site of origin from only WES or WGS results alone, as this type of analysis represents the most common analysis approach used in NGS and high throughput analyses.

The results in this chapter show there are certainly specific homogenous sets of cases with cancers types that share mutational patterns and are easily distinguished from other cancers (Table 17). This homogeneity is not universal, with 20 of the 34 cancer subtypes in this study not having unique clustering patterns. Based on these observations, is unlikely that all cancer subtypes have unique signatures that are homogenous throughout all cases and at the same time easily distinguished from all other cancer subtypes. There may, however, be some distinct patterns that distinguish subsets of the different cancers subtypes consistent with the work presented in this thesis. However, if possible, it will take a sophisticated methodology than presented in this chapter to be able to identify more complex or hidden relationships that are specific to the cancer types and as such may allow identification of specific cancers. The deep analysis of the mutations data is investigated in chapter 3 via the interrogation of multiple statistical and machine learning methodologies. This is performed with the intention of establishing a framework to use all dimensions of the data to help identify features which are unique to cancer subtypes and also allow the prediction of a cancer subtype or site of origin based on the revealed features.

Chapter III

Statistical and Machine learning prediction of cancer type

3.

3.1. Introduction

The work shown in chapter 2 has indicated that 17 of the 34 cancers in this study have mutation profiles which are both unique and consistent in at least one mutational dimension, throughout either all cases, termed the US cancers, or a subset of cases, termed the UC cancers. This is encouraging, as it suggests that SNVs and indels alone can provide sufficient information to determine cancer type from whole exome sequencing of these cancers. There are however 17 cancers for which no identifiable profile has been determined. Realising that cancer-specific signatures exist in certain cancers, the work in this chapter investigates whether more complex machine learning algorithms can identify mutational pattern specific to a greater number of cancer types or a subset of cancer types using the dimensions of mutations. These identified signatures may then be used to identify the origin of unknown cancers, with applications in CTCs and CUPs diagnosis.

3.2. Methods

3.2.1. Machine learning approaches

An integrated framework was developed that accepts the data matrix developed in chapter 2 (section 2.3.2, Table 5) and then applies statistical and machine learning (ML) algorithms to elucidate cancer subtype-specific features from the multidimensional data matrix. In this chapter, only 6 dimensions of the mutations (2.2.2) were used to investigate cancer specific signatures, as follows:

1. Trinucleotide mutations (proportions)
2. Indel mutations (proportions)
3. Mutated genes
4. Recurrent variants
5. Genomic distribution (proportions)
6. Multidimensional (proportions)

Specifically, where relevant, proportional representations of the data were used rather than the counts. The decision to not include counts was done as work in in chapter 2 showed that, overall, proportions were superior to counts in deciphering cancer specific profiles. This dataset that represents all 8820 cases derived from the TCGA project is termed the "learning" dataset.

Machine learning was performed using scikit-learn, a Python module that integrates a wide range of state-of-the-art machine learning algorithms for large-scale supervised and unsupervised problems (Pedregosa 2011). This package was selected for its robustness, excellent documentation and wide usage in the ML community. Its close integration with Numpy and Pandas also allowed rapid handling of the huge dataset. Figure 43 represents all the algorithms used in this section, it should be noted however that several of these individual algorithms were used in combination. The basis for the selections of these different algorithms was to cater to several possible data distribution patterns that could explain the underlying differences between the cancer subtypes. As seen in Figure 43, there are two broad classes of algorithms, supervised prediction techniques which are the algorithms that are actually used for the prediction of the subtype of a cancer, and dimensionality reduction algorithms, used in the process of reducing the number of random variables under consideration, which in this study, is used to identify only important factors of the datasets prior to prediction, a step that could potentially result in much more efficiently derived and meaningful predictions. The algorithms are also divided into decision tree methods, and those optimised for categorical data, respectively optimised for the trinucleotide, indel and genomic distribution data which are continuous and the mutated genes and recurrent variants data which are categorical (binary).

In terms of classifiers, the decision tree methods use branching model of decisions and possible consequences to predict the chance of outcomes. These methods are optimised for dependent variables and are computationally intensive. The methods used in this study are specifically ensemble decision tree approaches i.e. Random Forest, Bagging, Gradient

Boosting and Adaptive Boosting (AdaBoost). Ensemble methods average the results from of several decision trees, with the goal of improving prediction accuracy by reducing variance and avoidance of overfitting, and differ in the approach to averaging. Three naive Bayes classifiers were used, a class of classifiers based on Bayes' theorem using frequency tables to determine likelihood, specifically with the assumption of independence among features. Although considered less comprehensive than decision tree methods, these techniques have been shown to outperform more sophisticated methods, despite being computational efficient. The canonical naive Bayes algorithm is considered the most comprehensive approach, while the Gaussian model is suited for continuous data and the Bernoulli model optimised for binary variables. Support Vector Machines (SVM) were implemented as a linear classification, as well polynomial kernels with two separate degrees specified (3 and 10). As highlighted in Figure 43, SVM approaches do not provide probability estimates, however, probabilities can be calculated indirectly by using computationally expensive cross-validation. It should be noted that although the decision trees and SVM are often considered optimised for continuous and categorical data respectively both algorithms are actually applicable to both data types.

Dimensionality reduction was performed by principal components analysis (PCA), variance threshold (VT) and a Restricted Boltzmann machine (RBM). The PCA works by orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components, the variance threshold selects only features with a specified variance threshold, set at 0.2 in this study and the RBM is a generative stochastic artificial neural network that can learn a probability distribution over its set of inputs.

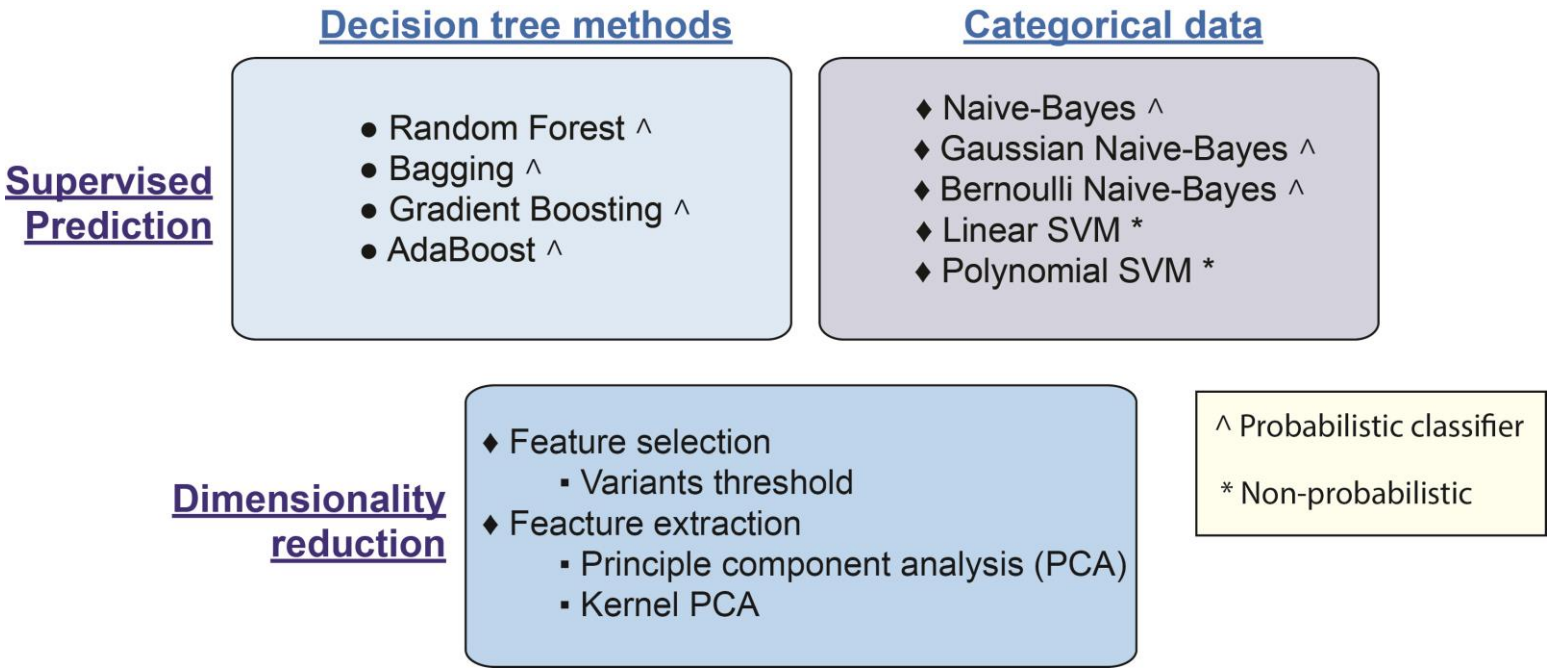

Figure 43: Machine learning algorithms used for mutational signature learning and cancer subtype prediction

3.2.2. Datasets used for determining prediction accuracy

Three datasets were used to test prediction accuracy. Firstly, a random sampling of 13 cases from each TCGA subtype, making up 442 cases (5% of the TCGA dataset), was used as the initial "TCGA test dataset" and for algorithm optimisation. This was done in three phases. Phase I involved the initial determination of the most accurate algorithm and dataset for cancer type prediction. This was followed by phase II, where optimisation using dimensionality reduction was investigated. Phase III involved a combinatorial approach, where the most accurate algorithms in phase I were combined with two mutation data dimension to achieve even greater accuracy.

The second dataset used was DNA mutations available via the ICGC database (International Cancer Genome Consortium et al. 2010). Only cancers with a similar cancer type in the TCGA dataset were chosen, resulting in 57 different cancer types from 54 studies derived from 14 countries including a total of 9155 individual cases (Table 18). The three most accurate predictors from phase III were analysed by an Area under the Receiver Operating Characteristic curve (AUROC), as this allows the selection of the best predictor while maintaining the potential for sensitivity and specificity modulation by threshold selection.

Lastly, the most optimised algorithm was applied to a CTC WES dataset (Ni et al. 2013). These data were derived from a study that sequenced CTCs, primary tumours and metastases in patients with lung adenocarcinomas and was used to truly determine if prediction can be made from CTCs.

Table 18: ICGC datasets used in Phase III of prediction optimisation

Cancer	Country of origin	TCGA equivalent cancer	Total cases
Acute lymphoblastic leukemia	US	LAML	27
Acute myeloid leukemia	KR,US	LAML	268
Benign liver tumour	FR	LIHC	30
Biliary tract cancer	JP	CHOL	239
Bladder cancer	CN	BLCA	103
Bladder urothelial cancer	US	BLCA	130
Bone cancer	UK	LAML	66
Brain glioblastoma multiforme	US	GBM	268
Brain lower grade glioma	US	LGG	283
Breast cancer	US	BRCA	955
Breast triple negative/lobular cancer	UK	BRCA	117
Cervical squamous cell carcinoma	US	CESC	194
Chronic lymphocytic leukemia	ES	LAML	218
Chronic myeloid disorders	UK	LAML	129
Colon adenocarcinoma	US	COAD	216
Colorectal cancer	CN	COAD	147
Early onset prostate cancer	DE	PRAD	11
Esophageal adenocarcinoma	UK	ESCA	119
Esophageal cancer	CN	ESCA	228
Gastric adenocarcinoma	US	STAD	289
Gastric cancer	CN	STAD	9
Head and neck thyroid carcinoma	US	HNSC	400
Kidney renal clear cell carcinoma	US	KIRC	404
Kidney renal papillary cell carcinoma	US	KIRP	159
Liver cancer	FR,JP	LIHC	748
Liver hepatocellular carcinoma	US	LIHC	188
Lung cancer	CN,KR	LUAD, LUSC	46
Lung squamous cell carcinoma	US	LUSC	178
Malignant lymphoma	DE	LAML	44
Neuroblastoma	US	GBM	108
Oral cancer	IN	HNSC	106
Ovarian cancer	AU	OV	93
Ovarian serous cystadenocarcinoma	US	OV	88
Pancreatic cancer	AU,CA,IT	PAAD	633
Pancreatic cancer endocrine neoplasms	AU	PAAD	52
Pediatric brain cancer	DE	LGG	246
Prostate adenocarcinoma	CA,US,UK	PRAD	488
Rectum adenocarcinoma	US	READ	80
Renal cancer	CN	READ	10
Renal cell cancer	FR	KIRC,KIRP	95
Skin adenocarcinoma	BR	SKCM	66
Skin cancer	AU	SKCM	183
Skin cutaneous melanoma	US	SKCM	335
Soft tissue cancer	FR	SARC	98
Thyroid cancer	SA	THCA	15
Uterine corpus endometrial carcinoma	US	UCEC	246

ICGC cancers with corresponding TCGA cancer types were included in this study. There are a total of 9155 cases as part of this dataset from 14 countries. AU: Australia, BR: Brazil, CA: Canada, CN: China, DE: Germany, ES: Spain, FR: France, IN: India, IT: Italy, JP: Japan, KR: Korea, SA: Saudi Arabia, UK: United Kingdom, US: United States

3.2.3. MutProfiler: Site of origin prediction web tool

The algorithm that most accurately predicted the tumour site of origin was included in MutProfiler, a web-based tool to aid researchers with summarising mutational signature data from their NGS results and predicting cancer type and/or subtype. MutProfiler is implemented in the Django web framework v1.8.0 on an Nginx web server via the Web Server Gateway Interface (WSGI) and currently hosted at http://172.16.203.178/mutprofiler/. The website is a pure python implementation using the Python programming language v3.4.1 with py-postgresql v1.1.0, Pandas v0.16.2, Numpy v1.9.2, Scikit-learn 0.16.1 and Matplotlib v1.4.3. It is currently hosted on a server running Ubuntu 14.04 with four 64-bit 15-core Intel Xeon processors (3.2 GHz each) and 512 GB of RAM. The current IP is non-static and therefore the IP address may change, but the website will be hosted at a permanent address within a 2-month period.

3.3. Results

Selection of the most accurate cancer type prediction algorithm was performed in four phases as follows:

In the phase I, the TCGA test dataset (442 randomly sampled cases) was used. The 9 supervised machine learning algorithms shown in Figure 43 were applied to the learning dataset (6 dimensions of data) to train the algorithm for cancer type prediction, resulting in 54 ML algorithm-dataset pairings. Each pairing was applied 442 times, i.e. once for each of the cases in the TCGA test dataset, with that specific case excluded from the learning dataset to be used only for prediction. This approach tests for the robustness of the ML algorithm-dataset pairing by avoiding correct prediction due to fitting to specific cases.

In phase II, an attempt was made to improve predictive outcomes and computational efficiency by applying dimensionality reduction procedures to the four most accurate pairings representing the multidimensional analysis, recurrent variants, trinucleotides mutations and mutated genes. Specifically, the RBM, PCA and VT dimensionality reduction methods were applied to the following pairings:

1. Bagging classifier + multidimensional data
2. Bagging classifier + mutated genes
3. Gaussian Naïve Bayes + recurrent variants
4. Gaussian Naïve Bayes + trinucleotide mutations
5. Random forest classifier + trinucleotide mutations

In phase III a combinatorial approach was made to improve prediction accuracy by combining the most accurate ML prediction algorithms with the two mutation datasets that generated the most accurate predictions from Phase I with said algorithm. This procedure revealed vastly improved prediction accuracies.

In phase IV, the three most accurate algorithms were tested against non-TCGA datasets. Firstly, the three most accurate algorithms were compared against all studies available via the ICGC database, from which the most accurate ML learning and dataset combination was determined. This was followed by utilization of this algorithm for the prediction of the cancer type from whole exome sequenced CTC results.

3.3.1. Phase I: Prediction of cancer type

Table 19 shows 11 of the most accurate prediction pairings overall out of the 54 combinations. There prediction accuracies within each cancer are also indicated. Overall, the indel proportions proved to be the worst prediction dataset of the six dimensions. The most accurate predictor with this dimension was the random forest classifier with 20% prediction accuracy, and the poorest predictor with this dimension was the linear SVM (9%). Overall, more accurate predictions were observed with the other dimension. Along with performance in indels, the linear SVM was also the worst predictor in 2 other dimensions, genomic distribution and trinucleotide distributions. This indicates that these datasets are highly complex and cannot be divided based on linear hyperplanes which work best with data which is linearly separable.

Ada Boosting was the worst predictor for two of the dimensions, which is unexpected as the other decision tree ensemble methods performed well. It is possible that the Ada Boosting could be refined to improve prediction accuracies, however this was not investigated further as prediction accuracies with the other ML algorithms were concentrated on instead.

Gaussian naive Bayes (GNB) was the most accurate algorithm when applied to the trinucleotide distributions and recurrent variants. These two datasets differ significantly in both size, where there are 96 trinucleotide categories and 42,030 variant categories,

and in data type, where the trinucleotide mutation is proportional data, ranging from 0 to 1, while the variant data is binary. The success with GNB suggest that the features in these datasets are either independent, or that dependant values largely co-occur, such that probabilities from frequency tables are adequate for prediction. The bagging classifier generated the most accurate predictions with the mutated genes and the multidimensional data, with 41% and 52% accuracies respectively.

The 11 most accurate pairings as shown in Table 19 are represented by all dimensions except indels, and by the bagging classifier, GNB, random forest classifier, gradient boosting classifier and the Bernoulli naive Bayes classifier. Although the prediction accuracies are highly encouraging, with the most accurate prediction at 52% accuracy, more accurate prediction would be desirable, especially for diagnostic applications, where the results could influence treatment strategy and affect patient survival and quality of life. The following sections are therefore attempts to refine prediction accuracy.

Table 19: Phase I - The ten most accurate prediction ML and dataset pairings

L algorithm	Bagging	Gnb	Gnb	Bagging	RFC	GBEst	RFC	GBEst	RFC	NBB	Bagging
Mutational dimension	multidimensional	recur vars	trinucleotides	genes	trinucleotides	distribution	recur vars	trinucleotides	genes	distribution	distribution
Mean	52	50	49	41	41	41	39	36	36	32	34
Adrenocortical carcinoma (ACC)	54	92	85	38	46	77	100	38	54	54	54
Bladder Urothelial Carcinoma (BLCA)	69	38	62	31	62	23	0	69	23	31	23
Breast invasive carcinoma (BRCA)	62	46	31	69	77	69	77	85	85	38	54
Cervical squamous cell carcinoma (CESC)	46	31	38	15	15	8	8	23	0	8	8
Cholangiocarcinoma (CHOL)	0	0	54	0	0	0	0	8	0	0	0
Colon adenocarcinoma (COAD)	69	38	77	65	54	46	38	35	46	0	54
Esophageal carcinoma (ESCA)	38	46	54	15	54	23	23	15	0	54	31
Glioblastoma multiforme (GBM)	77	69	62	31	62	23	54	38	8	69	38
Head and Neck squamous cell carcinoma (HNSC)	54	54	23	46	38	46	15	38	46	62	38
Kidney Chromophobe (KICH)	23	0	69	8	0	15	0	46	0	8	0
Kidney renal clear cell carcinoma (KIRC)	62	31	54	69	62	46	31	15	69	54	38
Kidney renal papillary cell carcinoma (KIRP)	38	46	38	23	15	15	31	15	0	15	0
Acute Myeloid Leukemia (LAML)	50	83	42	58	67	42	42	17	75	100	75
Brain Lower Grade Glioma (LGG)	62	85	31	62	54	69	62	46	77	77	69
Liver hepatocellular carcinoma (LIHC)	38	23	54	8	46	23	23	38	0	8	0
Lung adenocarcinoma (LUAD)	54	31	23	54	69	54	38	46	69	8	69
Lung squamous cell carcinoma (LUSC)	38	15	92	8	31	38	8	15	0	31	31
Ovarian serous cystadenocarcinoma (OV)	69	38	23	77	15	62	92	31	69	31	46
Pancreatic adenocarcinoma (PAAD)	69	100	54	46	23	46	85	31	85	69	38
Pheochromocytoma and Paraganglioma (PCPG)	46	92	54	62	0	23	38	15	69	62	46
Prostate adenocarcinoma (PRAD)	38	15	46	46	8	15	23	15	15	77	31
Rectum adenocarcinoma (READ)	0	17	25	0	0	17	8	33	0	8	0
Sarcoma (SARC)	54	77	46	23	46	23	23	54	0	46	8
Skin Cutaneous Melanoma (SKCM)	85	69	85	77	85	62	15	77	69	46	62
Stomach adenocarcinoma (STAD)	58	54	27	12	65	38	58	38	19	0	38
Testicular Germ Cell Tumors (TGCT)	46	69	85	8	85	31	8	62	0	15	0
Thyroid carcinoma (THCA)	77	69	23	85	23	62	92	23	62	38	54
Thymoma (THYM)	62	69	69	54	92	38	54	46	23	38	15
Uterine Corpus Endometrial Carcinoma (UCEC)	54	58	19	54	42	73	42	35	46	15	38
Uterine Carcinosarcoma (UCS)	23	31	23	8	0	38	0	15	0	0	8
Uveal Melanoma (UVM)	54	69	69	100	0	85	85	38	92	8	62

The numbers and coloured intensities represent the percentage of cases in each cancer type that were correctly predicted. The columns are arranged in descending order of mean prediction accuracy. **ML algorithms** - Bagging: Bagging decision tree classifier, Gnb: Gaussian Naive Bayes, RFC: Random forest classifier, GBEst: Gradient Boosting.

Mutational dimension - recur vars: Recurrent variants, genes: Mutated genes, distribution: Genomic distribution, trinucleotide: Trinucleotide mutations, multidimensional: Multidimensional mutational data.

3.3.2. Phase II: Dimensionality reduction reduced prediction accuracy

Table 20 shows the prediction accuracies after application of the three dimensionality reduction techniques to the most accurate predictors of phase I. The table is arranged in the same order as Table 19, i.e. according to the mean accuracy of the ML and mutation dimension pairing. Within each pairing, the original prediction accuracy (NDR) is shown along with the restricted Boltzmann machine (RBM), principal components analysis (PCA) and variance threshold (VT) transformed data. As seen, the dimensionality reduction (DR) procedures overall caused a reduction in prediction accuracy. However, in very few instances the DR procedure did improve accuracy as seen in the coloured cells of Table 20. Overall the least detrimental reduction was PCA, followed by RBM and the most detrimental was VT. PCA was the only DR that significantly improved prediction accuracy in several cancers.

PCA with the bagging classifier and multidimensional data pairing, generated an improvement in four cancer types, with improvements over the NDR ranging from 0.1 to 8.9 percentage points (pp), in testicular germ cell tumours (TGCT) and thyroid carcinoma (THCA) respectively. With the Gaussian Naive Bayes and recurrent variants pairing, prediction in 3 cancers were improved, with improvements ranging from 0.18 pp in sarcoma (SARC) to 10 pp in brain lower grade glioma (LGG). 6 of the cancers types had better predictions in the Gaussian Naive Bayes and trinucleotide mutation pairing, with the greatest improvement in testicular germ cell tumours at 6 pp improvement. Prediction in 7 of the cancers was improved in the bagging classifier and mutated genes pairing, significantly in kidney renal clear cell carcinoma (KIRC), colon adenocarcinoma (COAD) and pancreatic adenocarcinoma (PAAD), which improved by 13.4, 8.9 and 6.6 pp respectively.

The overall ineffectiveness of the DR procedures suggests that, generally, it is subtle differences between the cancers types that are the distinguishing features rather than large differences, and that these subtle differences are lost when dimensionality reduction is performed except in a few cancers types. Therefore, to create a single prediction algorithm for cancer type prediction it is necessary to not to lose dimensionality so as to cater to all possible cancers type that may present as a CTC sample or a CUP, where overall accuracy would be preferential.

Table 20: Phase II - Effect of dimensionality reduction of prediction accuracy

	Bagging				Gnb								Bagging			
	multidimensional				recur vars				trinucleotides				genes			
	NDR	RBM	PCA	VT	NDR	RBM	PCA	VT	NDR	RBM	PCA	VT	NDR	RBM	PCA	VT
Mean	51	26	37	11	50	25	34	14	50	21	39	16	40	22	28	7
ACC	54	20	45	18	92	41	37	43	85	33	0	51	38	20	39	12
BLCA	69	24	33	19	38	5	23	30	62	22	53	0	31	14	0	4
BRCA	62	22	66	0	46	16	0	0	31	7	16	0	69	32	40	26
CESC	46	34	37	0	31	29	17	0	38	13	36	11	15	7	13	7
CHOL	0	0	0	0	0	0	0	0	54	20	43	16	0	0	0	0
COAD	69	39	34	0	38	18	12	3	77	43	64	20	65	27	72	5
ESCA	38	21	27	2	46	30	52	8	54	18	35	0	15	12	16	0
GBM	77	5	54	0	69	64	61	0	62	4	66	25	31	16	22	0
HNSC	54	57	55	42	54	18	46	0	23	4	7	4	46	8	24	7
KICH	23	8	17	1	0	0	0	0	69	30	68	8	8	6	5	1
KIRC	62	21	39	16	31	19	16	30	54	6	28	3	69	46	83	12
KIRP	38	18	28	10	46	35	25	0	38	21	22	2	23	9	18	0
LAML	50	11	24	8	83	58	51	53	42	29	44	6	58	44	38	0
LGG	62	33	38	11	85	76	95	46	31	14	25	30	62	36	39	49
LIHC	38	29	27	8	23	14	16	10	54	24	55	5	8	7	5	1
LUAD	54	32	39	22	31	17	16	0	23	13	17	12	54	20	56	9
LUSC	38	14	19	13	15	18	9	2	92	97	81	13	8	2	6	0
OV	69	5	65	33	38	15	22	25	23	13	9	0	77	49	80	10
PAAD	69	50	54	19	100	53	51	26	54	23	44	35	46	27	55	30
PCPG	46	24	32	0	92	25	65	42	54	2	56	50	62	35	0	1
PRAD	38	31	24	0	15	7	8	0	46	12	27	36	46	41	17	0
READ	0	0	0	0	17	7	7	1	25	11	22	13	0	0	0	0
SARC	54	45	34	42	77	33	77	13	46	16	39	24	23	13	9	5
SKCM	85	64	59	23	69	42	40	26	85	58	67	55	77	25	70	7
STAD	58	23	28	4	54	23	35	18	27	14	22	17	12	8	8	1
TGCT	46	24	46	0	69	8	60	0	85	14	91	23	8	1	4	3
THCA	77	69	86	27	69	24	57	9	23	8	19	7	85	97	22	0
THYM	62	48	52	3	69	47	61	14	69	56	65	18	54	0	41	12
UCEC	54	12	45	1	58	0	21	13	19	1	15	3	54	27	44	15
UCS	23	17	14	0	31	19	27	24	23	1	7	0	8	2	2	0
UVM	54	6	15	15	69	25	50	0	69	22	74	12	100	55	37	0

Dimensionality reduction (DM) was performed on the four most accurate ML algorithm and mutation dimension pairings. The numbers represent the percentage of cases in each cancer type that were correctly predicted. DM was detrimental to prediction accuracy overall, but beneficial in very few cancer types, which are shown in white font with red background.

Prediction algorithms - Bagging: Bagging decision tree classifier, Gnb: Gaussian Naive Bayes

Mutational dimensions - recur vars: Recurrent variants, genes: Mutated genes, trinucleotide: Trinucleotide mutations, multidimensional: Multidimensional mutational data.

Dimensionality reduction algorithms - NDR: No dimensionality reduction, RBM: restricted Boltzmann machine, PCA: principal components analysis, VT: variance threshold (0.2)

3.3.3. Phase III: Prediction accuracy greatly improved by combinatorial approach

To improve the prediction accuracy a second approach was undertaken. Rather than lowering the number of features via dimensionality reduction, a combinatorial approach was taken, with the four most accurate algorithms separately combined with two of the most accurately predicted datasets for each algorithm. Using the same approach as in 3.3.1, cancer type prediction was made for cases from the TCGA test dataset. The ML training process was preformed independently for each case with this individual case excluded from the learning dataset. Table 21 shows the prediction accuracies as percent correct predictions from each ML + mutational dimension combination for each cancer with the mean of all cases also shown. The combinations are arranged in order of prediction accuracy.

The random forest classifier with the trinucleotide and recurrent variants dimensions (RFC+ trinucleotides+recur vars) generated the most accurate predictions with an overall accuracy of 88% across all cases. This was followed by 84%, 79% and 60% for the bagging classifier with mutated genes and mutational distribution (Bagging+genes+distribution), the Gaussian Naive Bayes with trinucleotides and recurrent variants (Gnb+trinucleotides+recur vars) and the gradient boosting with trinucleotides and mutational distribution (GBEst+trinucleotides+distribution) respectively. By taking this combinatorial approach, the best predictor provided a 36% improved prediction accuracy over the best predictor in the non-combinatorial approach (Table 19). Despite the overall improvement, prediction was very poor for both the cholangiocarcinoma (CHOL) cases where there were no correct predictions were made and cervical squamous cell carcinomas (CESC) where prediction accuracies ranged from 8 -23 %. Some pairings from non-combinatorial approach (section 3.3.1) did

generate better predictions for both these cancers, specifically Gaussian Naive Bayes with trinucleotides alone generated 54% accuracy for the CHOL, while the bagging classifier with the multidimensional data generated an accuracy of 46% with CESC. Although these ML-data dimension pairings generated superior accuracies for these cancers in particular, a diagnostic application requires a single algorithms to be able to make correct predictions the majority of the time whatever the cancer that is presented in the clinical setting. Therefore, despite the prediction limitations, the three most accurate algorithms shown in Table 21 were selected for accuracy testing in non-TCGA datasets, as these are the most accurate across all cancer types.

Table 21: Phase III - Prediction accuracy improved by combinatorial approach

ML algorithm	RFC	Bagging	Gnb	GBEst
Mutational dimension 1	trinucleotides	genes	trinucleotides	trinucleotides
Mutational dimension 2	recur vars	distribution	recur vars	distribution
Means	88	84	79	60
Adrenocortical carcinoma (ACC)	77	38	92	69
Bladder Urothelial Carcinoma (BLCA)	62	38	38	69
Breast invasive carcinoma (BRCA)	92	69	46	85
Cervical squamous cell carcinoma (CESC)	8	15	31	23
Cholangiocarcinoma (CHOL)	0	0	0	0
Colon adenocarcinoma (COAD)	50	62	38	50
Esophageal carcinoma (ESCA)	100	100	100	62
Glioblastoma multiforme (GBM)	100	100	100	62
Head and Neck squamous cell carcinoma (HNSC)	100	100	85	69
Kidney Chromophobe (KICH)	100	100	100	77
Kidney renal clear cell carcinoma (KIRC)	100	100	54	85
Kidney renal papillary cell carcinoma (KIRP)	100	100	92	46
Acute Myeloid Leukemia (LAML)	100	100	100	58
Brain Lower Grade Glioma (LGG)	100	92	85	62
Liver hepatocellular carcinoma (LIHC)	100	92	92	62
Lung adenocarcinoma (LUAD)	100	100	100	69
Lung squamous cell carcinoma (LUSC)	92	92	77	38
Ovarian serous cystadenocarcinoma (OV)	100	92	69	69
Pancreatic adenocarcinoma (PAAD)	100	100	100	92
Pheochromocytoma and Paraganglioma (PCPG)	92	85	92	62
Prostate adenocarcinoma (PRAD)	100	100	77	62
Rectum adenocarcinoma (READ)	83	83	58	25
Sarcoma (SARC)	100	100	100	15
Skin Cutaneous Melanoma (SKCM)	100	100	100	85
Stomach adenocarcinoma (STAD)	100	88	88	62
Testicular Germ Cell Tumors (TGCT)	85	85	85	62
Thyroid carcinoma (THCA)	100	100	62	69
Thymoma (THYM)	92	92	92	69
Uterine Corpus Endometrial Carcinoma (UCEC)	100	100	96	65
Uterine Carcinosarcoma (UCS)	92	92	92	62
Uveal Melanoma (UVM)	100	100	100	73

The numbers and coloured intensities represent the percentage of cases in each cancer type that were correctly predicted. Prediction accuracies was greatly improved with the combination of two dimensions of data with the various algorithms when compared with the same algorithms using just one dimension. Only cholangiocarcinoma proved impossible to predict, while adrenocortical carcinoma, cervical squamous cell carcinoma and colon adenocarcinoma had relative poor prediction accuracies.

ML algorithms - RFC: Random forest classifier, Bagging: Bagging decision tree classifier, Gnb: Gaussian Naive Bayes, GBEst: Gradient Boosting.

Mutational dimensions - recur vars: Recurrent variants, genes: Mutated genes, distribution: Genomic distribution, trinucleotide: Trinucleotide mutations, multidimensional: Multidimensional mutational data.

3.3.4. Phase IV: Prediction in WES dataset accurate, inaccurate in WGS datasets

30 whole exome sequencing (WES) and 25 whole genome sequencing (WGS) studies from the ICGC database were used to validate cancer type prediction with a non-TCGA dataset, shown in Table 22 and Table 23 respectively. These represent all cancers which have a corresponding cancer type in the TCGA. Overall, the most accurate prediction in both WES and WGS was the random forest classifier with trinucleotides and recurrent variants (RFC+ trinucleotides+recur vars) having overall 69% and 41% correct predicted samples in the WES and WGS studies respectively. As can be seen, the accuracies in the WES are largely superior to WGS studies, although the distinction is not absolute, for example, seven WES studies have no correct prediction, while the liver WGS study had a relative high accuracy of 88% with the random forest classifier with trinucleotides and recurrent variants. A WGS chronic lymphocytic leukaemia study also had the same accuracy when predicted using the Gaussian Naive Bayes with trinucleotides and recurrent variants (Gnb+trinucleotides+recur vars). It is likely that mutational patterns differ between WGS and WES samples, due to the difference in sequencing coverage, where only 1.5 of the genome is analysed in WES, while the entire genome is analysed in WGS. The coverage in WES studies tend to be much higher than WGS, which also equates to more reliable variant calling, and thus higher fidelities of mutational patterns. The large variations in accuracies may also be attributed to the fact that the different ICGC studies utilise a large number of different bioinformatics pipelines that may, in some studies, generate results inconsistent with the algorithms trained on the TCGA data.

In order to select the most optimised algorithm, an area under the receiver operating characteristic curve (AUROC) analysis was performed on the three ML-mutational dimension combination (Figure 44). This approach allows the choice of the

best estimator should there need for threshold modulation to increase specificity by the exclusion of non-reliable predictions. As can be seen the best predictor based purely on percent of correct predictions was RFC+trinucleotides+recur vars having an AUROC value of 0.85, which would qualify it as a robust estimator. Bagging+genes+distribution generated an AUROC of 0.76, a fair estimator. Gnb+trinucleotides+recur vars created a score of 0.5, which is a poor estimator, specifically an incorrect prediction is just as likely as a correct prediction. Due to the fact that the RFC+trinucleotides+recur vars combination is the most superior predictor, it has been implemented into the MutProfiler web tool, which is a user friendly cancer type prediction web tool which can be used in both research and clinical application where either the mutational dimensions are required or a cancer type prediction is needed (Figure 46).

Table 22: Phase IV - Prediction accuracies seen in the WES datasets

ML algorithm	RFC	Bagging	Gnb
Mutational dimension 1	trinucleotides	genes	trinucleotides
Mutational dimension 2	recur vars	distribution	recur vars
Samples mean accuracy	69	59	50
Ovarian Serous Cystadenocarcinoma	100	97	65
Lung Squamous Cell Carcinoma	99	57	81
Breast Cancer	99	87	36
Rectum Adenocarcinoma	99	78	99
Chronic Myeloid Disorders	98	100	95
Colon Adenocarcinoma	98	86	89
Acute Myeloid Leukemia	97	95	93
Brain Glioblastoma Multiforme	97	80	87
Prostate Adenocarcinoma	96	92	70
Skin Cutaneous melanoma	95	77	91
Brain Lower Grade Glioma	87	81	71
Gastric Adenocarcinoma	87	43	52
Kidney Renal Clear Cell Carcinoma	80	77	30
Bladder Urothelial Cancer	75	34	74
Breast Triple Negative/Lobular Cancer	69	49	7
Uterine Corpus Endometrial Carcinoma	49	57	74
Kidney Renal Papillary Cell Carcinoma	49	42	71
Bone Cancer	42	52	80
Liver Hepatocellular carcinoma	36	13	12
Gastric Cancer	33	0	11
Cervical Squamous Cell Carcinoma	30	38	45
Bladder Cancer	26	26	15
Oral Cancer	2	13	11
Acute Myeloid Leukemia	0	1	12
Benign Liver Tumour	0	0	5
Biliary Tract Cancer	0	1	3
Head and Neck Thyroid Carcinoma	0	0	0
Liver Cancer	0	0	0
Lung Cancer	0	6	0
Renal Cancer	0	0	0

The three most accurate combinatorial approaches were applied to the ICGC cancer datasets. Presented in this table are the results from the whole exome sequencing (WES) datasets. The numbers and coloured intensities represent the percentage of cases in each cancer type that were correctly predicted. The most accurate predictor was RFC with trinucleotide mutations and recurrent variants, with a samples mean accuracy of 69% in the WES datasets.
ML algorithms - RFC: Random forest classifier, Bagging: Bagging decision tree classifier, Gnb: Gaussian Naive Bayes.
Mutational dimensions - recur vars: Recurrent variants, genes: Mutated genes, distribution: Genomic distribution, trinucleotide: Trinucleotide mutations, multidimensional: Multidimensional mutational data.

Table 23: Phase IV - Prediction accuracies seen in the WGS datasets

ML algorithm	RFC	Bagging	Gnb
Mutational dimension 1	trinucleotides	genes	trinucleotides
Mutational dimension 2	recur vars	distribution	recur vars
Samples mean accuracy	41	37	30
Liver Cancer	88	28	0
Skin Cancer	76	30	63
Liver Cancer	56	16	1
Acute Lymphoblastic Leukemia	56	62	88
Skin Adenocarcinoma	48	11	26
Chronic Lymphocytic Leukemia	18	61	87
Esophageal Adenocarcinoma	10	1	7
Lung Cancer	10	10	30
Liver Cancer	7	6	2
Colorectal Cancer	4	22	6
Ovarian Cancer	0	0	36
Pancreatic Cancer Endocrine neoplasms	0	0	0
Pancreatic Cancer	0	0	0
Thyroid Cancer	0	13	0
Prostate Adenocarcinoma	0	2	2
Pancreatic Cancer	0	26	6
Pancreatic Cancer	0	38	2
Esophageal Cancer	0	3	1
Prostate Adenocarcinoma	0	6	4
Soft Tissue cancer	0	4	0
Early Onset Prostate Cancer	0	0	0
Renal Cell Cancer	0	18	0
Pediatric Brain Cancer	0	0	0
Malignant Lymphoma	0	36	66
Neuroblastoma	0	1	0

The three most accurate combinatorial approaches were applied to the ICGC cancer datasets. Presented in this table are the results from the whole genome sequencing (WGS) datasets. The numbers and coloured intensities represent the percentage of cases in each cancer type that were correctly predicted. The most accurate predictor was RFC with trinucleotide mutations and recurrent variants, with a samples mean accuracy of 41%.

ML algorithms - RFC: Random forest classifier, Bagging: Bagging decision tree classifier, Gnb: Gaussian Naive Bayes.

Mutational dimensions - recur vars: Recurrent variants, genes: Mutated genes, distribution: Genomic distribution, trinucleotide: Trinucleotide mutations, multidimensional: Multidimensional mutational data.

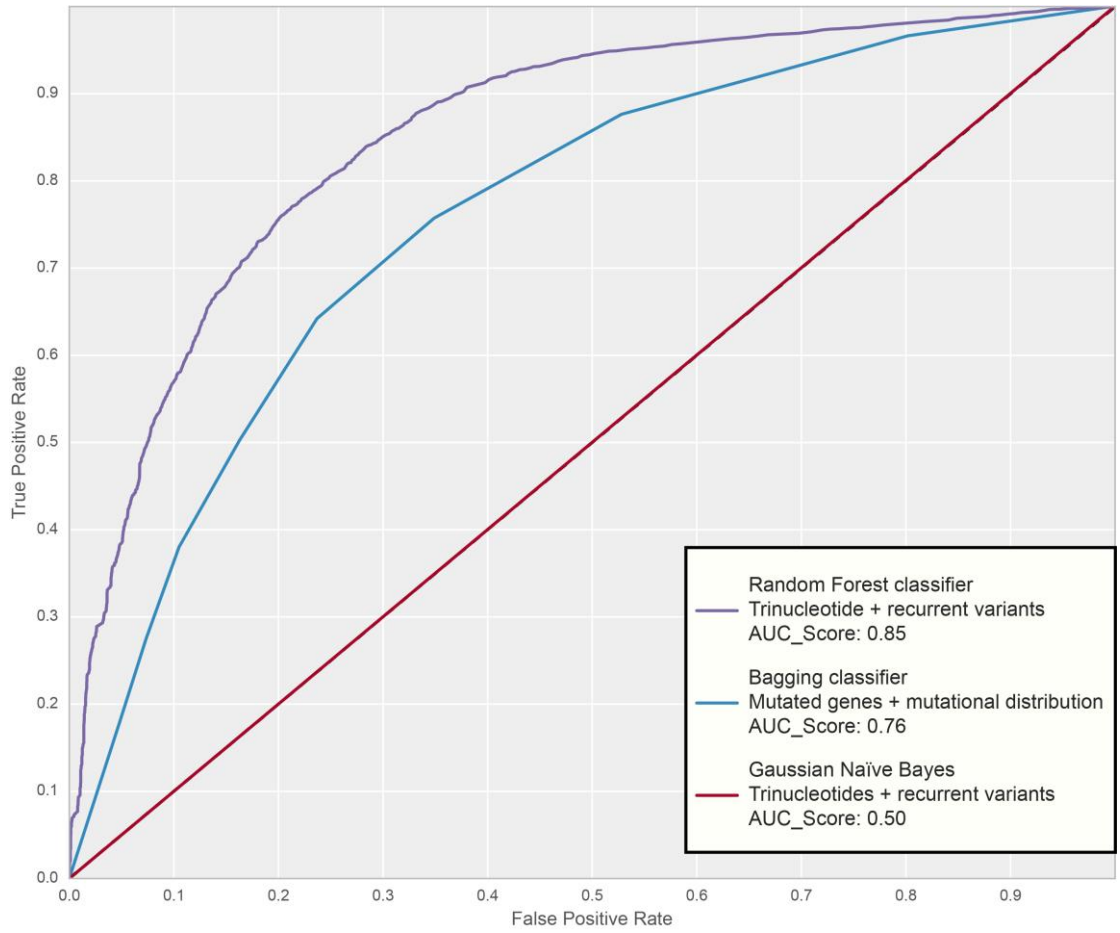

Figure 44: ROC and ROAUC results from the three best performing combinatorial predictors

The graph represents the ROC curves and corresponding ROAUC (AUC) for the three best classifiers. A diagonal line would represents the theoretical performance of a random classifier, and this corresponds precisely to the ROC of the Gaussian naive Bayes classifier. The predictor with the greatest ROAUC is the random forest classifier with trinucleotide mutations and recurrent variants with value of 0.85, suggesting that it is a robust classifier of cancer types when used across all utilised cancers from the ICGC dataset.

3.3.5. MutProfiler with optimised algorithm applied to a CTC WES study

In order to test whether MutProfiler and the optimised prediction methodology is truly applicable to a clinical setting, specifically with CTCs, a CTC WES dataset (Ni et al. 2013) was run through MutProfiler. Correct predictions are shown in red font in Table 24. Of the metastases, 3 of 4 samples (75%) were predicted correctly. 2 of the 5 CTC samples were correctly predicted (40%). Overall 50% of all samples were predicted correctly. The probabilities associated with the predictions, indicating the reliability are listed. These results are highly encouraging due to the correct predictions, however, the majority of CTCs (60%) were incorrectly predicted. This is likely due to the great challenge with CTC sequencing. Firstly, CTC sequencing can be contaminated by lymphocytes and thus the sequenced cell may in fact be non-cancerous. Furthermore, this study specifically employs single cell sequencing which is more prone to technical artefacts and noise introduced during single-cell isolation and genome interrogation than sequencing from multiple cells, and thus more likely to have biases in the mutational patterns (Gawad, Koh, and Quake 2016). Despite the limitations, these results do show the potential for cancer type prediction from CTCs, but also reveals the challenges. At this stage it would appear that for a clinical diagnostic application MutProfiler should be accompanied with other methods of cancer type detection, with both approaches being complementary.

Table 24: Accuracy of cancer prediction from CTC

Patient	Sample type	MutProfiler prediction	Probability
Patient1	CTC	Breast carcinoma	0.195
	Metastasis	Breast carcinoma	0.249
	Primary tumour	Breast carcinoma	0.311
Patient2	CTC	Kidney renal clear cell carcinoma	0.119
	Metastasis	**Lung adenocarcinoma**	0.119
Patient3	CTC	**Lung adenocarcinoma**	0.197
	Metastasis	**Lung adenocarcinoma**	0.212
Patient4	CTC	Breast carcinoma	0.224
Patient8	CTC	**Lung adenocarcinoma**	0.161
	Metastasis	**Lung adenocarcinoma**	0.199

Prediction accuracy using MutProfiler and the most optimised ML was applied to the Ni et al. 2013 dataset. Correct prediction was made in 50% of the cases. This includes correct predictions in 75% of the metastasis and 40% of the CTCs.

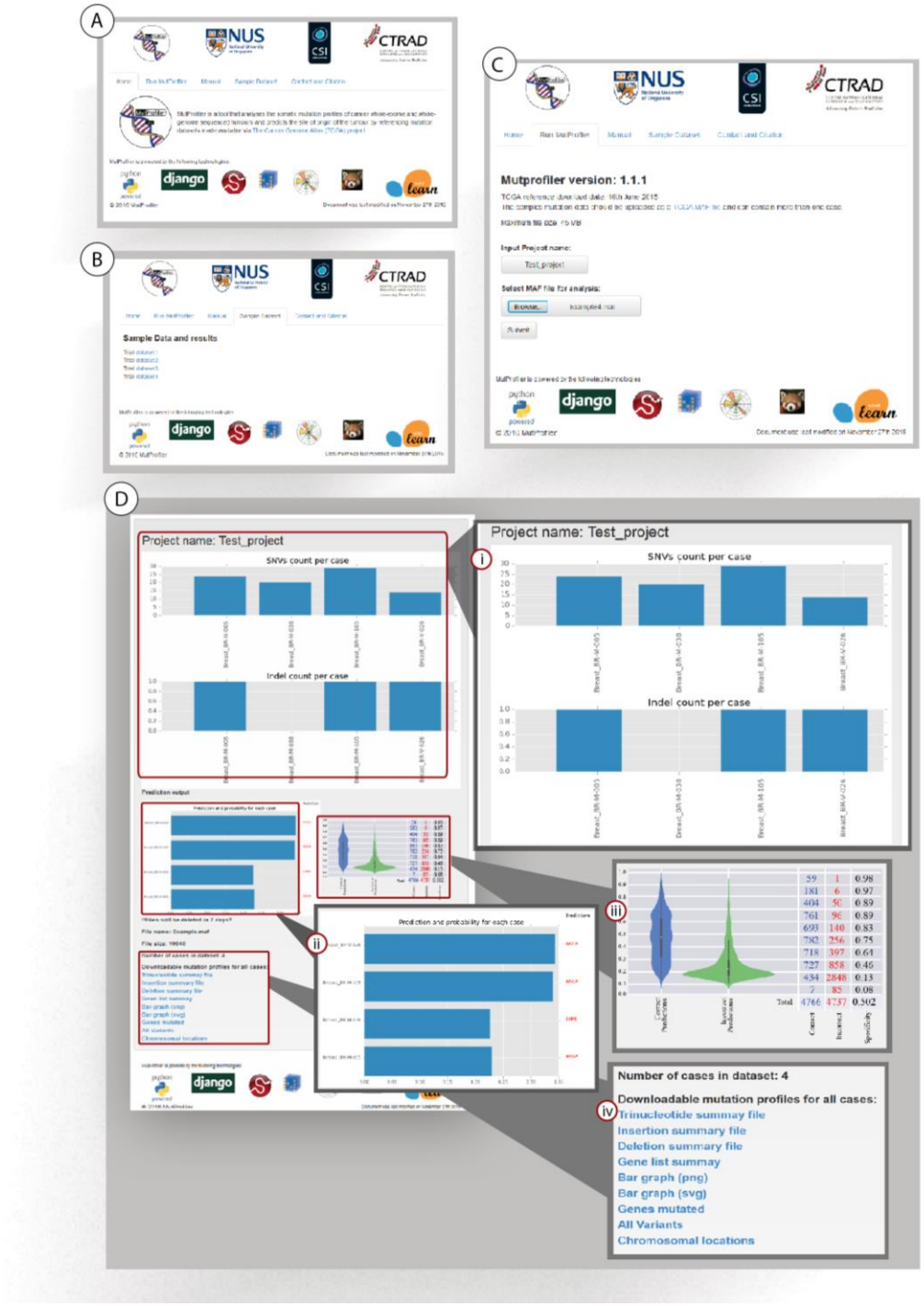

Figure 45: Interface and Usage

(A) MutProfiler implements a Random Forest classifier that utilised the TCGA trinucleotide mutations and recurrent variants as its learning dataset. (B) Several trial datasets are available to test out MutProfiler's functionality (C) The interface allows users to upload mutation data as a MAF file. (D i) A summary of the SNVs and indels will be displayed (D ii) The predicted cancers and the associated probability is included in the output. (D iii) A sensitivity and specificity plot is displayed to allow the user to accept or reject the prediction based on a personal risk assessment. (D iv) The data matrix for each of the analysis dimensions for all uploaded samples can be downloaded.

3.4. Discussion

3.4.1. A comprehensive list of ML algorithms was applied

The approaches listed in Figure 43 are a comprehensive list of ML prediction and dimensionality reduction approaches that had the potential to predict a tumour of origin. The prediction (supervised learning) techniques can be divided into either decision tree techniques (Random forest, Gradient Boosting, AdaBoost, Bagging), linear classifiers (Support Vector Machines) or Bayesian approaches (Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes). These classifiers differ at a fundamental level, where, as the name would suggest, decision trees seek to create learning and prediction processes through a process of decision where options fork into the various supervised categories. Linear models, however, assume that the supervised categories can be delineated via linear algebraic relation equations and in the case of polynomial classifiers that is done through Kernel methods where the data is mapped into high-dimensional feature spaces. The Bayesian techniques used frequency tables to establish the probabilities of features being associated with specific cancer types.

3.4.2. Strategy of Testing of Several Machine Learning Algorithms

ML has in many ways significantly impacted science (Wu et al. 2008) and medicine (Foster, Koprowski, and Skufca 2014) through the use of various methodologies. This study goes through the exercise of testing the prediction performance of several ML classifiers. Among the algorithms chosen are supervised methods that are considered to be the most influential in data mining (Wu et al. 2008). This approach, that is, the use of several algorithms to solve biological problems have been utilised by several studies and is considered a typical approach when specific

models are not required and/or when the nature of the dataset is not well understood (Tarca et al. 2007) and is also favourable when computational resources are not a limiting factor for the size of the dataset being analysed. For example, a 2014 study developed a method to predict the solubility of overexpressed recombinant proteins in Escherichia coli by analysing a number of protein features by the use of a large number of machine algorithms (Habibi et al. 2014). The algorithms included Support vector machines (SVM), Sequence pattern-based method, random forest, scoring card method (SCM), conditional inference trees and decision trees. The study concluded that several algorithms provided accurate predictions, with usability independent of the algorithm being the defining factor. Recently, a study (Singh and Singh 2016) tested six machine learning algorithms for their efficiency at classifing Riboswitches, which are small structured RNA elements which modulate the transcription or translation of genes. The study used J48, BayesNet, Naive Bayes, Multilayer Perceptron (MP), sequential minimal optimisation and hidden Markov model (HMM) ML algorithms with the MP algorithm being identified as superior to the rest via the use of several statistical measures including specificity, sensitivity and receiver operating characteristic. Antifungal peptides, part of a larger group of peptides known as host defensive peptides, have been shown to be safer and more effective weapon against fungal threats than chemical based treatments. Another recent study (Mousavizadegan and Mohabatkar 2016) compared five different machine learning algorithms frequently used for classification of biological data to classifying and predicting antifungal peptides, with the SVM algorithm having the highest performance values.

Some other examples include the evaluation of ML algorithms for virtual screening of lead molecules (Vyas et al. 2015; Hizukuri, Sawada, and Yamanishi 2015), quality assessment in three-dimensional breast ultrasound images (Schwaab et al. 2016), lung

cancer classification based on gene expression levels (Podolsky et al. 2016), predicting radiation therapy outcomes (Kang et al. 2015), histopathological and cell image classifier optimization (Abbas, Dijkstra, and Heskes 2014; Zachariah et al. 2014), among many others.

3.4.3. Evaluation of ML by ROC and AUROC

A receiver operating characteristic (ROC) curve is a plot that represents the capabilities of a binary system of classification, e.g. rates of success or failure. It is especially well suited to studying the performance of probabilistic detection and forecast systems. This methodology was developed in the field of radar signal detection (RSD) theory as the "receiver operating characteristics" (Peterson, Birdsall, and Fox 1954), which follows on from work in statistical quality control theory from Bell laboratories (Mason and Graham 2002). Since its implementation in RSD, this approach has been adopted in several fields of science such as psychology (D. and J. 1966; Tsoi et al. 2015; Wichchukit and O'Mahony 2010) and medicine (Zweig and Campbell 1993; Abruzzo et al. 2015; Lotta et al. 2015; Usher-Smith et al. 2016) amongst others. Creation of the ROC curves starts with generating a confusion matrix at each possible threshold (p-value) from the predictions of a probabilistic classifier. The actual ROC curve is generated by plotting a line chart with the number of true positive predictions (typically on the y-axis) against the number of false predictions (typically x-axis) from all the generated confusion (matching) matrices. In addition to the ROC line, a line in the diagonal of the plot is drawn to represent the ROC curve of a random predictor, i.e. it would have a 0.5 chance of a correct prediction at any given threshold. Assuming that the predictor is better than by chance, the ROC would be to the left of the diagonal line. The Area under the Receiver Operating Characteristic curve (AUROC), also known just as AUC, represents the area between the ROC curve and diagonal line, i.e. the

improvement of the predictor over random chance (Hanley and McNeil 1982). In practical terms, when using normalised units, the AUROC is a qualifier of any given predictor and can be used to test the performance of several predictors against each other. This approach is commonly used in determining and comparing the performance of supervised machine learning algorithms and typically used alongside overall prediction accuracy. The advantage of AUROC over just prediction accuracy is that prediction accuracy pertains to just one threshold, typically the inclusion of all data. As described, the AUROC, however, uses all possible thresholds and plots the sensitivity and specificity as such, the classifier is graded in a more robust manner. Following the AUROC selection, the specificity of the classifier can then be improved by selection of prediction thresholds, albeit at the expense of sensitivity. Due to this, the AUROC was utilised to determine the performance of the different classifiers in this study.

3.4.4. Dimensionality reduction was not beneficial

Three dimensionality reduction approaches were used in this study, namely PCA, RBM (which is a stochastic artificial neural network) and lastly feature selection was performed, by applying a VT of 0.2 to all data categories. The dimensionality reduction did not in any way improve the prediction outcomes, in fact, predictions seemed to suffer. This would indicate that the predictive successes seen are probably not dependent on the most variable components of the dataset, i.e. components that have the greatest variance, but rather that the determinant features are of low variance in the dataset and as such would be masked by dimensionality reduction procedures. This is especially telling considering that the VT of only 0.2 caused the largest reduction in prediction accuracy (Table 20). The fact that even the complex dimensionality reduction approach of RBM did not help would suggest that the defining mutational signatures of the cancers require a complex set of rules that requires that all, or most of

the categories of data remain distinct for accurate prediction. Other than the potential for improved prediction, dimensionality reduction may have been more computational efficient due to the comparison of fewer categories of data during cancer type prediction. It is unlikely that other approaches to dimensionality reduction or feature selection would benefit accuracy, therefore this approach was abandoned in favour of the combinatorial approach used in section 3.3.3.

3.4.5. Choice of prediction algorithms and CTC prediction

The most optimised algorithm, as used in MutProfiler, was selected from a multistep process of prediction optimisation. The first step (Phase I) revealed that correct predictions can be made from the individual mutational dimensions when this mutational data is paired with an appropriate ML algorithm. This verifies the hypothesis that there are indeed distinct cancer type or subtype mutational patterns beyond those seen in chapter 2, and that these distinct patterns have enabled cancer type prediction. Following this (Phase II), it was shown that dimensionality reduction was detrimental to prediction accuracy likely due to the loss of fidelity that is necessary to decipher the subtle characteristics that are required to identify cancer types. A combinatorial approach (Phase III) was then performed where the four best performing algorithms were combined with the two datasets that the provided the best predictions with those algorithms, specifically done to improve prediction for diagnostic usage. In Phase III, a AUROC analysis revealed that one combination in particular, the random forest classifier with the trinucleotide and recurrent variants dimensions (RFC+ trinucleotides+recur vars), stood out as the most robust predictor with a AUROC of 0.85. It was also revealed that predictions in WES samples were much better than in WGS samples, suggesting that WGS may require a separately trained algorithm. The RFC+ trinucleotides+recur vars combination was integrated into a web interface, thus

creating MutProfiler, a tool to generate mutational dimension data and perform cancer type prediction. MutProfiler was then applied to a CTC WGS study, where it was revealed the 2/5 CTC samples were correctly predicted (40% accuracy) suggesting that this method may eventually be used to non-invasively determine the site of origin/tumour types based solely on the mutational profile within CTC samples. Again it should be noted the test datasets including the CTC samples underwent different wet-lab and bioinformatics processes and therefore reduced prediction accuracy may be attributed to this. An ideal scenario would be to have a CTC or CUPs samples sequenced in conditions identical to their TCGA counterparts, and maybe even done as a standard clinical practice.

3.4.6. MutProfiler: Potential diagnostic tool

Figure 46 explains the rationale behind the development of MutProfiler, i.e. a tool that can accept mutational data and summarise this data in terms of mutations profiles, and to also perform a prediction of the cancer subtype.

The intent is that MutProfiler will be used in a clinical setting where a patient's cancer type prediction is required, specifically where the prediction of the correct cancer will help determine a correct diagnosis and direct a patient towards beneficial treatment strategies. Future versions of MutProfiler may also include prognosis or therapeutics outcome prediction by finding associations between mutational patterns of the different dimensions and these factors, increasing the diagnostic application.

The accuracy at predicting using WES studies suggests that the prediction methodology is robust with this type of sequencing analysis, however the predictions are not definitive, and therefore should be used alongside other techniques of cancer type prediction. This is further illustrated by the 40% correct prediction in the CTC

WES study. The output from this web tool including the mutations results from the various mutational dimensions is shown in Figure 45.

3.4.7. Currently available methodologies in cancer type prediction

Currently, pathology analysis and IHC remain the 'gold standard' for CUP diagnostic workup (Oien and Dennis 2012) and work by identifying IHC biomarkers have been identified on a candidate basis, as single genes involved in a particular process. This approach may take up to three rounds of analysis, where the first is to diagnose the CUP into the broad class of cancer types, e.g. whether it is a carcinoma (Cytokeratin AE1 / AE3), melanoma (S100), leukaemia (common leukocyte antigen) or sarcoma (Vimentin) by the staining of approximately 5 or 6 antibodies, with the identifying antibody shown in brackets. If diagnosed as a carcinoma, this is often followed by a second IHC staining panel to determine carcinoma subtype, e.g. adenocarcinoma (cytokeratin 7), squamous cell carcinoma (cytokeratin 5) etc. Adenocarcinomas are then treated to the third panel of antibodies to determine primary site, for example, prostate-specific antigen to identify prostate adenocarcinomas, or mesothelin to identify stomach adenocarcinomas. Although commonly implemented in clinical settings, the accuracy of these techniques are not well validated. The multistep process also requires a large amount of sectioned tissue and a long processing time. Table 25 shows a summary of the attributes of the IHC method, along with a summary of the other existing site of origin prediction methods.

Molecular techniques are being introduced with the promise of increased accuracy, speed, and reduced turnover time. These techniques also remove the subjectivity involved with pathological interpretation (Oien and Dennis 2012). MutProfiler is primed to play a role in addressing the needs of molecular profiling techniques by the use of DNA sequencing. Currently, the Cancer type ID from Biotheranostics (San

Diego, California, USA) is gaining acceptance as a standard for the identification of CUPS by molecular testing (Table 25, column 3). It is able to distinguish 28 main tumour types and 50 tumour subtypes, which represent approximately 95% of cancers by incidence. This is achieved by the use of a 92-gene real-time RT-PCR analysis TaqMan probe panel (Brachtel et al. 2016) requiring approximately 300 non-necrotic cells as starting material for RNA extraction. Accuracies of 100%, 92%, and 86% in FNA/cytology cell blocks, core biopsies, and small excisions. Although this approach has seemingly higher accuracies than MutProfiler, this approach is very much a complementary methodology. The use of RNA as required by this approach is not often feasible and certainly not a common part of clinical practice.

Recently TumourTracer (Marquard et al. 2015) a software package with similar functionality to MutProfiler has been published and it claims to be able to predict the tissue of origin of a tumour based on the mutational profile (Table 25, column 4). This software seems to have been developed in parallel to MutProfiler and employs the similar methodology of using machine learning approaches and mutational data as prediction features. In addition to mutational data, TumorTracer also uses copy number changes in the 232 cancer-related genes represented as trinary data (copy number loss, gain or normal copy number). TumorTracer produced accuracies of 46 %, 53 % and 89 % for three independent datasets. Whereas using the algorithm used with MutProfiler, accuracies of 88% were seen in the TCGA datasets as a whole, 69% in the ICGC WES studies, 41 % in the ICGC WGS studies and 40 % in the CTC study. This range is very similar to those seen in TumorTracer. It should be noted however the individual studies within the TCGA and ICGC may be considered separate data points, for which accuracies ranged from 0 to 100% using 87 different studies (31 from the TCGA, 55 from the ICGC and the CTC study). It is therefore very clear that MutProfiler has been

rigorously tested on a vastly greater number of studies and cases than TumorTracer, and it is therefore more likely that prediction accuracies shown for MutProfiler in this thesis are representative of real-world prediction accuracy compared to the prediction ranges determined for TumorTracer.

MutProfiler is also superior for several reasons. Firstly, TumourTracer utilises COSMIC version 68 as its mutational database to identify cancer type specific mutations. The cosmic database (S. a Forbes et al. 2011) is a publically available database of disease mutational data provide by the Sanger institute. While the COSMIC database itself is well curated, the mutation data is sourced from a wide range of studies of indeterminate quality and reliability. The work presented in this thesis has used the TCGA database which is a very highly curated database with representations from 31 cancer types at the time of this study. This work also utilises over 8000 cases for establishing the ML classifiers, as opposed to the 4975 used by TumorTracer.

Also, the TumorTracer's web interface is limited by the necessity of using MuTect for variant calling, making it technically challenging to use with other variant calling approaches, which would negate its use with the majority of sequencing data being generated. Support for Illumina-based methods e.g. Strelka variant calling would make it much more accessible. The web interface of MutProfiler, however, relies on MAF files, which are agnostic to variant calling platforms. Although the standardised out format of most variant callers e.g. GATK Unifiedgenotyper and Halotypecaller, Stelka, MuTect and VarScan are VCF files, conversion to the MAF format can be done by a variety of tools, notable Oncotator (Ramos et al. 2015), a tool that is available from the Broad institute. The use of MAF files specifically is necessitated by the need for gene-specific annotation. A direct comparison between MutProfiler and TumorTracer is challenging because of the need specifically for MuTect called variants.

MutProfiler is compatible with mutation annotation from three versions of the human genome, i.e. hg18 for purposes of being compatible with legacy datasets, hg19 which is not the newest version of the human genome, but currently the most commonly implemented in NGS pipeline (e.g. Illumina's basespace) and hg38, the newest version of the human genome and will become more vital to NGS platforms as this version of the genome matures and NGS systems are updated. MutProfiler does this by the inclusion of UCSC Liftover as discussed in section 2.2.2.3 which is a tool developed by University of California, Santa Cruz for coordinate conversion between genome builds. TumorTracer, however, is only implemented with the hg19 version of the genome.

This study has also gone further than TumorTracer, by not only looking at the 232 cancer specific genes which are based on an older dataset (Araya et al. 2015) and the trinucleotide changes, but this thesis has as looked in-depth into other features for data description and classification. We have investigated indels which were not considered by TumorTracer and studied all mutated genes in all 31 of the available cancers, thereby not biasing the feature selection to the cancers within the previous studies.

Admittedly this study has not looked at copy number changes as utilised by TumorTracer, however, this may not be a major limitation in practical implementation. Unlike copy number changes, variant analysis by WES is gradually being implemented in the clinical and diagnostic settings (Eliezer M Van Allen et al. 2014; Retterer et al. 2016). More and more NGS is allowing clinicians to make diagnoses and recommend treatment strategies based on mutations found in clinically relevant genes. Furthermore, WES is being implemented to allow for the discovery of potentially relevant mutations that do not yet have associated therapeutics. This clinical implementation will allow the DNA mutation aspect of MutProfiler to be its true strength. In addition, MutProfiler can

play a role in the complement of independent molecular profiling techniques and thus allow clinicians and patients to be better informed.

3.4.8. MutProfiler in combination with the other cancer type prediction methods

The costs associated with NGS procedures have been dramatically falling, outpacing even Moore's law (Muir et al. 2016). This along with its advantages compared to traditional diagnostic techniques have increased its usage in both cancer (Luthra et al. 2015) and non-cancer (Matthijs et al. 2016) diagnoses. In addition there is the potential for revealing additional diagnostic insight in the future through the identification of variants of unknown significance (Petersen et al. 2017). The inevitability of NGS, specifically WES, as a diagnostic tool places MutProfiler firmly in a position to aid in diagnosis of CTCs and CUPs. As the price point of sequencing reduces, the advantages of WES compared to IHC will only increase. The huge number of antibodies required to account for all possible tissue origins, and the lack statistics on prediction accuracy are large weaknesses compared to the other methodologies. The limitation of prediction accuracy does however indicate that the current version of MutProfiler is not sufficient. This is in fact true for all the site of origin prediction methods (Table 24), as none are infallible. With the push towards precision medicine, the pressure will only increase for more targeted and effective therapies with greater patient survival with fewer side-effects and lower costs (Pinato et al. 2017; Soong et al. 2016). Due to this, currently, the only viable approach would be to combine as many methods as possible for CUPs or CTC prediction with the potential of improved prediction via a consensus of the predictions, although this approach has to be tested. The combination is also a challenge in itself due to prohibitive costs and burdens on healthcare systems, the requirement for RNA for the Biotheranostics test and for the potential for confusion or distress due to conflicting results.

Table 25: Summary of cancer type prediction methodologies

	IHC	Biotheranostics	TumourTracer	MutProfiler
Status	**Gold standard**	Gaining clinical acceptance	Research	
Accuracy	Not well validated	**86 - 100 %**	46 - 89 %	**40 - 88 % (0 -100%)**
Priniciple	Protein expression	92-gene expression	Trinucleotide and copy number	Trinucleotide and Variant frequencies
CUPS	**y**	**y**	**y**	**y**
CTC	n	n	**y**	**y**
Cancers types	All	50	10	33
Turnaround time	1 - 3 weeks	Delivery (1 week)	**3 days**	
Sample type	Sectioned tissue	RNA	**DNA**	**DNA**
Limitation			MuTect mandatory	
Requirements	Histology lab	RNA extraction facility	Sequencing machines	
Reference database			COSMIC v68 4975 cases	TCGA database
Compatible genome builds			GRCh37	**GRCh36 (hg18)** **GRCh37 (hg19)** **GRCh38 (hg38)**
Compatible variant callers			MuTect	**All variant callers**
Validated datasets			3	**86 (TCGA/ICGC)**

Shown are the four established methods for cancer type prediction. The main strengths of each method are shown in blue font. The greatest strengths of MutProfiler are that it leverages on DNA sequencing (which is cheap to analyse and entering into the standard clinical analysis), highly accurate, has a low turnaround time, can be used with CTCs and is well validated, inclusive of the TCGA and ICGC studies. MutProfiler accuracies are based on databases (outside bracket) and on individual studies (within brackets).

3.4.9. Future work

Non-DNA based datasets have been shown to be capable of characterization of cancer site of origin prediction. A study used RNA expression by real-time reverse transcriptase-polymerase chain reaction (RT-PCR) (Greco et al. 2010), to predict the site of origin on CUPs, but this was tested on a very small dataset of only 28 patients with CUPs. The study achieved a 75% correct prediction however the prediction was poorly explained. Another study has however shown that cell of origin chromatin organisation seems to drive mutation patterns in cancer genomes, linking there two separate genomic events (Polak et al. 2015). The authors studied the mutations from 173 cancer genomes from eight different cancer types that represent a wide range of tissues of origin, carcinogenic mechanisms, and mutational signatures, namely, melanoma, multiple myeloma, lung adenocarcinoma, liver cancer, colorectal cancer, glioblastoma, oesophageal adenocarcinoma and lung squamous cell carcinoma and compared this to 424 epigenetic features derived from 106 different cell types from 45 different tissue types. The study developed a predictor based on enrichment of epigenomic variables from a single cell type among the top 20 variables selected by the random forest analysis.

Future work concerning MutProfiler and subtype prediction will involve the inclusion of other datasets to improve site of origin prediction and thus allow user catered prediction according to the datasets that are available. These will include the integration of RNA expression and methylation, among other features. The ML algorithms used in this study are robust enough to be used with other datasets and as such, including this additional information will be an extension of what has been done already. Again the TCGA database will be used as the source of the additional, owing to it rich source of samples that have several types of genomic information.

MutProfiler's prediction will also be continually updated by including newer versions of the TCGA project mutations to refine the predictor learning process and perhaps even include mutations from other databases such as COSMIC.

There is also the potential that mutation patterns of the different dimensions may have associations with clinical outcome i.e. prognosis or treatment outcome, similar to how increase mutation burden is associated with improved immunotherapy response in lung cancer (N. A. Rizvi et al. 2015), melanomas (E. M. Van Allen et al. 2015) and other cancers (Le et al. 2015). Survival data is available for many studies in both the TCGA and the ICGC databases, and thus this will be investigated further. This will ensure that MutProfiler is not only relevant to CTC and CUPs applications, but also to the clinical analysis of tumour biopsies and resections as well. In this way, future versions of MutProfiler will not only be much more accurate in prediction cancer type, but will also be a much more comprehensive tools for clinical diagnosis.

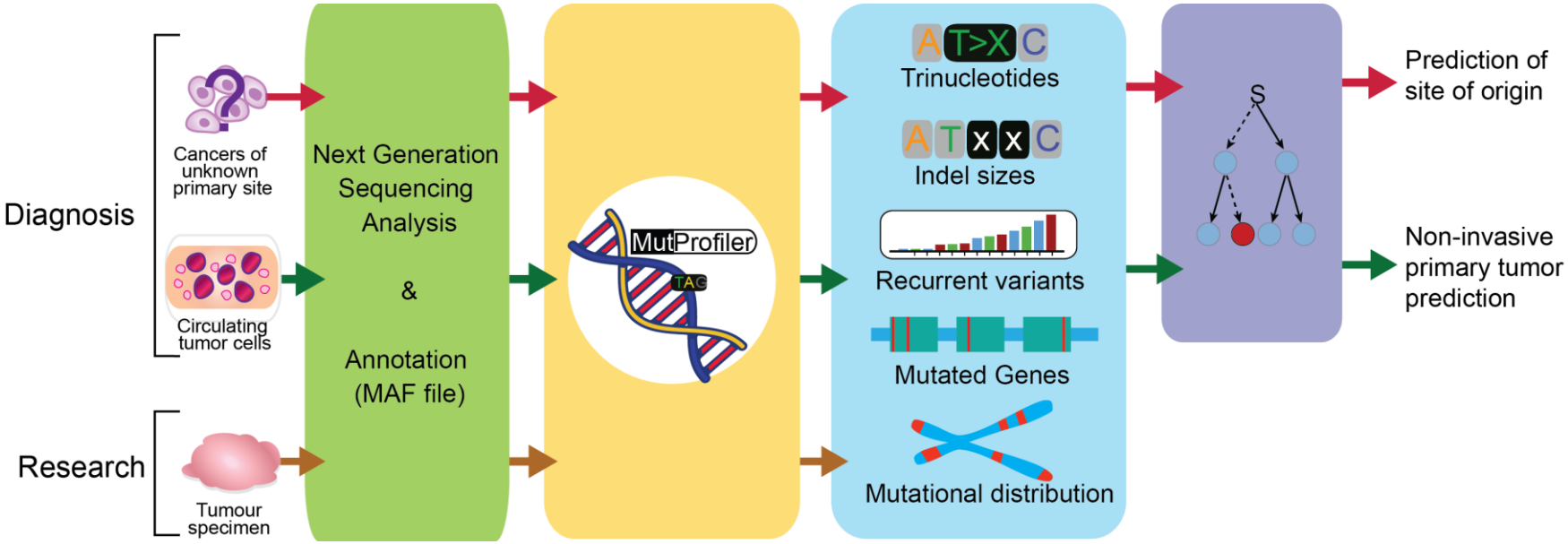

Figure 46: Applications of MutProfiler: Mutation summaries and cancer subtype prediction

MutProfiler can be used for diagnosis to determine the site of origin of a CUP or for a non-invasive tumour diagnosis from CTCs. Data concerning the individual dimensions can be used in further research work.

4. Conclusion

The novel work presented in this thesis is the most comprehensive and up-to-date summary of SNV changes in cancers as it looks at the trinucleotide mutations in 31 cancer subtypes (34 with MSI segregation), with the greatest number of cases. A potentially new mutational signatures was revealed existing in a clusters of testicular germ cell tumours (TGCT). The indel sizes were studied in much greater detail than ever before with only one other study with comparable analysis, albeit a specific focus in this issue (Yang et al. 2010). The genomic distributions features that are unique to specific cancer of subsets of cases have been revealed for the first time, and these observations have to be followed up with studies on DNA accessibility based on epigenetic factors. Genes and variants have been established to be cancer specific, but this work reveals how there are genes and variants that have especially enriched mutation rates compared to other cancers rather than a comparison against non-cancer tissues.

The choice to use proportions to study trinucleotides, indels and genomic distributions, in addition to counts, was to normalize for the variant calling sensitivity of different variant callers and sequencing centres and therefore allow the results to be relevant to all studies not just studies that specifically use the same software packages utilized by the TCGA datasets. Furthermore, this approach has been used by landmark studies looking at mutational signatures (Nik-Zainal, Van Loo, et al. 2012; Alexandrov et al. 2015). The impact of mutational load on the cancers mutation profiles, however, is undeniable as shown in this thesis work (Figure 6) and previous studies (Lawrence et al. 2013) and therefore was included alongside the proportions. Moreover, the effect mutational load has on immunotherapy, for example, makes it an especially pertinent aspect of research (Naiyer A Rizvi et al. 2015).

Work on the various dimensions has proven that there may be several consistent mechanisms that drive the development of certain cancers, seen in all the dimensions studied. Certain cancers can be distinguished by hierarchical clustering alone, while it is unlikely to be possible for other cancers (Table 17).

The work in chapter 3, then establishes that even in cancers where relationships may not be obvious by methods such as hierarchical clustering, deep learning machine learning algorithms can help establish consistent mutational signature events that distinguish different cancer types from each other, even in cancers from non-TCGA datasets which have undergone different bioinformatics procedures. The lung CTC dataset was a prime example of the potential of this tool to predicting cancer site of origin in a clinical setting, but also exposes the limitations of the prediction methodology. More work will be done to refine the prediction accuracy and improve its clinical relevance.

References

Abbas, Syed Saiden, Tjeerd M H Dijkstra, and Tom Heskes. 2014. “A Comparative Study of Cell Classifiers for Image-Based High-Throughput Screening.” *BMC Bioinformatics* 15: 342. doi:10.1186/1471-2105-15-342. http://www.ncbi.nlm.nih.gov/pubmed/25336059.

Abecasis, Goncalo R, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. 2012. “An Integrated Map of Genetic Variation from 1,092 Human Genomes.” *Nature* 491 (7422) (November 1): 56–65. doi:10.1038/nature11632. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3498066&tool=pmcentrez&rendertype=abstract.

Abruzzo, Provvidenza M, Alessandro Ghezzo, Alessandra Bolotta, Carla Ferreri, Renato Minguzzi, Arianna Vignini, Paola Visconti, and Marina Marini. 2015. “Perspective Biological Markers for Autism Spectrum Disorders: Advantages of the Use of Receiver Operating Characteristic Curves in Evaluating Marker Sensitivity and Specificity.” *Disease Markers* 2015: 329607. doi:10.1155/2015/329607.

Adam, Rosalyn M, and David J DeGraff. 2015. “Molecular Mechanisms of Squamous Differentiation in Urothelial Cell Carcinoma: A Paradigm for Molecular Subtyping of Urothelial Cell Carcinoma of the Bladder.” *Urologic Oncology*. doi:10.1016/j.urolonc.2015.06.006. http://www.ncbi.nlm.nih.gov/pubmed/26254697.

Adusumalli, Swarnaseetha, Mohd Feroz Mohd Omar, Richie Soong, and Touati Benoukraf. 2015. “Methodological Aspects of Whole-Genome Bisulfite Sequencing Analysis.” *Briefings in Bioinformatics* 16 (3) (May 27): 369–79. doi:10.1093/bib/bbu016. http://bib.oxfordjournals.org/content/early/2014/05/27/bib.bbu016.short.

Alexandrov, Ludmil B, Serena Nik-Zainal, Hoi Cheong Siu, Suet Yi Leung, and Michael R Stratton. 2015. “A Mutational Signature in Gastric Cancer Suggests Therapeutic Strategies.” *Nature Communications* 6: 8683. doi:10.1038/ncomms9683. http://www.nature.com/ncomms/2015/151029/ncomms9683/full/ncomms9683.html\nhttp://www.nature.com/doifinder/10.1038/ncomms9683.

Alexandrov, Ludmil B, Serena Nik-Zainal, David C Wedge, Peter J Campbell, and Michael R Stratton. 2013. “Deciphering Signatures of Mutational Processes Operative in Human Cancer.” *Cell Reports* 3 (1) (January 31): 246–59. doi:10.1016/j.celrep.2012.12.008. http://www.ncbi.nlm.nih.gov/pubmed/23318258.

Alexandrov, Ludmil B, and Michael R Stratton. 2014. “Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes.” *Current Opinion in Genetics & Development* 24 (February): 52–60. doi:10.1016/j.gde.2013.11.014. http://www.ncbi.nlm.nih.gov/pubmed/24657537.

Alexandrov, Ludmil B., Serena Nik-Zainal, David C. Wedge, Samuel a. J. R. Aparicio, Sam Behjati, Andrew V. Biankin, Graham R. Bignell, et al. 2013. “Signatures of Mutational Processes in Human Cancer.” *Nature* 500 (7463) (August 22): 415–21. doi:10.1038/nature12477. http://www.nature.com/doifinder/10.1038/nature12477.

Alunni-Fabbroni, Marianna, and Maria Teresa Sandri. 2010. “Circulating Tumour Cells in Clinical Practice: Methods of Detection and Possible Characterization.” *Methods* 50 (4) (April): 289–297. doi:10.1016/j.ymeth.2010.01.027.

http://linkinghub.elsevier.com/retrieve/pii/S1046202310000423.

Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. 2016. “Deep Learning for Computational Biology.” *Molecular Systems Biology* 12 (7) (July): 878. doi:10.15252/msb.20156651.
http://msb.embopress.org/lookup/doi/10.15252/msb.20156651.

Asan, Yu Xu, Hui Jiang, Chris Tyler-Smith, Yali Xue, Tao Jiang, Jiawei Wang, et al. 2011. “Comprehensive Comparison of Three Commercial Human Whole-Exome Capture Platforms.” *Genome Biology* 12 (9) (January): R95. doi:10.1186/gb-2011-12-9-r95. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3308058&tool=pmcentrez&rendertype=abstract.

Ascierto, Paolo A, John M Kirkwood, Jean-Jacques Grob, Ester Simeone, Antonio M Grimaldi, Michele Maio, Giuseppe Palmieri, Alessandro Testori, Francesco M Marincola, and Nicola Mozzillo. 2012. “The Role of BRAF V600 Mutation in Melanoma.” *Journal of Translational Medicine* 10 (1): 85. doi:10.1186/1479-5876-10-85. http://translational-medicine.biomedcentral.com/articles/10.1186/1479-5876-10-85.

Ashworth, T. R. 1869. “A Case of Cancer in Which Cells Similar to Those in the Tumours Were Seen in the Blood after Death.” *Australian Medical Journal* 14: 146–7.

Avery, O T, C M Macleod, and M McCarty. 1944. “Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type III.” *The Journal of Experimental Medicine* 79: 137–158. doi:10.1084/jem.79.2.137.

Ayer, Turgay, Oguzhan Alagoz, Jagpreet Chhatwal, Jude W. Shavlik, Charles E. Kahn, and Elizabeth S. Burnside. 2010. “Breast Cancer Risk Estimation with Artificial Neural Networks Revisited.” *Cancer* 116 (14) (April 27): 3310–3321. doi:10.1002/cncr.25081. http://doi.wiley.com/10.1002/cncr.25081.

Babraham/Bioinformatics. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Banerjee, Sulagna, and Ashok Saluja. 2015. “Minnelide, a Novel Drug for Pancreatic and Liver Cancer.” *Pancreatology : Official Journal of the International Association of Pancreatology (IAP) ... [et Al.]* 15 (4 Suppl): S39–43. doi:10.1016/j.pan.2015.05.472. http://www.ncbi.nlm.nih.gov/pubmed/26122306.

Banerji, Shantanu, Kristian Cibulskis, Claudia Rangel-Escareno, Kristin K. Brown, Scott L. Carter, Abbie M. Frederick, Michael S. Lawrence, et al. 2012. “Sequence Analysis of Mutations and Translocations across Breast Cancer Subtypes.” *Nature* 486 (7403) (June 20): 405–9. doi:10.1038/nature11154. http://www.nature.com/doifinder/10.1038/nature11154.

Bass, Adam J., Vesteinn Vésteinn Thorsson, Ilya Shmulevich, Sheila M. Reynolds, Michael Miller, Brady Bernard, Toshinori Hinoue, et al. 2014. “Comprehensive Molecular Characterization of Gastric Adenocarcinoma.” *Nature* 513 (7517) (September 11): 202–9. doi:10.1038/nature13480. http://www.nature.com/doifinder/10.1038/nature13480.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical*

Society. Series B (Methodological) 57 (1): 289–300. http://www.jstor.org/stable/2346101.

Bhattacharya, Bhaskar, Mohd Feroz Mohd Omar, and Richie Soong. 2016. "The Warburg Effect and Drug Resistance." *British Journal of Pharmacology* 173 (6) (March): 970–9. doi:10.1111/bph.13422. http://doi.wiley.com/10.1111/bph.13422.

Bhattacharyya, N P, A Skandalis, A Ganesh, J Groden, and M Meuth. 1994. "Mutator Phenotypes in Human Colorectal Carcinoma Cell Lines." *Proceedings of the National Academy of Sciences of the United States of America* 91 (14) (July 5): 6319–23. http://www.ncbi.nlm.nih.gov/pubmed/8022779.

Blatti, C., M. Kazemian, S. Wolfe, M. Brodsky, and S. Sinha. 2015. "Integrating Motif, DNA Accessibility and Gene Expression Data to Build Regulatory Maps in an Organism." *Nucleic Acids Research* 43 (8) (April 30): 3998–4012. doi:10.1093/nar/gkv195. http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv195.

Boland, C. Richard, and Ajay Goel. 2010. "Microsatellite Instability in Colorectal Cancer." *Gastroenterology* 138 (6) (June): 2073–2087.e3. doi:10.1053/j.gastro.2009.12.064. http://www.ncbi.nlm.nih.gov/pubmed/20420947.

Bonn, Stefan, Robert P Zinzen, Charles Girardot, E Hilary Gustafson, Alexis Perez-Gonzalez, Nicolas Delhomme, Yad Ghavi-Helm, Bartek Wilczyński, Andrew Riddell, and Eileen E M Furlong. 2012. "Tissue-Specific Analysis of Chromatin State Identifies Temporal Signatures of Enhancer Activity during Embryonic Development." *Nature Genetics* 44 (2) (January 8): 148–156. doi:10.1038/ng.1064. http://www.nature.com/doifinder/10.1038/ng.1064.

Borodovsky, Alexandra, Meghan J. Seltzer, and Gregory J. Riggins. 2012. "Altered Cancer Cell Metabolism in Gliomas with Mutant IDH1 or IDH2." *Current Opinion in Oncology* 24 (1) (January): 83–89. doi:10.1097/CCO.0b013e32834d816a. http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001622-201201000-00015.

Bralten, Linda B C, Nanne K Kloosterhof, Rutger Balvers, Andrea Sacchetti, Lariesa Lapre, Martine Lamfers, Sieger Leenstra, et al. 2011. "IDH1 R132H Decreases Proliferation of Glioma Cell Lines in Vitro and in Vivo." *Annals of Neurology* 69 (3) (March): 455–63. doi:10.1002/ana.22390. http://www.ncbi.nlm.nih.gov/pubmed/21446021.

Brash, D E, J A Rudolph, J A Simon, A Lin, G J McKenna, H P Baden, A J Halperin, and J Pontén. 1991. "A Role for Sunlight in Skin Cancer: UV-Induced p53 Mutations in Squamous Cell Carcinoma." *Proceedings of the National Academy of Sciences of the United States of America* 88 (22) (November 15): 10124–8. http://www.ncbi.nlm.nih.gov/pubmed/1946433.

Briasoulis, E, C Tolis, J Bergh, and N Pavlidis. 2005. "ESMO Minimum Clinical Recommendations for Diagnosis, Treatment and Follow-up of Cancers of Unknown Primary Site (CUP)." *Annals of Oncology : Official Journal of the European Society for Medical Oncology / ESMO* 16 Suppl 1: i75–i76. doi:10.1093/annonc/mdi804.

Brose, Marcia S, Patricia Volpe, Michael Feldman, Madhu Kumar, Irum Rishi, Renee Gerrero, Eugene Einhorn, et al. 2002. "BRAF and RAS Mutations in Human Lung Cancer and Melanoma." *Cancer Research* 62 (23) (December 1): 6997–7000. http://www.ncbi.nlm.nih.gov/pubmed/12460918.

Bunting, Samuel F, Elsa Callén, Nancy Wong, Hua-Tang Chen, Federica Polato, Amanda Gunn, Anne Bothmer, et al. 2010. “53BP1 Inhibits Homologous Recombination in Brca1-Deficient Cells by Blocking Resection of DNA Breaks.” *Cell* 141 (2) (April 16): 243–54. doi:10.1016/j.cell.2010.03.012. http://www.ncbi.nlm.nih.gov/pubmed/20362325.

Burger, Gerard, Ameen Abu-Hanna, Nicolette de Keizer, and Ronald Cornet. 2016. “Natural Language Processing in Pathology: A Scoping Review.” *Journal of Clinical Pathology* (July 22). doi:10.1136/jclinpath-2016-203872. http://www.ncbi.nlm.nih.gov/pubmed/27451435.

Burns, Michael B, Lela Lackey, Michael A Carpenter, Anurag Rathore, Allison M Land, Brandon Leonard, Eric W Refsland, et al. 2013. “APOBEC3B Is an Enzymatic Source of Mutation in Breast Cancer.” *Nature* 494 (7437) (February 21): 366–70. doi:10.1038/nature11881. http://www.ncbi.nlm.nih.gov/pubmed/23389445.

Cancer Genome Atlas Network. 2012. “Comprehensive Molecular Characterization of Human Colon and Rectal Cancer.” *Nature* 487 (7407) (July 19): 330–7. doi:10.1038/nature11252. http://www.ncbi.nlm.nih.gov/pubmed/22810696.

Cancer Genome Atlas Research Network, Cyriac Kandoth, Nikolaus Schultz, Andrew D Cherniack, Rehan Akbani, Yuexin Liu, Hui Shen, et al. 2013. “Integrated Genomic Characterization of Endometrial Carcinoma.” *Nature* 497 (7447) (May 2): 67–73. doi:10.1038/nature12113. http://www.ncbi.nlm.nih.gov/pubmed/23636398.

Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. 2013. “The Cancer Genome Atlas Pan-Cancer Analysis Project.” *Nature Genetics* 45 (10) (October): 1113–20. doi:10.1038/ng.2764. http://www.ncbi.nlm.nih.gov/pubmed/24071849.

CGA. 2013. “The Cancer Genome Analysis (CGA) Group, Broad Institute of Harvard and MIT.” https://www.broadinstitute.org/cancer/cga/Home.

Chapman, Michael A., Michael S. Lawrence, Jonathan J. Keats, Kristian Cibulskis, Carrie Sougnez, Anna C. Schinzel, Christina L. Harview, et al. 2011. “Initial Genome Sequencing and Analysis of Multiple Myeloma.” *Nature* 471 (7339) (March 24): 467–472. doi:10.1038/nature09837. http://www.nature.com/doifinder/10.1038/nature09837.

Check Hayden, Erika. 2014. “Is the $1,000 Genome for Real?” *Nature* (January 15). doi:10.1038/nature.2014.14530.
http://www.nature.com/doifinder/10.1038/nature.2014.14530.

Chen, Yen-Chen, Wan-Chi Ke, and Hung-Wen Chiu. 2014. “Risk Classification of Cancer Survival Using ANN with Gene Expression Data from Multiple Laboratories.” *Computers in Biology and Medicine* 48: 1–7. doi:10.1016/j.compbiomed.2014.02.006. http://www.ncbi.nlm.nih.gov/pubmed/24631783.

Cheng, Xiaoping, Hongmin Cai, Yue Zhang, Bo Xu, and Weifeng Su. 2015. “Optimal Combination of Feature Selection and Classification via Local Hyperplane Based Learning Strategy.” *BMC Bioinformatics* 16: 219. doi:10.1186/s12859-015-0629-6. http://www.ncbi.nlm.nih.gov/pubmed/26159165.

Cibulskis, Kristian, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe,

Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. 2013. “Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples.” *Nature Biotechnology* (February 10): 1–9. doi:10.1038/nbt.2514. http://www.ncbi.nlm.nih.gov/pubmed/23396013.

Cingolani, Pablo, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J. Land, Xiangyi Lu, and Douglas M. Ruden. 2012. “A Program for Annotating and Predicting the Effects of Single Nucleotide Polymorphisms, SnpEff: SNPs in the Genome of Drosophila Melanogaster Strain W 1118; Iso-2; Iso-3.” *Fly* 6 (2): 80–92. doi:10.4161/fly.19695.

Cock, Peter J A, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. 2010. “The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants.” *Nucleic Acids Research* 38 (6) (April): 1767–71. doi:10.1093/nar/gkp1137. http://www.ncbi.nlm.nih.gov/pubmed/20015970.

Cohen, S. J., C. J A Punt, N. Iannotti, B. H. Saidman, K. D. Sabbath, N. Y. Gabrail, J. Picus, et al. 2009. “Prognostic Significance of Circulating Tumor Cells in Patients with Metastatic Colorectal Cancer.” *Annals of Oncology* 20 (7): 1223–1229. doi:10.1093/annonc/mdn786.

Cotney, J., J. Leng, S. Oh, L. E. DeMare, S. K. Reilly, M. B. Gerstein, and J. P. Noonan. 2012. “Chromatin State Signatures Associated with Tissue-Specific Gene Expression and Enhancer Activity in the Embryonic Limb.” *Genome Research* 22 (6) (June 1): 1069–1080. doi:10.1101/gr.129817.111. http://genome.cshlp.org/cgi/doi/10.1101/gr.129817.111.

Cristofanilli, Massimo. 2006. “Circulating Tumor Cells, Disease Progression, and Survival in Metastatic Breast Cancer.” *Seminars in Oncology* 33 (SUPPL. 9): 9–14. doi:10.1053/j.seminoncol.2006.03.016.

D., Green, and Swets J. 1966. *Signal Detection Theory and Psychophysics*. New York: J. Wiley.

Davies, A A, J Y Masson, M J McIlwraith, A Z Stasiak, A Stasiak, A R Venkitaraman, and S C West. 2001. “Role of BRCA2 in Control of the RAD51 Recombination and DNA Repair Protein.” *Molecular Cell* 7 (2) (February): 273–82. http://www.ncbi.nlm.nih.gov/pubmed/11239456.

De Bono, Johann S., Howard I. Scher, R. Bruce Montgomery, Christopher Parker, M. Craig Miller, Henk Tissing, Gerald V. Doyle, Leon W W M Terstappen, Kenneth J. Pienta, and Derek Raghavan. 2008. “Circulating Tumor Cells Predict Survival Benefit from Treatment in Metastatic Castration-Resistant Prostate Cancer.” *Clinical Cancer Research* 14 (19): 6302–6309. doi:10.1158/1078-0432.CCR-08-0872.

Demirci, Ferhat, Pinar Akan, Tuncay Kume, Ali Riza Sisman, Zubeyde Erbayraktar, and Suleyman Sevinc. 2016. “Artificial Neural Network Approach in Laboratory Test Reporting: Learning Algorithms.” *American Journal of Clinical Pathology* 146 (2) (August): 227–37. doi:10.1093/ajcp/aqw104. http://www.ncbi.nlm.nih.gov/pubmed/27473741.

DePristo, Mark a, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, et al. 2011. “A Framework for Variation Discovery and Genotyping Using next-Generation DNA Sequencing Data.” *Nature Genetics* 43 (5) (May):

491–8. doi:10.1038/ng.806. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3083463&tool=pmcentrez&rendertype=abstract.

Drier, Yotam, Michael S Lawrence, Scott L Carter, Chip Stewart, Stacey B Gabriel, Eric S Lander, Matthew Meyerson, Rameen Beroukhim, and Gad Getz. 2013. "Somatic Rearrangements across Cancer Reveal Classes of Samples with Distinct Patterns of DNA Breakage and Rearrangement-Induced Hypermutability." *Genome Research* 23 (2) (February): 228–35. doi:10.1101/gr.141382.112. http://www.ncbi.nlm.nih.gov/pubmed/23124520.

Drobetsky, E A, A J Grosovsky, and B W Glickman. 1987. "The Specificity of UV-Induced Mutations at an Endogenous Locus in Mammalian Cells." *Proceedings of the National Academy of Sciences of the United States of America* 84 (24) (December): 9103–7. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=299700&tool=pmcentrez&rendertype=abstract.

Durbeej, Bo, and Leif A Eriksson. 2003. "On the Formation of Cyclobutane Pyrimidine Dimers in UV-Irradiated DNA: Why Are Thymines More Reactive?" *Photochemistry and Photobiology* 78 (2) (August): 159–67. http://www.ncbi.nlm.nih.gov/pubmed/12945584.

Ellegren, Hans, Nick G C Smith, and Matthew T Webster. 2003. "Mutation Rate Variation in the Mammalian Genome." *Current Opinion in Genetics & Development* 13 (6) (December): 562–8. http://www.ncbi.nlm.nih.gov/pubmed/14638315.

Ewing, B, L Hillier, M C Wendl, and P Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment." *Genome Research* 8 (3) (March): 175–85. http://www.ncbi.nlm.nih.gov/pubmed/9521921.

Ewing, Brent, and Phil Green. 1998. "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities." *Genome Research* 8 (3) (March 1): 186–94. doi:10.1101/gr.8.3.186. http://www.ncbi.nlm.nih.gov/pubmed/9521922.

Exarchos, Konstantinos P., Yorgos Goletsis, and Dimitrios I. Fotiadis. 2012. "Multiparametric Decision Support System for the Prediction of Oral Cancer Reoccurrence." *IEEE Transactions on Information Technology in Biomedicine* 16 (6): 1127–1134. doi:10.1109/TITB.2011.2165076.

Ferlay, J, I Soerjomataram I, R Dikshit, S Eser, C Mathers, M Rebelo, D M Parkin, D Forman D, and F Bray. 2014. "Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012." *International Journal of Cancer. Journal International Du Cancer* (September 13). doi:10.1002/ijc.29210. http://www.ncbi.nlm.nih.gov/pubmed/25220842.

Forbes, Simon A, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, et al. 2015. "COSMIC: Exploring the World's Knowledge of Somatic Mutations in Human Cancer." *Nucleic Acids Research* 43 (Database issue): D805–11. doi:10.1093/nar/gku1075. http://www.ncbi.nlm.nih.gov/pubmed/25355519.

Forbes, Simon a, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, et al. 2011. "COSMIC: Mining Complete Cancer Genomes in the Catalogue

of Somatic Mutations in Cancer." *Nucleic Acids Research* 39 (Database issue) (January): D945–50. doi:10.1093/nar/gkq929. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013785&tool=pmcentrez&rendertype=abstract.

Foster, Kenneth R, Robert Koprowski, and Joseph D Skufca. 2014. "Machine Learning, Medical Diagnosis, and Biomedical Engineering Research - Commentary." *BioMedical Engineering OnLine* 13 (1): 94. doi:10.1186/1475-925X-13-94. http://biomedical-engineering-online.biomedcentral.com/articles/10.1186/1475-925X-13-94.

Gaidzik, Verena I, Lars Bullinger, Richard F Schlenk, Andreas S Zimmermann, Jürgen Röck, Peter Paschka, Andrea Corbacioglu, et al. 2011. "RUNX1 Mutations in Acute Myeloid Leukemia: Results from a Comprehensive Genetic and Clinical Analysis from the AML Study Group." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 29 (10) (April 1): 1364–72. doi:10.1200/JCO.2010.30.7926. http://www.ncbi.nlm.nih.gov/pubmed/21343560.

Goh, Liang, Geng Bo Chen, Ioana Cutcutache, Benjamin Low, Bin Tean Teh, Steve Rozen, and Patrick Tan. 2011. "Assessing Matched Normal and Tumor Pairs in next-Generation Sequencing Studies." *PloS One* 6 (3) (January): e17810. doi:10.1371/journal.pone.0017810. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3060821&tool=pmcentrez&rendertype=abstract.

Govindan, Ramaswamy, Li Ding, Malachi Griffith, Janakiraman Subramanian, Nathan D Dees, Krishna L Kanchi, Christopher A Maher, et al. 2012. "Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers." *Cell* 150 (6) (September 14): 1121–34. doi:10.1016/j.cell.2012.08.024. http://www.ncbi.nlm.nih.gov/pubmed/22980976.

Greco, F A, W K Vaughn, and J D Hainsworth. 1986. "Advanced Poorly Differentiated Carcinoma of Unknown Primary Site: Recognition of a Treatable Syndrome." *Annals of Internal Medicine* 104 (4) (April): 547–53. http://www.ncbi.nlm.nih.gov/pubmed/3006571.

Greco, F. A., D. R. Spigel, D. A. Yardley, M. G. Erlander, X.-J. Ma, and J. D. Hainsworth. 2010. "Molecular Profiling in Unknown Primary Cancer: Accuracy of Tissue of Origin Prediction." *The Oncologist* 15 (5) (May 1): 500–506. doi:10.1634/theoncologist.2009-0328. http://theoncologist.alphamedpress.org/cgi/doi/10.1634/theoncologist.2009-0328.

Greenman, Christopher, Philip Stephens, Raffaella Smith, Gillian L Dalgliesh, Christopher Hunter, Graham Bignell, Helen Davies, et al. 2007. "Patterns of Somatic Mutation in Human Cancer Genomes." *Nature* 446 (7132) (March 8): 153–8. doi:10.1038/nature05610. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2712719&tool=pmcentrez&rendertype=abstract.

Guo, Guangwu, Xiaojuan Sun, Chao Chen, Song Wu, Peide Huang, Zesong Li, Michael Dean, et al. 2013. "Whole-Genome and Whole-Exome Sequencing of Bladder Cancer Identifies Frequent Alterations in Genes Involved in Sister Chromatid Cohesion and Segregation." *Nature Genetics* 45 (12): 1459–63. doi:10.1038/ng.2798. http://www.ncbi.nlm.nih.gov/pubmed/24121792.

Habibi, Narjeskhatoon, Siti Z Mohd Hashim, Alireza Norouzi, and Mohammed Samian. 2014. "A Review of Machine Learning Methods to Predict the Solubility of Overexpressed

Recombinant Proteins in Escherichia Coli." *BMC Bioinformatics* 15 (1): 134. doi:10.1186/1471-2105-15-134. http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-134.

Haiman, Christopher A., Daniel O. Stram, Lynne R. Wilkens, Malcolm C. Pike, Laurence N. Kolonel, Brian E. Henderson, and Loïc Le Marchand. 2006. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." *New England Journal of Medicine* 354 (4) (January 26): 333–342. doi:10.1056/NEJMoa033250. http://www.nejm.org/doi/abs/10.1056/NEJMoa033250.

Hainaut, P, M Olivier, and G P Pfeifer. 2001. "TP53 Mutation Spectrum in Lung Cancers and Mutagenic Signature of Components of Tobacco Smoke: Lessons from the IARC TP53 Mutation Database." *Mutagenesis* 16 (6) (November): 551–3; author reply 555–6. http://www.ncbi.nlm.nih.gov/pubmed/11682648.

Hainaut, P, and G P Pfeifer. 2001. "Patterns of p53 G-->T Transversions in Lung Cancers Reflect the Primary Mutagenic Signature of DNA-Damage by Tobacco Smoke." *Carcinogenesis* 22 (3) (March): 367–74. http://www.ncbi.nlm.nih.gov/pubmed/11238174.

Hammerman, Peter S, D Neil Hayes, Matthew D Wilkerson, Nikolaus Schultz, Ron Bose, Andy Chu, Eric a Collisson, et al. 2012. "Comprehensive Genomic Characterization of Squamous Cell Lung Cancers." *Nature* 489 (7417) (September 27): 519–25. doi:10.1038/nature11404. http://www.ncbi.nlm.nih.gov/pubmed/22960745.

Hanley, J A, and B J McNeil. 1982. "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology* 143 (1) (April): 29–36. doi:10.1148/radiology.143.1.7063747. http://pubs.rsna.org/doi/abs/10.1148/radiology.143.1.7063747.

Hao, Yujun, Yardena Samuels, Qingling Li, Dawid Krokowski, Bo-Jhih Guan, Chao Wang, Zhicheng Jin, et al. 2016. "Oncogenic PIK3CA Mutations Reprogram Glutamine Metabolism in Colorectal Cancer." *Nature Communications* 7: 11971. doi:10.1038/ncomms11971. http://www.ncbi.nlm.nih.gov/pubmed/27321283.

Hayes, Daniel F., Massimo Cristofanilli, G. Thomas Budd, Matthew J. Ellis, Alison Stopeck, M. Craig Miller, Jeri Matera, W. Jeffrey Allard, Gerald V. Doyle, and Leon W W M Terstappen. 2006. "Circulating Tumor Cells at Each Follow-up Time Point during Therapy of Metastatic Breast Cancer Patients Predict Progression-Free and Overall Survival." *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research* 12 (14 Pt 1): 4218–4224. doi:10.1158/1078-0432.CCR-05-2821.

He, Pengcheng. 2013. "Prognostic Significance of NPM1 Mutations in Acute Myeloid Leukemia: A Meta-Analysis." *Molecular and Clinical Oncology* (December 10). doi:10.3892/mco.2013.222. http://www.spandidos-publications.com/10.3892/mco.2013.222.

Heffernan, Rhys, Kuldip Paliwal, James Lyons, Abdollah Dehzangi, Alok Sharma, Jihua Wang, Abdul Sattar, Yuedong Yang, and Yaoqi Zhou. 2015. "Improving Prediction of Secondary Structure, Local Backbone Angles, and Solvent Accessible Surface Area of Proteins by Iterative Deep Learning." *Scientific Reports* 5: 11476. doi:10.1038/srep11476. http://www.ncbi.nlm.nih.gov/pubmed/26098304.

Heinz, Sven, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. 2010. "Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities." *Molecular Cell* 38 (4): 576–589. doi:10.1016/j.molcel.2010.05.004.

Helleday, Thomas, Saeed Eshtad, and Serena Nik-Zainal. 2014. "Mechanisms Underlying Mutational Signatures in Human Cancers." *Nature Reviews. Genetics* 15 (9) (September): 585–98. doi:10.1038/nrg3729. http://www.ncbi.nlm.nih.gov/pubmed/24981601.

Hizukuri, Yoshiyuki, Ryusuke Sawada, and Yoshihiro Yamanishi. 2015. "Predicting Target Proteins for Drug Candidate Compounds Based on Drug-Induced Gene Expression Data in a Chemical Structure-Independent Manner." *BMC Medical Genomics* 8: 82. doi:10.1186/s12920-015-0158-1. http://www.ncbi.nlm.nih.gov/pubmed/26684652.

Hoang, Margaret L, Chung-Hsin Chen, Viktoriya S Sidorenko, Jian He, Kathleen G Dickman, Byeong Hwa Yun, Masaaki Moriya, et al. 2013. "Mutational Signature of Aristolochic Acid Exposure as Revealed by Whole-Exome Sequencing." *Science Translational Medicine* 5 (197) (August 7): 197ra102. doi:10.1126/scitranslmed.3006200. http://www.ncbi.nlm.nih.gov/pubmed/23926200.

Hodis, Eran, Ian R Watson, Gregory V Kryukov, Stefan T Arold, Marcin Imielinski, Jean-philippe Theurillat, Elizabeth Nickerson, et al. 2012. "A Landscape of Driver Mutations in Melanoma." *Cell* 150 (2) (July 20): 251–63. doi:10.1016/j.cell.2012.06.024. http://www.ncbi.nlm.nih.gov/pubmed/22817889.

Homer, Nils, Barry Merriman, and Stanley F. Nelson. 2009. "BFAST: An Alignment Tool for Large Scale Genome Resequencing." Edited by Chad Creighton. *PLoS ONE* 4 (11) (November 11): e7767. doi:10.1371/journal.pone.0007767. http://dx.plos.org/10.1371/journal.pone.0007767.

Hou, Jian-Mei, Alastair Greystoke, Lee Lancashire, Jeff Cummings, Tim Ward, Ruth Board, Eitan Amir, et al. 2009. "Evaluation of Circulating Tumor Cells and Serological Cell Death Biomarkers in Small Cell Lung Cancer Patients Undergoing Chemotherapy." *The American Journal of Pathology* 175 (2): 808–816. doi:10.2353/ajpath.2009.090078.

Hu, Bing, Nassim El Hajj, Scott Sittler, Nancy Lammert, Robert Barnes, and Aurelia Meloni-Ehrig. 2012. "Gastric Cancer: Classification, Histology and Application of Molecular Pathology." *Journal of Gastrointestinal Oncology* 3 (3) (September): 251–61. doi:10.3978/j.issn.2078-6891.2012.021. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3418539&tool=pmcentrez&rendertype=abstract.

Huang, Kie Kyon, Kang Won Jang, Sangwoo Kim, Han Sang Kim, Sung-Moo Kim, Hyeong Ju Kwon, Hye Ryun Kim, et al. 2016. "Exome Sequencing Reveals Recurrent REV3L Mutations in Cisplatin-Resistant Squamous Cell Carcinoma of Head and Neck." *Scientific Reports* 6: 19552. doi:10.1038/srep19552. http://www.ncbi.nlm.nih.gov/pubmed/26790612.

Hunter, Chris, Raffaella Smith, Daniel P Cahill, Philip Stephens, Claire Stevens, Jon Teague, Chris Greenman, et al. 2006. "A Hypermutation Phenotype and Somatic MSH6 Mutations in Recurrent Human Malignant Gliomas after Alkylator Chemotherapy." *Cancer Research* 66

(8) (April 15): 3987–91. doi:10.1158/0008-5472.CAN-06-0127. http://www.ncbi.nlm.nih.gov/pubmed/16618716.

Hurd, Paul J, and Christopher J Nelson. 2009. "Advantages of next-Generation Sequencing versus the Microarray in Epigenetic Research." *Briefings in Functional Genomics & Proteomics* 8 (3) (May): 174–83. doi:10.1093/bfgp/elp013. http://www.ncbi.nlm.nih.gov/pubmed/19535508.

IARC. 2015. "Globocan Website." http://globocan.iarc.fr/Default.aspx.

Iggo, R, K Gatter, J Bartek, D Lane, and A L Harris. 1990. "Increased Expression of Mutant Forms of p53 Oncogene in Primary Lung Cancer." *Lancet* 335: 675–679. doi:0140-6736(90)90801-B [pii].

Imielinski, Marcin, Alice H Berger, Peter S Hammerman, Bryan Hernandez, Trevor J Pugh, Eran Hodis, Jeonghee Cho, et al. 2012. "Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing." *Cell* 150 (6) (September 14): 1107–20. doi:10.1016/j.cell.2012.08.029. http://www.ncbi.nlm.nih.gov/pubmed/22980975.

International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, et al. 2010. "International Network of Cancer Genome Projects." *Nature* 464 (7291) (April 15): 993–8. doi:10.1038/nature08987. http://www.ncbi.nlm.nih.gov/pubmed/20393554.

Jia, Wei-Hua, Ben Zhang, Keitaro Matsuo, Aesun Shin, Yong-Bing Xiang, Sun Ha Jee, Dong-Hyun Kim, et al. 2012. "Genome-Wide Association Analyses in East Asians Identify New Susceptibility Loci for Colorectal Cancer." *Nature Genetics* 45 (2) (December 23): 191–196. doi:10.1038/ng.2505. http://www.nature.com/doifinder/10.1038/ng.2505.

Kandoth, Cyriac, Michael D. McLellan, Fabio Vandin, Kai Ye, Beifang Niu, Charles Lu, Mingchao Xie, et al. 2013. "Mutational Landscape and Significance across 12 Major Cancer Types." *Nature* 502 (7471) (October 17): 333–9. doi:10.1038/nature12634. http://www.nature.com/doifinder/10.1038/nature12634.

Kang, John, Russell Schwartz, John Flickinger, and Sushil Beriwal. 2015. "Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective." *International Journal of Radiation Oncology, Biology, Physics* 93 (5) (December 1): 1127–35. doi:10.1016/j.ijrobp.2015.07.2286. http://www.ncbi.nlm.nih.gov/pubmed/26581149.

Karran, P. 1996. "Microsatellite Instability and DNA Mismatch Repair in Human Cancer." *Seminars in Cancer Biology* 7 (1) (February): 15–24. doi:10.1006/scbi.1996.0003. http://www.ncbi.nlm.nih.gov/pubmed/8695762.

Keller, Irene, Douda Bensasson, and Richard A. Nichols. 2007. "Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes." *PLoS Genetics* 3 (2): e22. doi:10.1371/journal.pgen.0030022. http://dx.plos.org/10.1371/journal.pgen.0030022.

Kellis, M., B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, et al. 2014. "Defining Functional DNA Elements in the Human Genome." *Proceedings of the National Academy of Sciences* 111 (17) (April 29): 6131–6138. doi:10.1073/pnas.1318948111. http://www.pnas.org/cgi/doi/10.1073/pnas.1318948111.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and a. D. Haussler. 2002. “The Human Genome Browser at UCSC.” *Genome Research* 12 (6) (May 16): 996–1006. doi:10.1101/gr.229102. http://www.genome.org/cgi/doi/10.1101/gr.229102.

Kerkhofs, Thomas M A, M Hester T Ettaieb, Ilse G C Hermsen, and Harm R Haak. 2015. “Developing Treatment for Adrenocortical Carcinoma.” *Endocrine-Related Cancer*. doi:10.1530/ERC-15-0318. http://www.ncbi.nlm.nih.gov/pubmed/26259571.

Khan, S. A., B. R. Davidson, R. D. Goldin, N. Heaton, J. Karani, S. P. Pereira, W. M. C. Rosenberg, et al. 2012. “Guidelines for the Diagnosis and Treatment of Cholangiocarcinoma: An Update.” *Gut* 61 (12) (December 1): 1657–1669. doi:10.1136/gutjnl-2011-301748. http://gut.bmj.com/cgi/doi/10.1136/gutjnl-2011-301748.

Kim, Juhyeon, and Hyunjung Shin. 2013. “Breast Cancer Survivability Prediction Using Labeled, Unlabeled, and Pseudo-Labeled Patient Data.” *Journal of the American Medical Informatics Association* 20 (4): 613–618. doi:10.1136/amiajnl-2012-001570. http://jamia.bmj.com/content/20/4/613\nhttp://jamia.bmj.com/content/20/4/613.short?g=w_jamia_current_tab\nhttp://www.ncbi.nlm.nih.gov/pubmed/23467471.

Kim, Woojae, Ku Sang Kim, Jeong Eon Lee, Dong-Young Noh, Sung-Won Kim, Yong Sik Jung, Man Young Park, and Rae Woong Park. 2012. “Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine.” *Journal of Breast Cancer*. doi:10.4048/jbc.2012.15.2.230.

Kimelman, D., and C. Bjornson. 2004. “Vertebrate Mesoderm Induction: From Frogs to Mice.” In *Gastrulation: From Cells to Embryo*, edited by Claudio D. Stern, 363. CSHL Press.

Kirsch-Volders, M., S. Bonassi, Z. Herceg, A. Hirvonen, L. Möller, and D. H. Phillips. 2010. “Gender-Related Differences in Response to Mutagens and Carcinogens.” *Mutagenesis*. doi:10.1093/mutage/geq008.

Koboldt, Daniel C, Qunyuan Zhang, David E Larson, Dong Shen, Michael D McLellan, Ling Lin, Christopher a Miller, Elaine R Mardis, Li Ding, and Richard K Wilson. 2012. “VarScan 2: Somatic Mutation and Copy Number Alteration Discovery in Cancer by Exome Sequencing.” *Genome Research* 22 (3) (March): 568–76. doi:10.1101/gr.129684.111. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3290792&tool=pmcentrez&rendertype=abstract.

Korde, L. A., J. A. Zujewski, L. Kamin, S. Giordano, S. Domchek, W. F. Anderson, J. M. S. Bartlett, et al. 2010. “Multidisciplinary Meeting on Male Breast Cancer: Summary and Research Recommendations.” *Journal of Clinical Oncology* 28 (12) (April 20): 2114–2122. doi:10.1200/JCO.2009.25.5729. http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2009.25.5729.

Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. “Machine Learning Applications in Cancer Prognosis and Prediction.” *Computational and Structural Biotechnology Journal* 13: 8–17. doi:10.1016/j.csbj.2014.11.005.
http://linkinghub.elsevier.com/retrieve/pii/S2001037014000464.

Krebs, M. G., J.-M. Hou, T. H. Ward, F. H. Blackhall, and C. Dive. 2010. “Circulating Tumour Cells: Their Utility in Cancer Management and Predicting Outcomes.” *Therapeutic*

Advances in Medical Oncology 2 (6) (November 1): 351–365. doi:10.1177/1758834010378414. http://tam.sagepub.com/cgi/doi/10.1177/1758834010378414.

Kuipers, E J, and P Sipponen. 2006. "Helicobacter Pylori Eradication for the Prevention of Gastric Cancer." *Helicobacter* 11 Suppl 1: 52–57. doi:10.1111/j.1478-405X.2006.00425.x.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, et al. 2015. "Integrative Analysis of 111 Reference Human Epigenomes." *Nature* 518 (7539) (February 18): 317–330. doi:10.1038/nature14248. http://www.nature.com/doifinder/10.1038/nature14248.

Kuraguchi, M, W Edelmann, K Yang, M Lipkin, R Kucherlapati, and A M Brown. 2000. "Tumor-Associated Apc Mutations in Mlh1-/- Apc1638N Mice Reveal a Mutational Signature of Mlh1 Deficiency." *Oncogene* 19 (50) (November 23): 5755–63. doi:10.1038/sj.onc.1203962. http://www.ncbi.nlm.nih.gov/pubmed/11126362.

Landrum, Melissa J, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2016. "ClinVar: Public Archive of Interpretations of Clinically Relevant Variants." *Nucleic Acids Research* 44 (D1) (January 4): D862–8. doi:10.1093/nar/gkv1222. http://www.ncbi.nlm.nih.gov/pubmed/26582918.

Landrum, Melissa J, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M Church, and Donna R Maglott. 2014. "ClinVar: Public Archive of Relationships among Sequence Variation and Human Phenotype." *Nucleic Acids Research* 42 (Database issue) (January): D980–5. doi:10.1093/nar/gkt1113. http://www.ncbi.nlm.nih.gov/pubmed/24234437.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Lawrence, Michael S, Petar Stojanov, Paz Polak, Gregory V Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L Carter, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457) (July 11): 214–8. doi:10.1038/nature12213. http://www.ncbi.nlm.nih.gov/pubmed/23770567.

Levis, M. 2011. "FLT3/ITD AML and the Law of Unintended Consequences." *Blood* 117 (26) (June 30): 6987–6990. doi:10.1182/blood-2011-03-340273. http://www.bloodjournal.org/cgi/doi/10.1182/blood-2011-03-340273.

Ley, Timothy J., Li Ding, Matthew J. Walter, Michael D. McLellan, Tamara Lamprecht, David E. Larson, Cyriac Kandoth, et al. 2010. "DNMT3A Mutations in Acute Myeloid Leukemia." *New England Journal of Medicine* 363 (25) (December 16): 2424–2433. doi:10.1056/NEJMoa1005143. http://www.nejm.org/doi/abs/10.1056/NEJMoa1005143.

Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics (Oxford, England)* 25 (14) (July 15): 1754–60. doi:10.1093/bioinformatics/btp324. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2705234&tool=pmcentrez&rendertype=abstract.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16) (August 15): 2078–9. doi:10.1093/bioinformatics/btp352. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723002&tool=pmcentrez&rendertype=abstract.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11): 1851–1858. doi:10.1101/gr.078212.108.

Li, Min, Christian Marin-Muller, Uddalak Bharadwaj, Kwong-Hon Chow, Qizhi Yao, and Changyi Chen. 2009. “MicroRNAs: Control and Loss of Control in Human Physiology and Disease.” *World Journal of Surgery* 33 (4) (April): 667–684. doi:10.1007/s00268-008-9836-x. http://link.springer.com/10.1007/s00268-008-9836-x.

Liu, Chenbin, Francis Tsow, Yi Zou, and Nongjian Tao. 2016. “Particle Pollution Estimation Based on Image Analysis.” *PloS One* 11 (2): e0145955. doi:10.1371/journal.pone.0145955. http://www.ncbi.nlm.nih.gov/pubmed/26828757.

Liu, Pengyuan, Carl Morrison, Liang Wang, Donghai Xiong, Peter Vedell, Peng Cui, Xing Hua, et al. 2012. “Identification of Somatic Mutations in Non-Small Cell Lung Carcinomas Using Whole-Exome Sequencing.” *Carcinogenesis* 33 (7): 1270–1276. doi:10.1093/carcin/bgs148.

Loeb, Lawrence A., and Curtis C. Harris. 2008. “Advances in Chemical Carcinogenesis: A Historical Review and Prospective.” *Cancer Research*. doi:10.1158/0008-5472.CAN-08-2852.

Löffler, Harald, Nicole Pfarr, Mark Kriegsmann, Volker Endris, Thomas Hielscher, Philipp Lohneis, Gunnar Folprecht, et al. 2016. “Molecular Driver Alterations and Their Clinical Relevance in Cancer of Unknown Primary Site.” *Oncotarget* (June 14). doi:10.18632/oncotarget.10035. http://www.ncbi.nlm.nih.gov/pubmed/27322425.

Long, Fei, Jia-Hang Su, Bin Liang, Li-Li Su, and Shu-Juan Jiang. 2015. “Identification of Gene Biomarkers for Distinguishing Small-Cell Lung Cancer from Non-Small-Cell Lung Cancer Using a Network-Based Approach.” *BioMed Research International* 2015: 685303. doi:10.1155/2015/685303. http://www.ncbi.nlm.nih.gov/pubmed/26290870.

Lotta, Luca A, Ali Abbasi, Stephen J Sharp, Anna-Stina Sahlqvist, Dawn Waterworth, Julia M Brosnan, Robert A Scott, Claudia Langenberg, and Nicholas J Wareham. 2015. “Definitions of Metabolic Health and Risk of Future Type 2 Diabetes in BMI Categories: A Systematic Review and Network Meta-Analysis.” *Diabetes Care* 38 (11) (November): 2177–87. doi:10.2337/dc15-1218.

Lynch, Thomas J T.J., D.W. Daphne W Bell, Raffaella Sordella, Sarada Gurubhagavatula, Ross A R.A. Okimoto, B.W. Brian W Brannigan, P.L. Patricia L Harris, et al. 2004. “Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib.” *The New England Journal of Medicine* 350 (21) (May 20): 2129–39. doi:10.1056/NEJMoa040938. http://www.nejm.org/doi/full/10.1056/NEJMoa040938.

Manning, G, DB B Whyte, R Martinez, T Hunter, and S Sudarsanam. 2002. “The Protein Kinase Complement of the Human Genome.” *Science (New York, N.Y.)* 298 (5600) (December 6): 1912–34. doi:10.1126/science.1075762. http://stke.sciencemag.org/cgi/content/abstract/sci;298/5600/1912.

Martínez-Alcántara, A., E. Ballesteros, C. Feng, M. Rojas, H. Koshinsky, V. Y. Fofanov, P. Havlak, and Y. Fofanov. 2009. “PIQA: Pipeline for Illumina G1 Genome Analyzer Data Quality Assessment.” *Bioinformatics* 25 (18): 2438–2439. doi:10.1093/bioinformatics/btp429.

Mason, S. J., and N. E. Graham. 2002. “Areas beneath the Relative Operating Characteristics (ROC) and Relative Operating Levels (ROL) Curves: Statistical Significance and Interpretation.” *Quarterly Journal of the Royal Meteorological Society* 128 (584) (July 15): 2145–2166. doi:10.1256/003590002320603584. http://doi.wiley.com/10.1256/003590002320603584.

Massard, Christophe, Yohann Loriot, and Karim Fizazi. 2011. “Carcinomas of an Unknown Primary Origin—diagnosis and Treatment.” *Nature Reviews Clinical Oncology* 8 (12) (November 1): 701–710. doi:10.1038/nrclinonc.2011.158. http://www.nature.com/doifinder/10.1038/nrclinonc.2011.158.

McCann, J. 2000. “Gender Differences in Cancer That Don’t Make Sense--Or Do They?” *Journal of the National Cancer Institute* 92 (19) (October 4): 1560–1562. doi:10.1093/jnci/92.19.1560. http://jnci.oxfordjournals.org/cgi/doi/10.1093/jnci/92.19.1560.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data.” *Genome Research* 20 (9) (September): 1297–303. doi:10.1101/gr.107524.110. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract.

McLaren, William, Bethan Pritchard, Daniel Rios, Yuan Chen, Paul Flicek, and Fiona Cunningham. 2010. “Deriving the Consequences of Genomic Variants with the Ensembl API and SNP Effect Predictor.” *Bioinformatics* 26 (16) (August 15): 2069–2070. doi:10.1093/bioinformatics/btq330. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2916720&tool=pmcentrez&rendertype=abstract.

Mertes, Florian, Abdou Elsharawy, Sascha Sauer, Joop M L M van Helvoort, P J van der Zaag, Andre Franke, Mats Nilsson, Hans Lehrach, and Anthony J Brookes. 2011. “Targeted Enrichment of Genomic DNA Regions for next-Generation Sequencing.” *Briefings in Functional Genomics* 10 (6) (November 26): 374–86. doi:10.1093/bfgp/elr033. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245553&tool=pmcentrez&rendertype=abstract.

Meyer, Clifford A., and X. Shirley Liu. 2014. “Identifying and Mitigating Bias in next-Generation Sequencing Methods for Chromatin Biology.” *Nature Reviews Genetics* 15 (11) (September 16): 709–721. doi:10.1038/nrg3788. http://www.nature.com/doifinder/10.1038/nrg3788.

Meyerson, Matthew, Stacey Gabriel, and Gad Getz. 2010. “Advances in Understanding Cancer Genomes through Second-Generation Sequencing.” *Nature Reviews. Genetics* 11 (10) (October): 685–96. doi:10.1038/nrg2841. http://www.ncbi.nlm.nih.gov/pubmed/20847746.

Mills, Ryan E., Christopher T. Luttig, Christine E. Larkins, Adam Beauchamp, Circe Tsui, W. Stephen Pittard, and Scott E. Devine. 2006. “An Initial Map of Insertion and Deletion (INDEL) Variation in the Human Genome.” *Genome Research* 16 (9) (September): 1182–1190. doi:10.1101/gr.4565806. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1557762&tool=pmcentrez&rendertype=abstract.

Mousavizadegan, Maryam, and Hassan Mohabatkar. 2016. “An Evaluation on Different Machine Learning Algorithms for Classification and Prediction of Antifungal Peptides.” *Medicinal Chemistry (Shariqah (United Arab Emirates))* (February 29). http://www.ncbi.nlm.nih.gov/pubmed/26924627.

Moynahan, M E, J W Chiu, B H Koller, and M Jasin. 1999. “Brca1 Controls Homology-Directed DNA Repair.” *Molecular Cell* 4 (4) (October): 511–8. http://www.ncbi.nlm.nih.gov/pubmed/10549283.

Moynahan, M E, A J Pierce, and M Jasin. 2001. “BRCA2 Is Required for Homology-Directed Repair of Chromosomal Breaks.” *Molecular Cell* 7 (2) (February): 263–72. http://www.ncbi.nlm.nih.gov/pubmed/11239455.

Muzny, Donna M., Matthew N. Bainbridge, Kyle Chang, Huyen H. Dinh, Jennifer a. Drummond, Gerald Fowler, Christie L. Kovar, et al. 2012. “Comprehensive Molecular Characterization of Human Colon and Rectal Cancer.” *Nature* 487 (7407) (July 19): 330–7. doi:10.1038/nature11252. http://www.ncbi.nlm.nih.gov/pubmed/22810696.

Napolitano, Giulio, Adele Marshall, Peter Hamilton, and Anna T Gavin. 2016. “Machine Learning Classification of Surgical Pathology Reports and Chunk Recognition for Information Extraction Noise Reduction.” *Artificial Intelligence in Medicine* 70 (June): 77–83. doi:10.1016/j.artmed.2016.06.001. http://www.ncbi.nlm.nih.gov/pubmed/27431038.

NCBI. 2015. “dbVar: Database of Genomic Structural Variation.” http://www.ncbi.nlm.nih.gov/dbvar.

Ng, Andrea K, and Lois B Travis. “Subsequent Malignant Neoplasms in Cancer Survivors.” *Cancer Journal (Sudbury, Mass.)* 14 (6): 429–34. doi:10.1097/PPO.0b013e31818d8779. http://www.ncbi.nlm.nih.gov/pubmed/19060610.

Ng, Sarah B, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, et al. 2009. “Targeted Capture and Massively Parallel Sequencing of 12 Human Exomes.” *Nature* 461 (7261) (September 10): 272–6. doi:10.1038/nature08250. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2844771&tool=pmcentrez&rendertype=abstract.

Ni, Xiaohui, Minglei Zhuo, Zhe Su, Jianchun Duan, Yan Gao, Zhijie Wang, Chenghang Zong, et al. 2013. “Reproducible Copy Number Variation Patterns among Single Circulating Tumor Cells of Lung Cancer Patients.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (52): 21083–8. doi:10.1073/pnas.1320659110.

http://www.ncbi.nlm.nih.gov/pubmed/24324171.

Nik-Zainal, Serena, Ludmil B Alexandrov, David C Wedge, Peter Van Loo, Christopher D Greenman, Keiran Raine, David Jones, et al. 2012. “Mutational Processes Molding the Genomes of 21 Breast Cancers.” *Cell* 149 (5) (May 25): 979–93. doi:10.1016/j.cell.2012.04.024. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3414841&tool=pmcentrez&rendertype=abstract.

Nik-Zainal, Serena, Peter Van Loo, David C Wedge, Ludmil B Alexandrov, Christopher D Greenman, King Wai Lau, Keiran Raine, et al. 2012. “The Life History of 21 Breast Cancers.” *Cell* 149 (5) (May 25): 994–1007. doi:10.1016/j.cell.2012.04.023. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3428864&tool=pmcentrez&rendertype=abstract.

Oien, K. A., and J. L. Dennis. 2012. “Diagnostic Work-up of Carcinoma of Unknown Primary: From Immunohistochemistry to Molecular Profiling.” *Annals of Oncology* 23 (SUPPL. 10). doi:10.1093/annonc/mds357.

Olivier, Magali, Monica Hollstein, and Pierre Hainaut. 2010. “TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use.” *Cold Spring Harbor Perspectives in Biology*. doi:10.1101/cshperspect.a001008.

Osato, Motomi. 2004. “Point Mutations in the RUNX1/AML1 Gene: Another Actor in RUNX Leukemia.” *Oncogene* 23 (24) (May 24): 4284–4296. doi:10.1038/sj.onc.1207779. http://www.nature.com/doifinder/10.1038/sj.onc.1207779.

Paez, J Guillermo, Pasi A Jänne, Jeffrey C Lee, Sean Tracy, Heidi Greulich, Stacey Gabriel, Paula Herman, et al. 2004. “EGFR Mutations in Lung Cancer: Correlation with Clinical Response to Gefitinib Therapy.” *Science (New York, N.Y.)* 304 (5676) (June 4): 1497–500. doi:10.1126/science.1099314. http://www.ncbi.nlm.nih.gov/pubmed/15118125.

Pandey, Gaurav, Bin Zhang, and Le Jian. 2013. “Predicting Submicron Air Pollution Indicators: A Machine Learning Approach.” *Environmental Science. Processes & Impacts* 15 (5) (May): 996–1005. doi:10.1039/c3em30890a. http://www.ncbi.nlm.nih.gov/pubmed/23535697.

Pansky, Ben. 1982. “Germ Layers and Their Derivatives.” In *Review of Medical Embryology*, 24. Macmillan USA. http://discovery.lifemapsc.com/library/review-of-medical-embryology/chapter-25-germ-layers-and-their-derivatives.

Pao, William, Vincent Miller, Maureen Zakowski, Jennifer Doherty, Katerina Politi, Inderpal Sarkaria, Bhuvanesh Singh, et al. 2004. “EGF Receptor Gene Mutations Are Common in Lung Cancers From ‘never Smokers’ and Are Associated with Sensitivity of Tumors to Gefitinib and Erlotinib.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (36) (September 7): 13306–11. doi:10.1073/pnas.0405220101. http://www.ncbi.nlm.nih.gov/pubmed/15329413.

Parla, Jennifer S, Ivan Iossifov, Ian Grabill, Mona S Spector, Melissa Kramer, and W Richard McCombie. 2011. “A Comparative Analysis of Exome Capture.” *Genome Biology* 12 (9) (January): R97. doi:10.1186/gb-2011-12-9-r97.

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3308060&tool=pmcentrez&rendertype=abstract.

Pentheroudakis, George, Yael Spector, Dimitrios Krikelis, Vassiliki Kotoula, Eti Meiri, Vassiliki Malamou-Mitsi, George Fountzilas, et al. 2013. "Global microRNA Profiling in Favorable Prognosis Subgroups of Cancer of Unknown Primary (CUP) Demonstrates No Significant Expression Differences with Metastases of Matched Known Primary Tumors." *Clinical and Experimental Metastasis* 30 (4): 431–439. doi:10.1007/s10585-012-9548-3.

Peters, Nikki, and Katrina Armstrong. 2005. "Racial Differences in Prostate Cancer Treatment Outcomes: A Systematic Review." *Cancer Nursing* 28 (2): 108–118.

Peterson, W., T. Birdsall, and W. Fox. 1954. "The Theory of Signal Detectability." *Transactions of the IRE Professional Group on Information Theory* 4 (4) (September): 171–212. doi:10.1109/TIT.1954.1057460. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1057460.

Pfeifer, Gerd P, Mikhail F Denissenko, Magali Olivier, Natalia Tretyakova, Stephen S Hecht, and Pierre Hainaut. 2002. "Tobacco Smoke Carcinogens, DNA Damage and p53 Mutations in Smoking-Associated Cancers." *Oncogene* 21 (48) (October 21): 7435–51. doi:10.1038/sj.onc.1205803. http://www.ncbi.nlm.nih.gov/pubmed/12379884.

Pfeifer, Gerd P, Young-Hyun You, and Ahmad Besaratinia. 2005. "Mutations Induced by Ultraviolet Light." *Mutation Research* 571 (1-2) (April 1): 19–31. doi:10.1016/j.mrfmmm.2004.06.057. http://www.ncbi.nlm.nih.gov/pubmed/15748635.

Pleasance, Erin D, R Keira Cheetham, Philip J Stephens, David J McBride, Sean J Humphray, Chris D Greenman, Ignacio Varela, et al. 2010. "A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome." *Nature* 463 (7278) (January 14): 191–6. doi:10.1038/nature08658. http://www.ncbi.nlm.nih.gov/pubmed/20016485.

Pleasance, Erin D, Philip J Stephens, Sarah O'Meara, David J McBride, Alison Meynert, David Jones, Meng-Lay Lin, et al. 2010. "A Small-Cell Lung Cancer Genome with Complex Signatures of Tobacco Exposure." *Nature* 463 (7278) (January 14): 184–90. doi:10.1038/nature08629. http://www.ncbi.nlm.nih.gov/pubmed/20016488.

Podolsky, Maxim D, Anton A Barchuk, Vladimir I Kuznetcov, Natalia F Gusarova, Vadim S Gaidukov, and Segrey A Tarakanov. 2016. "Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels." *Asian Pacific Journal of Cancer Prevention : APJCP* 17 (2): 835–8. http://www.ncbi.nlm.nih.gov/pubmed/26925688.

Polak, Paz, Rosa Karlić, Amnon Koren, Robert Thurman, Richard Sandstrom, Michael S. Lawrence, Alex Reynolds, et al. 2015. "Cell-of-Origin Chromatin Organization Shapes the Mutational Landscape of Cancer." *Nature* 518 (7539) (February 18): 360–4. doi:10.1038/nature14221. http://www.nature.com/doifinder/10.1038/nature14221.

Poon, Song Ling, Mi Ni Huang, Yang Choo, John R McPherson, Willie Yu, Hong Lee Heng, Anna Gan, et al. 2015. "Mutation Signatures Implicate Aristolochic Acid in Bladder Cancer Development." *Genome Medicine* 7 (1): 38. doi:10.1186/s13073-015-0161-3. http://www.ncbi.nlm.nih.gov/pubmed/26015808.

Poon, Song Ling, See-Tong Pang, John R McPherson, Willie Yu, Kie Kyon Huang, Peiyong Guan, Wen-Hui Weng, et al. 2013. “Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool.” *Science Translational Medicine* 5 (197) (August 7): 197ra101. doi:10.1126/scitranslmed.3006086. http://www.ncbi.nlm.nih.gov/pubmed/23926199.

Qiu, Bo, and M. Celeste Simon. 2015. “Oncogenes Strike a Balance between Cellular Growth and Homeostasis.” *Seminars in Cell & Developmental Biology* (August). doi:10.1016/j.semcdb.2015.08.005. http://linkinghub.elsevier.com/retrieve/pii/S1084952115001494.

Ramos, Alex H., Lee Lichtenstein, Manaswi Gupta, Michael S. Lawrence, Trevor J. Pugh, Gordon Saksena, Matthew Meyerson, and Gad Getz. 2015. “Oncotator: Cancer Variant Annotation Tool.” *Human Mutation* 36 (4) (April): E2423–E2429. doi:10.1002/humu.22771. http://doi.wiley.com/10.1002/humu.22771.

Reddy, E P, R K Reynolds, E Santos, and M Barbacid. 1982. “A Point Mutation Is Responsible for the Acquisition of Transforming Properties by the T24 Human Bladder Carcinoma Oncogene.” *Nature* 300 (5888): 149–152. doi:10.1038/300149a0.

Rizvi, Naiyer A, Matthew D Hellmann, Alexandra Snyder, Pia Kvistborg, Vladimir Makarov, Jonathan J Havel, William Lee, et al. 2015. “Cancer Immunology. Mutational Landscape Determines Sensitivity to PD-1 Blockade in Non-Small Cell Lung Cancer.” *Science (New York, N.Y.)* 348 (6230) (April 3): 124–8. doi:10.1126/science.aaa1348. http://www.ncbi.nlm.nih.gov/pubmed/25765070.

Roberts, Steven a, Michael S Lawrence, Leszek J Klimczak, Sara a Grimm, David Fargo, Petar Stojanov, Adam Kiezun, et al. 2013. “An APOBEC Cytidine Deaminase Mutagenesis Pattern Is Widespread in Human Cancers.” *Nature Genetics* 45 (9) (September 14): 970–6. doi:10.1038/ng.2702. http://www.ncbi.nlm.nih.gov/pubmed/23852170.

Roberts, Steven A, Joan Sterling, Cole Thompson, Shawn Harris, Deepak Mav, Ruchir Shah, Leszek J Klimczak, et al. 2012. “Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions.” *Molecular Cell* 46 (4) (May 25): 424–35. doi:10.1016/j.molcel.2012.03.030. http://www.ncbi.nlm.nih.gov/pubmed/22607975.

Rodin, Sergei N, and Andrei S Rodin. 2005. “Origins and Selection of p53 Mutations in Lung Carcinogenesis.” *Seminars in Cancer Biology* 15 (2) (April): 103–12. doi:10.1016/j.semcancer.2004.08.005. http://www.ncbi.nlm.nih.gov/pubmed/15652455.

Rosenbloom, Kate R, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, et al. 2015. “The UCSC Genome Browser Database: 2015 Update.” *Nucleic Acids Research* 43 (Database issue): D670–81. doi:10.1093/nar/gku1177. http://www.ncbi.nlm.nih.gov/pubmed/25428374.

Rowley, J D. 1973. “Letter: A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia Identified by Quinacrine Fluorescence and Giemsa Staining.” *Nature* 243: 290–293. doi:10.1038/243290a0.

Roy, David M., Logan A. Walsh, and Timothy A. Chan. 2014. “Driver Mutations of Cancer

Epigenomes." *Protein & Cell* 5 (4) (April 14): 265–296. doi:10.1007/s13238-014-0031-6. http://link.springer.com/10.1007/s13238-014-0031-6.

Saunders, Christopher T, Wendy S W Wong, Sajani Swamy, Jennifer Becq, Lisa J Murray, and R Keira Cheetham. 2012. "Strelka: Accurate Somatic Small-Variant Calling from Sequenced Tumor-Normal Sample Pairs." *Bioinformatics (Oxford, England)* 28 (14) (July 15): 1811–7. doi:10.1093/bioinformatics/bts271. http://www.ncbi.nlm.nih.gov/pubmed/22581179.

Schlesinger-Raab, Anne, André L Mihaljevic, Silvia Egert, Rebecca Emeny, Karl-Walter Jauch, Jörg Kleeff, Alexander Novotny, et al. 2015. "Outcome of Gastric Cancer in the Elderly: A Population-Based Evaluation of the Munich Cancer Registry." *Gastric Cancer : Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*. doi:10.1007/s10120-015-0527-7. http://www.ncbi.nlm.nih.gov/pubmed/26260874.

Schnitt, Stuart J. 2010. "Classification and Prognosis of Invasive Breast Cancer: From Morphology to Molecular Taxonomy." *Modern Pathology* 23 (May): S60–S64. doi:10.1038/modpathol.2010.33. http://www.nature.com/doifinder/10.1038/modpathol.2010.33.

Schwaab, Julia, Yago Diez, Arnau Oliver, Robert Martí, Jan van Zelst, Albert Gubern-Mérida, Ahmed Bensouda Mourri, Johannes Gregori, and Matthias Günther. 2016. "Automated Quality Assessment in Three-Dimensional Breast Ultrasound Images." *Journal of Medical Imaging (Bellingham, Wash.)* 3 (2) (April): 027002. doi:10.1117/1.JMI.3.2.027002. http://www.ncbi.nlm.nih.gov/pubmed/27158633.

Seiser, and Federico Innocenti. 2015. "Hidden Markov Model-Based CNV Detection Algorithms for Illumina Genotyping Microarrays." *Cancer Informatics* (January): 77. doi:10.4137/CIN.S16345. http://www.la-press.com/hidden-markov-model-based-cnv-detection-algorithms-for-illumina-genoty-article-a4624.

Sen, R, L Nayak, and R K De. 2016. "A Review on Host-Pathogen Interactions: Classification and Prediction." *European Journal of Clinical Microbiology & Infectious Diseases : Official Publication of the European Society of Clinical Microbiology* (July 29). doi:10.1007/s10096-016-2716-7. http://www.ncbi.nlm.nih.gov/pubmed/27470504.

Seo, Ji Yeon, Eun Hyo Jin, Hyun Jin Jo, Hyuk Yoon, Cheol Min Shin, Young Soo Park, Nayoung Kim, Hyun Chae Jung, and Dong Ho Lee. 2015. "Clinicopathologic and Molecular Features Associated with Patient Age in Gastric Cancer." *World Journal of Gastroenterology : WJG* 21 (22): 6905–13. doi:10.3748/wjg.v21.i22.6905. http://www.ncbi.nlm.nih.gov/pubmed/26078567.

Shen, Tony, Stefan Hans Pajaro-Van de Stadt, Nai Chien Yeat, and Jimmy C-H Lin. 2015. "Clinical Applications of next Generation Sequencing in Cancer: From Panels, to Exomes, to Genomes." *Frontiers in Genetics* 6: 215. doi:10.3389/fgene.2015.00215. http://www.ncbi.nlm.nih.gov/pubmed/26136771.

Sherry, S T, M Ward, and K Sirotkin. 1999. "dbSNP-Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation." *Genome Research* 9 (8): 677–679. doi:10.1101/gr.9.8.677.

Shevelev, Igor V, and Ulrich Hübscher. 2002. “The 3’ 5' Exonucleases.” *Nature Reviews. Molecular Cell Biology* 3 (5) (May): 364–76. doi:10.1038/nrm804. http://www.ncbi.nlm.nih.gov/pubmed/11988770.

Shinohara, Masanobu, Katsuhiro Io, Keisuke Shindo, Masashi Matsui, Takashi Sakamoto, Kohei Tada, Masayuki Kobayashi, Norimitsu Kadowaki, and Akifumi Takaori-Kondo. 2012. “APOBEC3B Can Impair Genomic Stability by Inducing Base Substitutions in Genomic DNA in Human Cells.” *Scientific Reports* 2: 806. doi:10.1038/srep00806. http://www.ncbi.nlm.nih.gov/pubmed/23150777.

Siegel, Rebecca, Deepa Naishadham, and Ahmedin Jemal. 2012. “Cancer Statistics, 2012.” *CA: A Cancer Journal for Clinicians* 62 (1) (January): 10–29. doi:10.3322/caac.20138. http://doi.wiley.com/10.3322/caac.20138.

Simeon, Vittorio, Katia Todoerti, Francesco La Rocca, Antonella Caivano, Stefania Trino, Marta Lionetti, Luca Agnelli, et al. 2015. “Molecular Classification and Pharmacogenetics of Primary Plasma Cell Leukemia: An Initial Approach toward Precision Medicine.” *International Journal of Molecular Sciences* 16 (8): 17514–34. doi:10.3390/ijms160817514. http://www.ncbi.nlm.nih.gov/pubmed/26263974.

Singh, Swadha, and Raghvendra Singh. 2016. “Application of Supervised Machine Learning Algorithms for the Classification of Regulatory RNA Riboswitches.” *Briefings in Functional Genomics* (April 3): elw005. doi:10.1093/bfgp/elw005. http://bfgp.oxfordjournals.org/lookup/doi/10.1093/bfgp/elw005.

Sisu, C., B. Pei, J. Leng, A. Frankish, Y. Zhang, S. Balasubramanian, R. Harte, et al. 2014. “Comparative Analysis of Pseudogenes across Three Phyla.” *Proceedings of the National Academy of Sciences*: 1407293111–. doi:10.1073/pnas.1407293111. http://www.pnas.org/cgi/content/long/1407293111v1.

Smith, Harold C, Ryan P Bennett, Ayse Kizilyer, William M McDougall, and Kimberly M Prohaska. 2012. “Functions and Regulation of the APOBEC Family of Proteins.” *Seminars in Cell & Developmental Biology* 23 (3) (May): 258–68. doi:10.1016/j.semcdb.2011.10.004. http://www.ncbi.nlm.nih.gov/pubmed/22001110.

Sommer, C., and D. W. Gerlich. 2013. “Machine Learning in Cell Biology - Teaching Computers to Recognize Phenotypes.” *Journal of Cell Science* 126 (24) (December 15): 5529–5539. doi:10.1242/jcs.123604. http://jcs.biologists.org/cgi/doi/10.1242/jcs.123604.

Stephens, Philip J., Patrick S. Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C. Wedge, Serena Nik-Zainal, et al. 2012. “The Landscape of Cancer Genes and Mutational Processes in Breast Cancer.” *Nature* 486 (7403) (June 21): 400–4. doi:10.1038/nature11017. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3428862&tool=pmcentrez&rendertype=abstract.

Stojadinovic, Alexander, Aviram Nissan, John Eberhardt, Terence C. Chua, Joerg O W Pelz, and Jesus Esquivel. 2011. “Development of a Bayesian Belief Network Model for Personalized Prognostic Risk Assessment in Colon Carcinomatosis.” *American Surgeon* 77 (2): 221–230.

Suspène, Rodolphe, Marie-Ming Aynaud, Denise Guétard, Michel Henry, Grace Eckhoff,

Agnès Marchio, Pascal Pineau, Anne Dejean, Jean-Pierre Vartanian, and Simon Wain-Hobson. 2011. “Somatic Hypermutation of Human Mitochondrial and Nuclear DNA by APOBEC3 Cytidine Deaminases, a Pathway for DNA Catabolism.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (12) (March 22): 4858–63. doi:10.1073/pnas.1009687108. http://www.ncbi.nlm.nih.gov/pubmed/21368204.

Syn, Nicholas Li-Xun, Wei-Peng Yong, Boon-Cher Goh, and Soo-Chin Lee. 2016. “Evolving Landscape of Tumor Molecular Profiling for Personalized Cancer Therapy: A Comprehensive Review.” *Expert Opinion on Drug Metabolism & Toxicology* (June 13): 1–12. doi:10.1080/17425255.2016.1196187. http://www.ncbi.nlm.nih.gov/pubmed/27249175.

Szikriszt, Bernadett, Ádám Póti, Orsolya Pipek, Marcin Krzystanek, Nnennaya Kanu, János Molnár, Dezső Ribli, et al. 2016. “A Comprehensive Survey of the Mutagenic Impact of Common Cancer Cytotoxics.” *Genome Biology* 17 (1): 99. doi:10.1186/s13059-016-0963-7. http://www.ncbi.nlm.nih.gov/pubmed/27161042.

Tarca, Adi L., Vincent J. Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. 2007. “Machine Learning and Its Applications to Biology.” *PLoS Computational Biology* 3 (6): e116. doi:10.1371/journal.pcbi.0030116. http://dx.plos.org/10.1371/journal.pcbi.0030116.

TCGA. 2013. “Mutation Annotation Format (MAF) Specification - TCGA - National Cancer Institute - Confluence Wiki.” https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification.

Tevfik Dorak, M., and Ebru Karpuzoglu. 2012. “Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue.” *Frontiers in Genetics* 3 (NOV): 1–11. doi:10.3389/fgene.2012.00268. http://journal.frontiersin.org/article/10.3389/fgene.2012.00268/abstract.

The Cancer Genome Atlas. 2013. “The Cancer Genome Atlas - Data Portal.” https://tcga-data.nci.nih.gov/tcga/.

Tomita-Mitchell, A, A G Kat, L A Marcelino, X C Li-Sucholeiki, J Goodluck-Griffith, and W G Thilly. 2000. “Mismatch Repair Deficient Human Cells: Spontaneous and MNNG-Induced Mutational Spectra in the HPRT Gene.” *Mutation Research* 450 (1-2) (May 30): 125–38. http://www.ncbi.nlm.nih.gov/pubmed/10838138.

Tomlinson, I. 2012. “Colorectal Cancer Genetics: From Candidate Genes to GWAS and Back Again.” *Mutagenesis* 27 (2) (March 1): 141–142. doi:10.1093/mutage/ger072. http://www.mutage.oxfordjournals.org/cgi/doi/10.1093/mutage/ger072.

Travis, Lois B. 2006. “The Epidemiology of Second Primary Cancers.” *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 15 (11) (November): 2020–6. doi:10.1158/1055-9965.EPI-06-0414. http://www.ncbi.nlm.nih.gov/pubmed/17057028.

Travis, W. D., E. Brambilla, and G. J. Riely. 2013. “New Pathologic Classification of Lung Cancer: Relevance for Clinical Practice and Clinical Trials.” *Journal of Clinical Oncology* 31 (8) (March 10): 992–1001. doi:10.1200/JCO.2012.46.9270.

http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2012.46.9270.

Tsoi, Kelvin K. F., Joyce Y. C. Chan, Hoyee W. Hirai, Samuel Y. S. Wong, and Timothy C. Y. Kwok. 2015. "Cognitive Tests to Detect Dementia." *JAMA Internal Medicine* 175 (9) (September): 1450. doi:10.1001/jamainternmed.2015.2152.

Tuononen, Katja, Satu Mäki-Nevala, Virinder Kaur Sarhadi, Aino Wirtanen, Mikko Rönty, Kaisa Salmenkivi, Jenny M Andrews, et al. 2013. "Comparison of Targeted next-Generation Sequencing (NGS) and Real-Time PCR in the Detection of EGFR, KRAS, and BRAF Mutations on Formalin-Fixed, Paraffin-Embedded Tumor Material of Non-Small Cell Lung Carcinoma-Superiority of NGS." *Genes, Chromosomes & Cancer* 52 (5) (May): 503–11. doi:10.1002/gcc.22047. http://www.ncbi.nlm.nih.gov/pubmed/23362162.

Usher-Smith, Juliet A, Fiona M Walter, Jon D Emery, Aung K Win, and Simon J Griffin. 2016. "Risk Prediction Models for Colorectal Cancer: A Systematic Review." *Cancer Prevention Research (Philadelphia, Pa.)* 9 (1) (January): 13–26. doi:10.1158/1940-6207.CAPR-15-0274.

Vanderbilt-Ingram Cancer Center. 2010. "My Cancer Genome." http://www.mycancergenome.org.

Varadhachary, Gauri R., and Martin N. Raber. 2014. "Cancer of Unknown Primary Site." *New England Journal of Medicine* 371 (8) (August 21): 757–765. doi:10.1056/NEJMra1303917. http://www.nejm.org/doi/abs/10.1056/NEJMra1303917.

Visel, Axel, Edward M. Rubin, and Len A. Pennacchio. 2009. "Genomic Views of Distant-Acting Enhancers." *Nature* 461 (7261) (September 10): 199–205. doi:10.1038/nature08451. http://www.nature.com/doifinder/10.1038/nature08451.

Vyas, Renu, Sanket Bapat, Esha Jain, Sanjeev S Tambe, Muthukumarasamy Karthikeyan, and Bhaskar D Kulkarni. 2015. "A Study of Applications of Machine Learning Based Classification Methods for Virtual Screening of Lead Molecules." *Combinatorial Chemistry & High Throughput Screening* 18 (7): 658–72. http://www.ncbi.nlm.nih.gov/pubmed/26138573.

Walker, Brian A, Christopher P Wardell, Alex Murison, Eileen M Boyle, Dil B Begum, Nasrin M Dahir, Paula Z Proszek, et al. 2015. "APOBEC Family Mutational Signatures Are Associated with Poor Prognosis Translocations in Multiple Myeloma." *Nature Communications* 6: 6997. doi:10.1038/ncomms7997. http://www.ncbi.nlm.nih.gov/pubmed/25904160.

Walser, Jean-Claude, Loïc Ponger, and Anthony V Furano. 2008. "CpG Dinucleotides and the Mutation Rate of Non-CpG DNA." *Genome Research* 18 (9) (September): 1403–14. doi:10.1101/gr.076455.108. http://www.ncbi.nlm.nih.gov/pubmed/18550801.

Wang, Kai, Junsuo Kan, Siu Tsan Yuen, Stephanie T Shi, Kent Man Chu, Simon Law, Tsun Leung Chan, et al. 2011. "Exome Sequencing Identifies Frequent Mutation of ARID1A in Molecular Subtypes of Gastric Cancer." *Nature Genetics* (October) (October 30): 1–7. doi:10.1038/ng.982. http://www.ncbi.nlm.nih.gov/pubmed/22037554.

Wang, Zhong, Mark Gerstein, and Michael Snyder. 2009. "RNA-Seq: A Revolutionary Tool for Transcriptomics." *Nature Reviews. Genetics* 10 (1) (January): 57–63.

doi:10.1038/nrg2484. http://www.nature.com/doifinder/10.1038/nrg2484.

Watson, JD, and Francis HC FHC Crick. 1953. “Molecular Structure of Nucleic Acids.” *Nature* 171: 737–8. doi:10.1038/171737a0. http://www.nature.com/physics/looking-back/crick/.

Weissmann, S, T Alpermann, V Grossmann, A Kowarsch, N Nadarajah, C Eder, F Dicker, et al. 2012. “Landscape of TET2 Mutations in Acute Myeloid Leukemia.” *Leukemia* 26 (5) (May): 934–42. doi:10.1038/leu.2011.326. http://www.ncbi.nlm.nih.gov/pubmed/22116554.

Wellcome Trust Sanger Institute. 2016a. “COSMIC: Signatures of Mutational Processes in Human Cancer (Mutation Signatures).” http://cancer.sanger.ac.uk/cosmic/signatures.

Wellcome Trust Sanger Institute. 2016b. “COSMIC: Expert Curation of Genes.” http://cancer.sanger.ac.uk/cosmic/curation.

Wichchukit, Sukanya, and Michael O’Mahony. 2010. “A Transfer of Technology from Engineering: Use of ROC Curves from Signal Detection Theory to Investigate Information Processing in the Brain during Sensory Difference Testing.” *Journal of Food Science* 75 (9) (November): R183–R193. doi:10.1111/j.1750-3841.2010.01863.x.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. “Top 10 Algorithms in Data Mining.” *Knowledge and Information Systems* 14 (1) (January 4): 1–37. doi:10.1007/s10115-007-0114-2. http://link.springer.com/10.1007/s10115-007-0114-2.

Xi, Ruibin, Tae-Min Kim, and Peter J Park. 2010. “Detecting Structural Variations in the Human Genome Using next Generation Sequencing.” *Briefings in Functional Genomics* 9 (5-6) (December): 405–15. doi:10.1093/bfgp/elq025. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3080742&tool=pmcentrez&rendertype=abstract.

Yang, Haiwang, Yan Zhong, Cheng Peng, Jian-Qun Chen, and Dacheng Tian. 2010. “Important Role of Indels in Somatic Mutations of Human Cancer Genes.” *BMC Medical Genetics* 11 (January): 128. doi:10.1186/1471-2350-11-128. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2940769&tool=pmcentrez&rendertype=abstract.

Yen, Angela, and Manolis Kellis. 2015. “Systematic Chromatin State Comparison of Epigenomes Associated with Diverse Properties Including Sex and Tissue Type.” *Nature Communications* 6 (August 18): 7973. doi:10.1038/ncomms8973. http://www.nature.com/doifinder/10.1038/ncomms8973.

Zachariah, Nishant, Sonal Kothari, Senthil Ramamurthy, Adeboye O Osunkoya, and May D Wang. 2014. “Evaluation of Performance Metrics for Histopathological Image Classifier Optimization.” *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference* 2014: 1933–6. doi:10.1109/EMBC.2014.6943990. http://www.ncbi.nlm.nih.gov/pubmed/25570358.

Zang, Zhi Jiang, Ioana Cutcutache, Song Ling Poon, Shen Li Zhang, John R McPherson, Jiong Tao, Vikneswari Rajasegaran, et al. 2012. “Exome Sequencing of Gastric

Adenocarcinoma Identifies Recurrent Somatic Mutations in Cell Adhesion and Chromatin Remodeling Genes." *Nature Genetics* 44 (5) (May 8): 570–4. doi:10.1038/ng.2246. http://www.nature.com/doifinder/10.1038/ng.2246.

Zhang, Ben, Wei-Hua Jia, Koichi Matsuda, Sun-Seog Kweon, Keitaro Matsuo, Yong-Bing Xiang, Aesun Shin, et al. 2014. "Large-Scale Genetic Study in East Asians Identifies Six New Loci Associated with Colorectal Cancer Risk." *Nature Genetics* 46 (6) (May 18): 533–542. doi:10.1038/ng.2985. http://www.nature.com/doifinder/10.1038/ng.2985.

Zhang, J., J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, et al. 2011. "International Cancer Genome Consortium Data Portal--a One-Stop Shop for Cancer Genomics Data." *Database* 2011 (September 19): bar026–bar026. doi:10.1093/database/bar026. http://database.oxfordjournals.org/cgi/doi/10.1093/database/bar026.

Zhang, Tongwu, Yingfeng Luo, Kan Liu, Linlin Pan, Bing Zhang, Jun Yu, and Songnian Hu. 2011. "BIGpre: A Quality Assessment Package for Next-Generation Sequencing Data." *Genomics, Proteomics and Bioinformatics* 9 (6): 238–244. doi:10.1016/S1672-0229(11)60027-2.

Zweig, M H, and G Campbell. 1993. "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine." *Clinical Chemistry* 39 (4) (April): 561–77.

Appendices

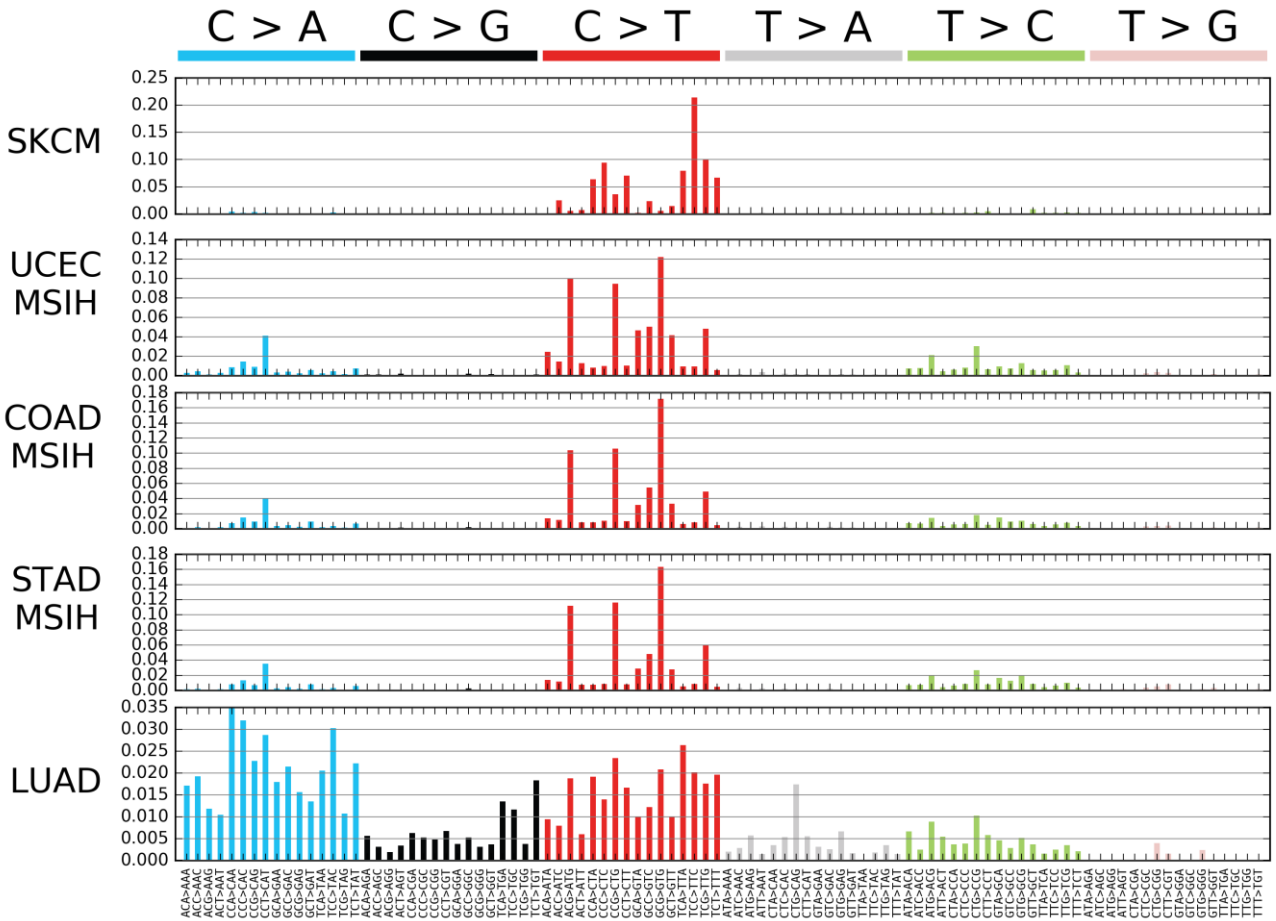

Figure 47: Bar graphs of trinucleotide mutations proportions in SKCM, UCEC-MSIH, COAD-MSIH, STAD-MSIH and LUAD

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

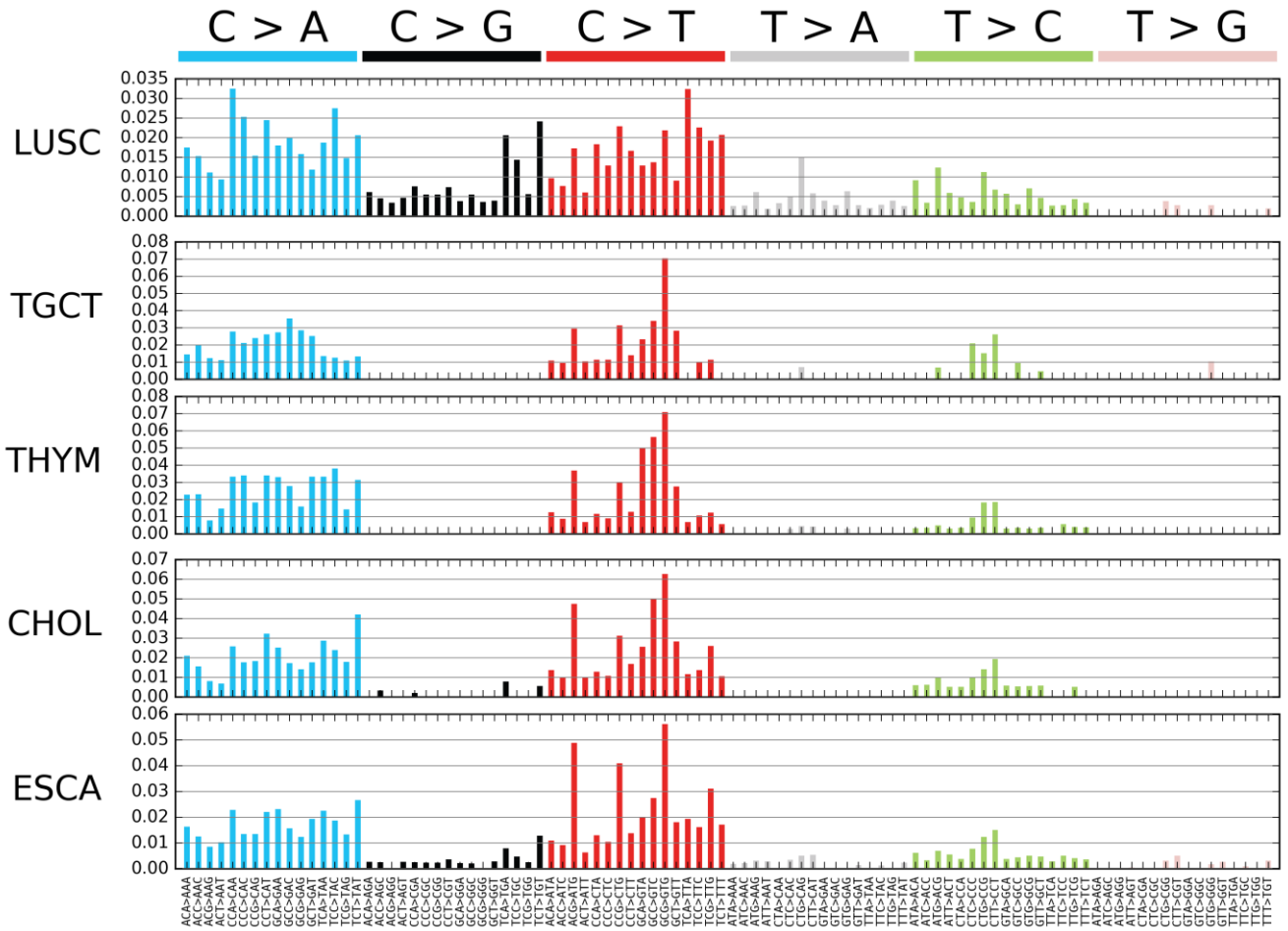

Figure 48: Bar graphs of trinucleotide mutations proportions in LUSC, TGCT, THYM, CHOL and ESCA

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

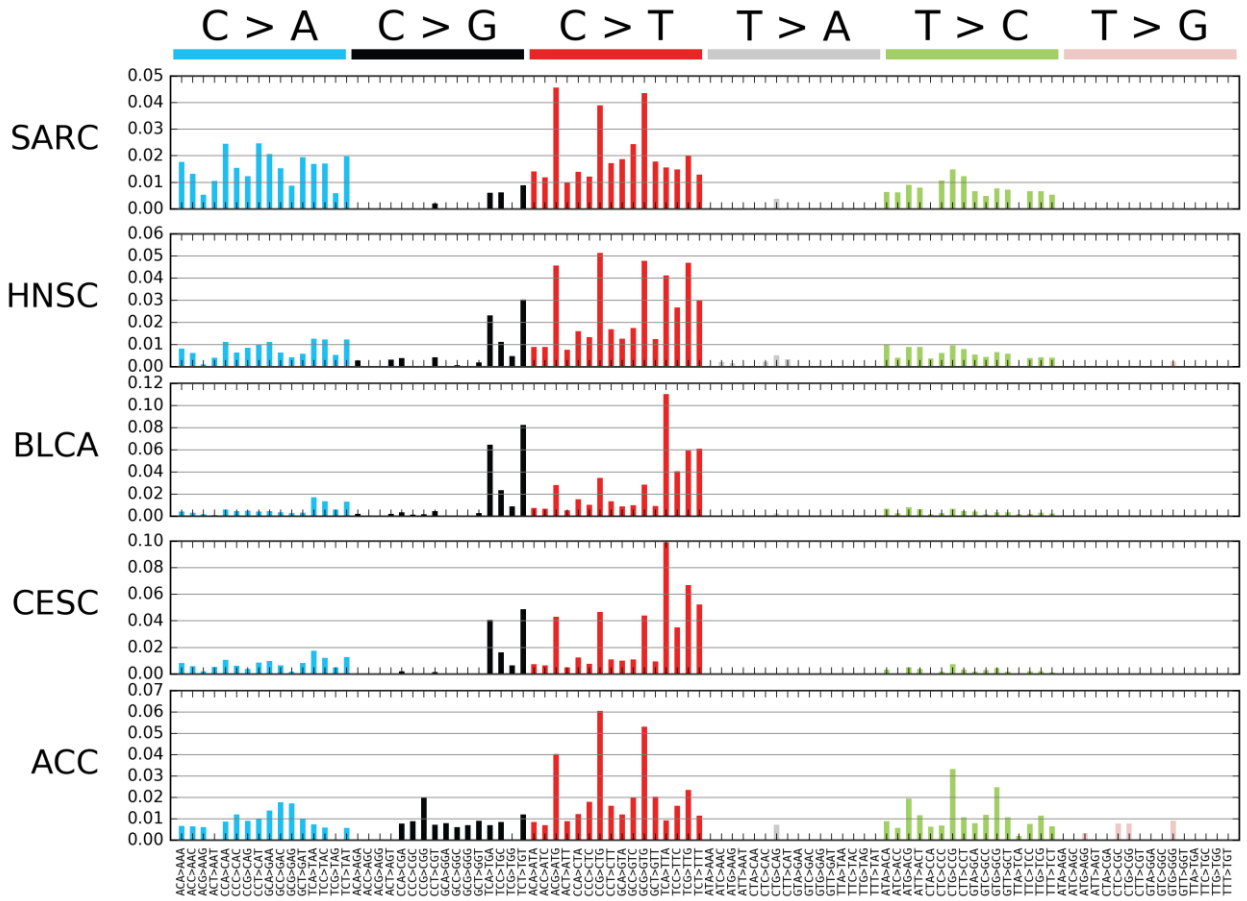

Figure 49: Bar graphs of trinucleotide mutations proportions in SARC, HNSC, BLCA, CESC and ACC

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

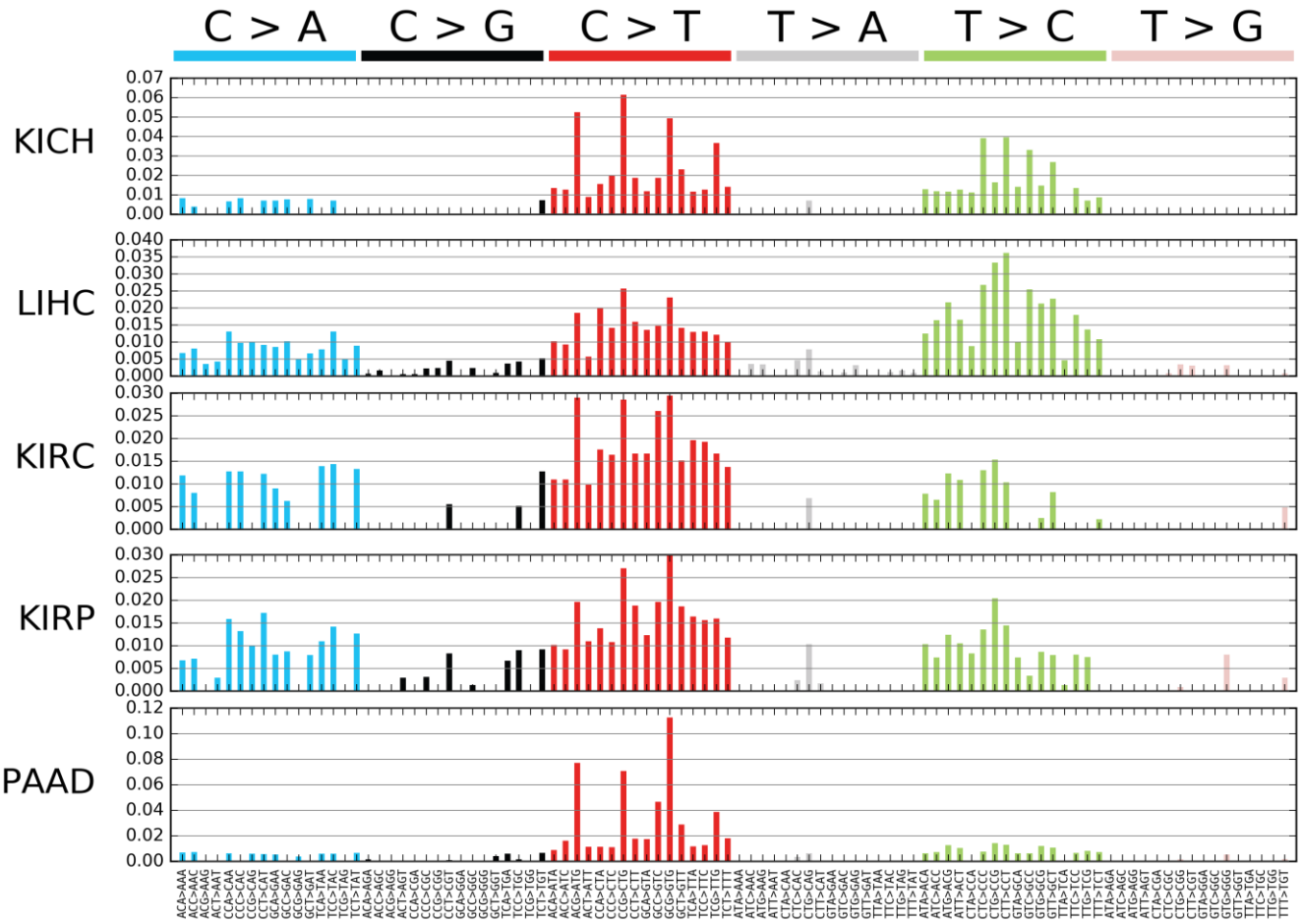

Figure 50: Bar graphs of trinucleotide mutations proportions in KICH, LIHC, KIRC, KIRP and PAAD

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

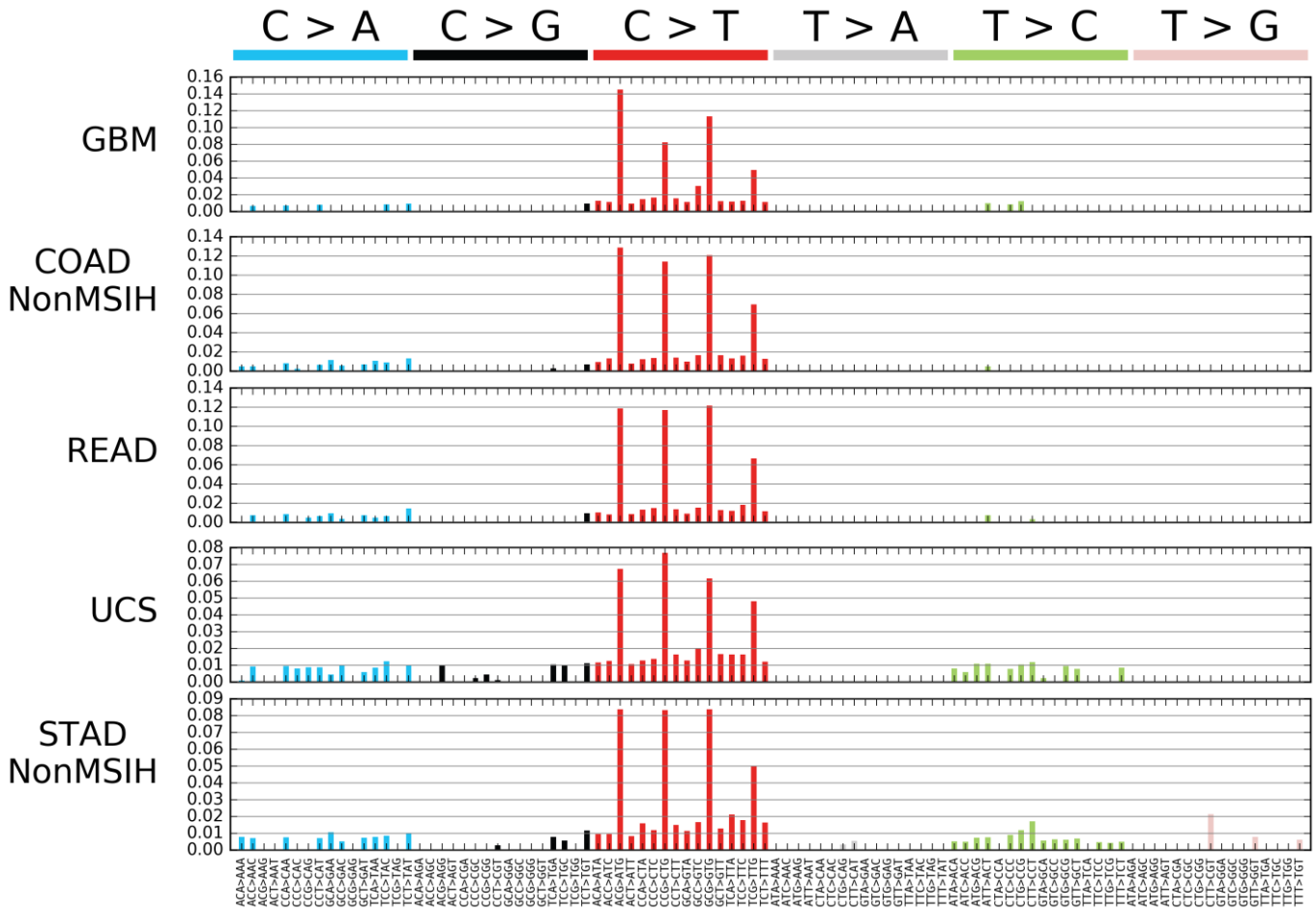

Figure 51: Bar graphs of trinucleotide mutations proportions in GBM, COAD-NonMSIH, READ, UCS and STAD-NonMSIH

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

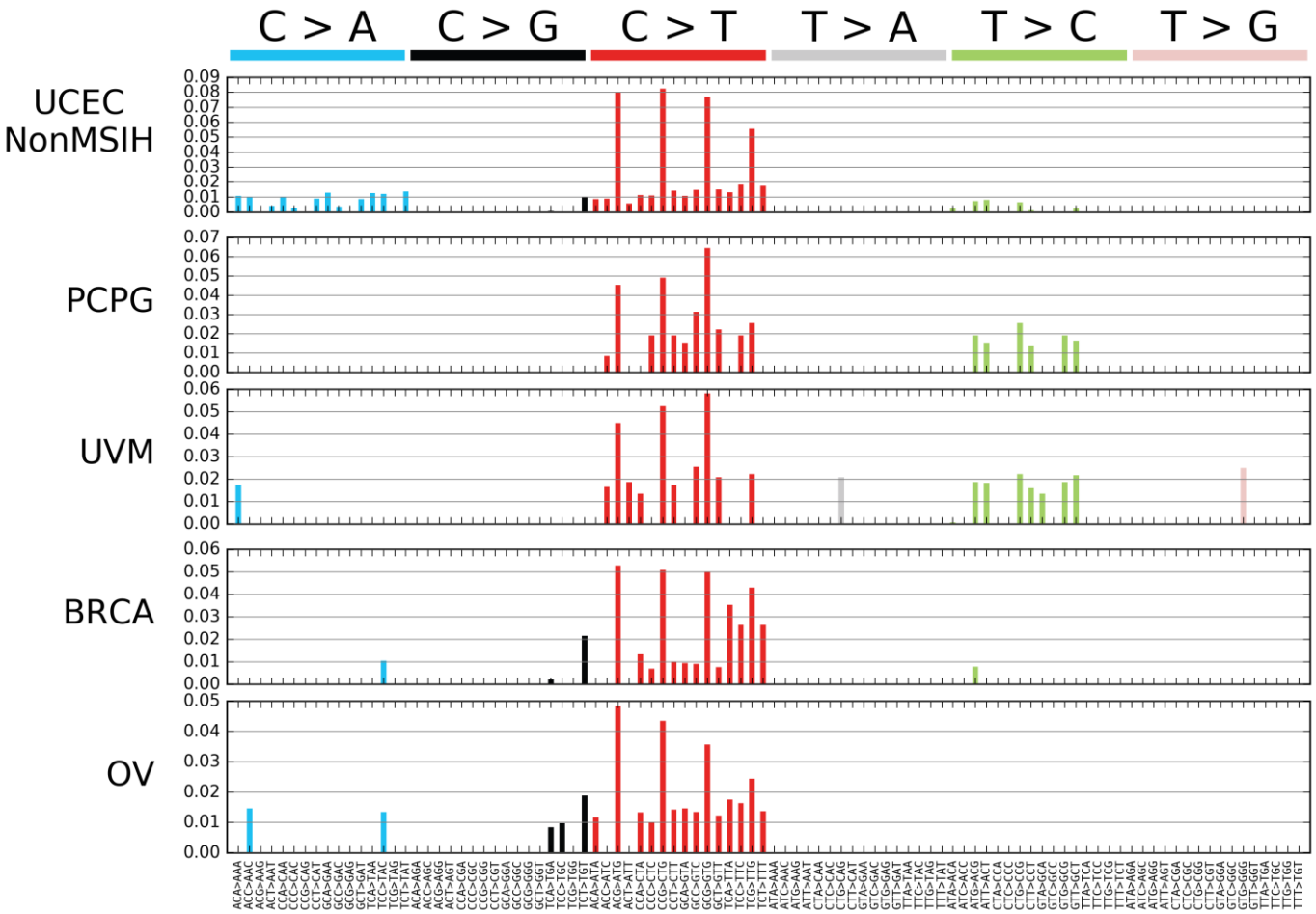

Figure 52: Bar graphs of trinucleotide mutations proportions in UCEC-NonMSIH, PCPG, UVM, BRCA and OV

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutations proportions clustering as shown in section 2.3.3.1)

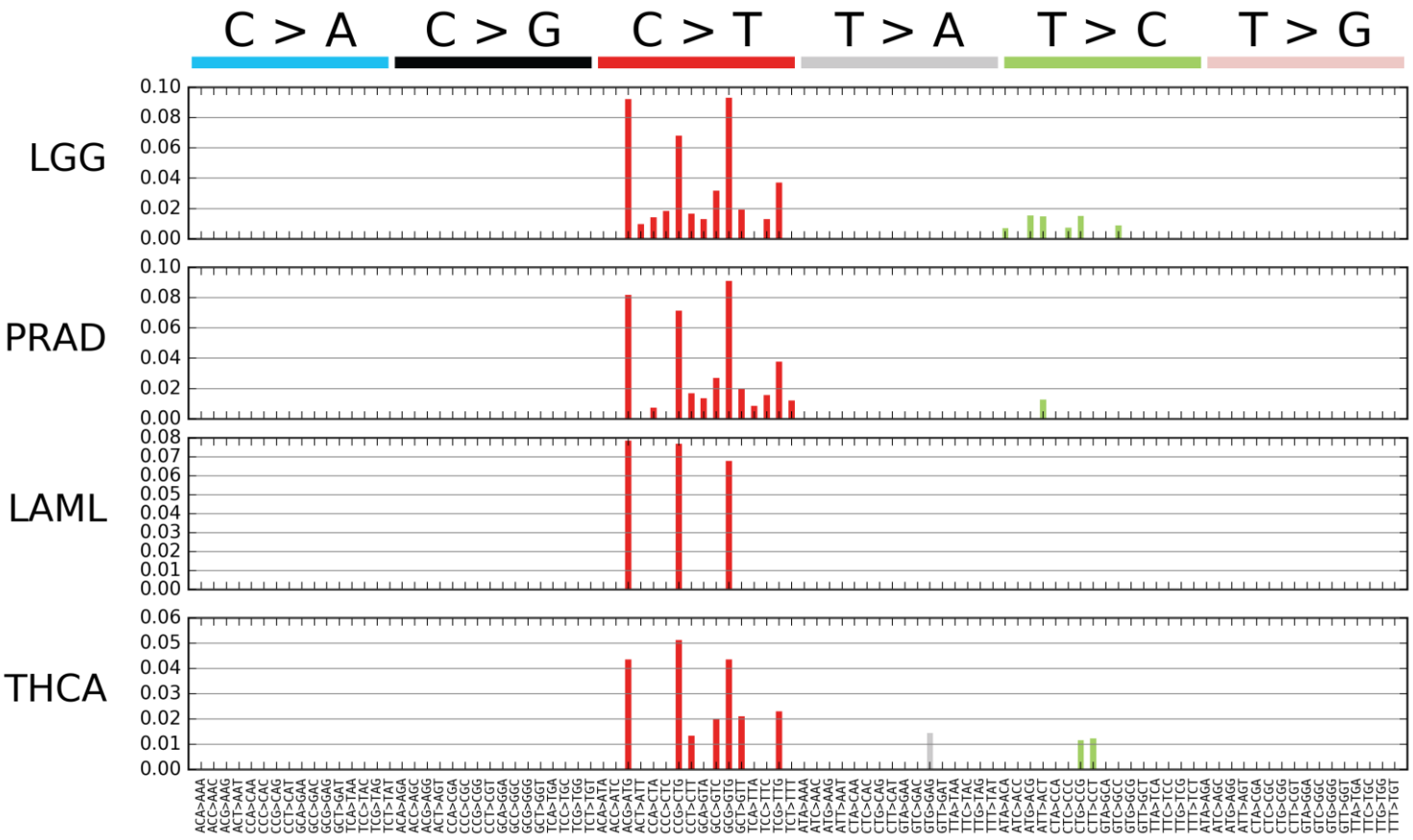

Figure 53: Bar graphs of trinucleotide mutations proportions in LGG, PRAD, LAML and THCA

(Cancers in Figure 47 to Figure 53 are ordered according to consensus trinucleotide mutation proportions clustering as shown in section 2.3.3.1)

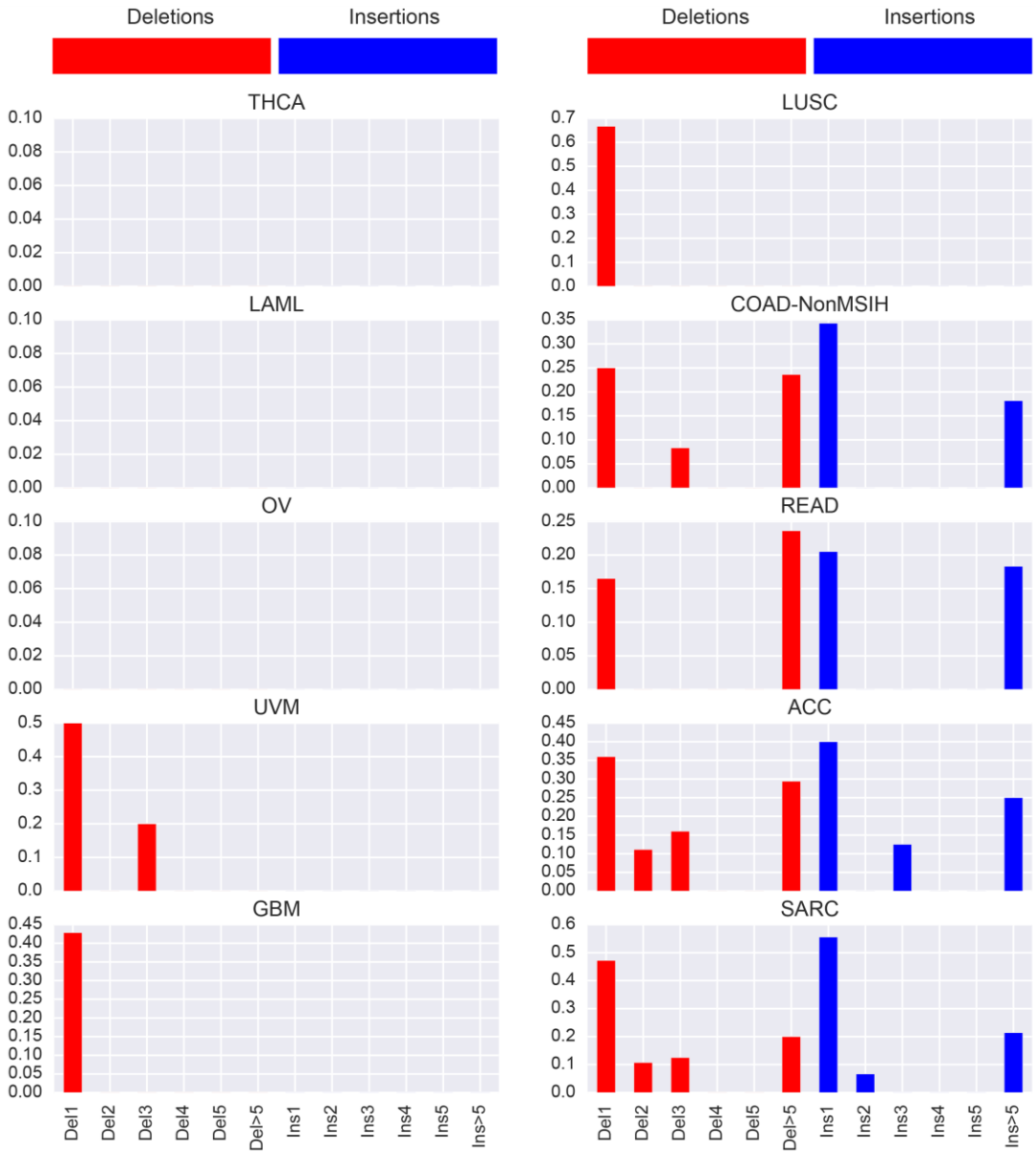

Figure 54: Bar graphs of indel size distribution in THCA, LAML, OV, UVM, GBM, LUSC, COAD-nonMSIH, READ, ACC and SARC

(Cancers in Figure 54 to Figure 57 are ordered according to consensus indel proportions clustering as shown in section 2.3.3.2)

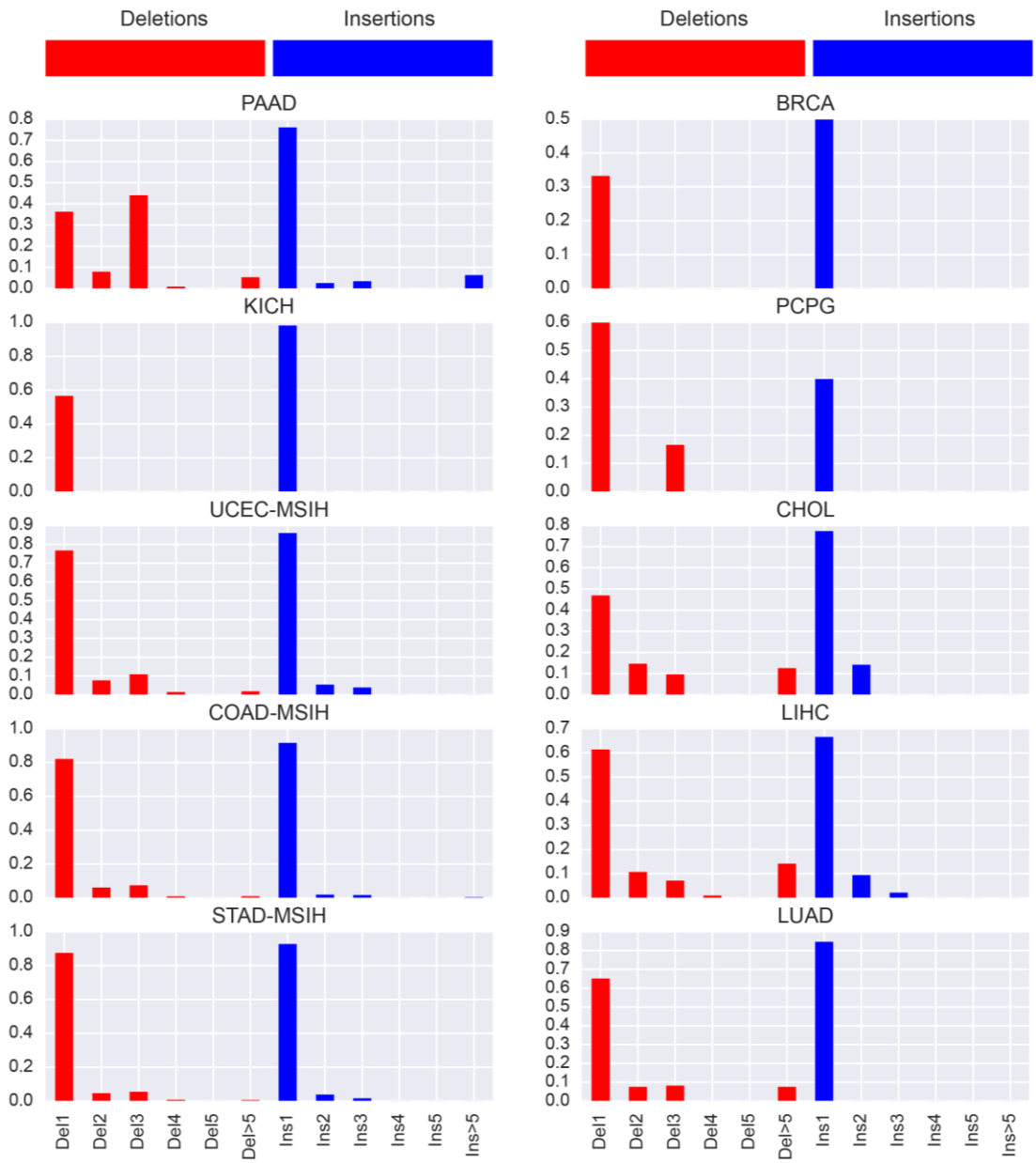

Figure 55: Bar graphs of indel size distribution in PAAD, KICH, UCEC-MSIH, COAD-MSIH, STAD-MSIH, BRCA, PCPG, CHOL, LIHC and LUAD

(Cancers in Figure 54 to Figure 57 are ordered according to consensus indel proportions clustering as shown in section 2.3.3.2)

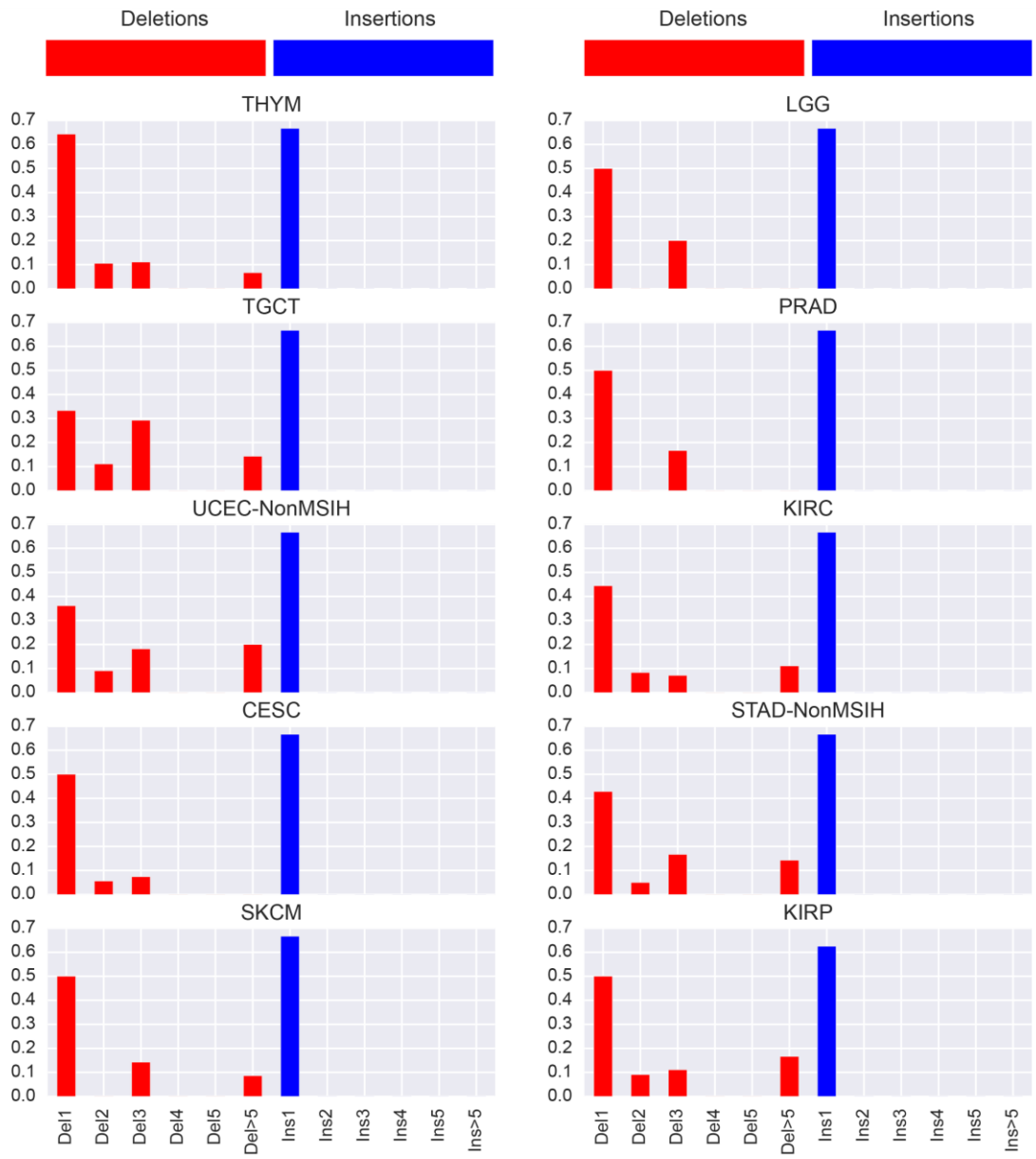

Figure 56: Bar graphs of indel size distribution in THYM, TGCT, UCEC-NonMSIH, CESC, SKCM, LGG, PRAD, KIRC, STAD-NonMSIH and KIRP

(Cancers in Figure 54 to Figure 57 are ordered according to consensus indel proportions clustering as shown in section 2.3.3.2)

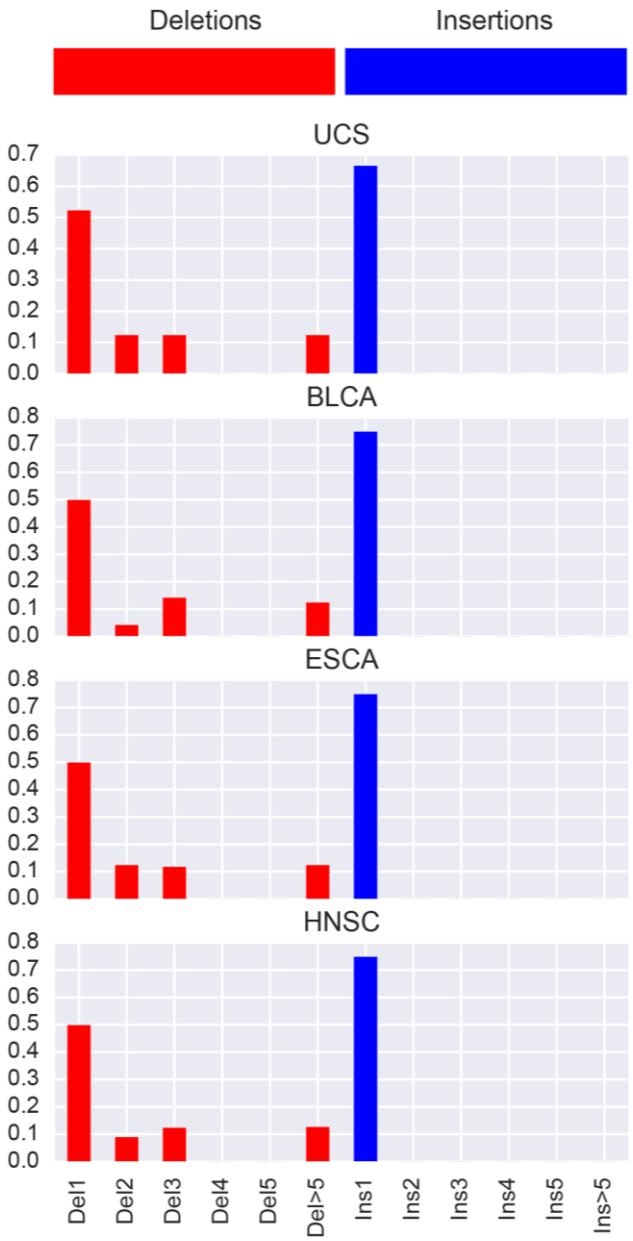

Figure 57: Bar graphs of indel size distribution in UCS, BLCA, ESCA and HNSC

(Cancers in Figure 54 to Figure 57 are ordered according to consensus indel proportions clustering as shown in section 2.3.3.2)