Database-Driven Whole Genome Profiling for Stratifying Triple Negative Breast Cancers (TNBC)

by

Rebecca Asiimwe

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

Database-Driven Whole Genome Profiling for Stratifying Triple Negative Breast Cancers (TNBC)

Submitted by Rebecca Asiimwe in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Examining Committee:

Dr. Sohrab P. Shah

Supervisor

Dr. David Huntsman

Supervisory Committee Member

Dr. William Hsiao

Supervisory Committee Member

Dr. Martin Hirst

Committee Chair

Abstract

Whole genome sequencing of cancers for variant discovery and patient stratification generates vast amounts of data including on the order of 10ˆ6 relevant features per sample. The current practice is to store this data in flat files whose structure complicates tasks required to optimally store, query and conduct integrative data mining and analysis of orthogonally collected data such as phenotype and clinical outcomes. In this study we designed, developed and optimized an object-relational database to support optimal storage, integration, querying, analysis and visualization of largescale whole genome profiling data at the level of genome-wide individual somatic variants (CNAs, SNVs, SVs and indels). We structured variant data from analytics pipelines and implemented a PostgreSQL database in which we bulk-loaded clinical outcomes and somatic variants from 88 Triple Negative Breast cancers (TNBCs). Our focus on TNBC was driven by the current and urgent need for better characterization of the genetic, molecular and clinical biomarkers of this heterogeneous, more aggressive and difficult to treat breast cancer subtype for which there are limited treatment options. Secondly, our inclination to whole genome sequencing (WGS) was attributed to the ability of WGS approaches to provide an in-depth analysis and elucidation of the landscape of mutations occurring across the genome that may reflect specific mutational processes as targetable vulnerabilities in human cancers. However, a whole genome sequencing study in TNBC at scale to investigate genomic properties as a stratification tool has not been undertaken. Hinged on these notions, we applied the developed database and present its indispensable utility in supporting optimal access, exploration, analysis and visualization of genomic contents of patient tumours to support quality control, inference of patterns of mutations and genomic events underpinning a patient's disease, population level aggregation analysis, gene mutation visualization and patient stratification. Furthermore, we developed Genome-Miner, a web-based database user interface to additionally support interactive and convenient access, sharing, interrogation and visualization of collected data across various research groups. We anticipate the database infrastructure we present will have utility in other whole genome studies and push the field beyond the use of flat files for managing whole genome datasets in cancer.

Lay Summary

Since the inception of DNA sequencing in the 1970s, various sequencing technologies have been introduced to help biologists understand the genetic makeup of individuals towards optimized treatment of diseases especially complex diseases such as cancer. These sequencing technologies generate vast amounts of data that are mostly stored in flat files whose structure does not support optimal storage, access, exploration, analysis and visualization of vast amounts of related genomic and clinical outcomes data. Using whole genome profiling data from 88 Triple Negative Breast Cancers (TNBC), we designed, developed, optimized and implemented a postgreSQL database and further developed Genome-Miner, a database-driven and web-based tool to support the optimal storage, exploration, analysis, visualization and global sharing of clinical outcomes and genomic contents of cancers in whole genome studies. We present the indispensable application of databases for quality control, population level aggregation analysis, gene mutation visualization and patient stratification.

Preface

This thesis was written under the guidance of my supervisor, Dr. Sohrab P. Shah in collaboration with my distinguished supervisory and examining committee comprised of Dr. David Huntsman, Dr. William Hsiao and Dr. Martin Hirst. The intellectual contents herein are original and written by Rebecca Asiimwe with contributions from the following:- Project conception, main methods and direction were provided by my supervisor Dr. Sohrab Shah, sample and data collection was done by Dr. Samuel Aparicio who also provided key insights and advice on the project; collection of clinical data from various institutions (Breast Cancer Outcomes Unit (BCOU), Alberta and Montreal) was done by Dr. Steven McKinney; sample and library preparation and construction was done by Damian Yap and Adrian Wan who further submitted libraries to the Genome Sciences Center (GSC) for Whole Genome Sequencing (WGS). Quality control (QC) and variant calling from the data from Whole Genome Sequencing was done by the Shah Lab. I completed all tasks pertaining downstream analysis, that is:- structuring and modelling the data from variant calling and data analytics pipelines, database design and development, bulk data loading, database management and optimization, downstream QC, identification of significantly mutated genes in the cohort, analysis of the genomic variants and clinical data in the database, TNBC subgroup discovery and analyses. Towards subgroup discovery, Tyler Funnel provided data on mutation signatures derived from the multi-modal correlated topic model (MMCTM) as one of the key input parameters for patient stratification. I further developed a database interface (Genom-Miner) to support database access and user defined data analyses, visualizations and sharing among various research groups, an idea conceived by my supervisor.

Table of Contents

Appendices

List of Tables

List of Figures

Glossary

ALOH	Amplified Loss of Heterozygosity
ASCNA	Allele-Specific Copy Number Amplification
BAM	Binary Alignment Map
BCNA	Balanced Copy Number Amplification
BCOU	Breast Cancer Outcomes Unit
BCS	Breast-Conserving Surgery
BLIA	Basal-Like Immune-Activated
BLIS	Basal-Like Immune-Suppressed
bp	Base Pair
Cl-FBI	Clustered Foldback Inversions
Cl-SV	Clustered Structural Variants
CNA	Copy Number Aberration
COSMIC	The Catalogue of Somatic Mutations in Cancer
CSV	Comma-Separated Values
DBMS	Database Management System
dbSNP	Single Nucleotide Polymorphism Database
DCIS	Ductal Carcinoma In Suti
DDR	DNA Damage Repair
DLOH	Hemizygous Deletion LOH
DMFS	Distant Metastasis Free Survival
DS	Disease Free Survival
DSS	Disease Specific Survival
EMBL	European Molecular Biology Laboratory
ER	Estrogen Receptor
ERD	Entity Relationship Diagram

FBI	Foldback Inversions
GAIN	Gain/Duplication of 1 Allele
GSC	Genome Sciences Center
HER2	Human Epidermal Growth Factor Receptor 2
HET	Diploid Heterozygous
HGSC	High-Grade Serous Carcinoma
HOMD	Homozygous Deletion
HPRD	Human Protein Reference Database
HRD	Homologous Recombination Deficiency
IDC	Infiltrating Ductal Carcinoma
IHC	Immunohistochemical
IM	Immunomodulatory
Indel	Insertions and Deletions
L-Del	Large Deletions
L-Dup	Large Duplications
LAR	Luminal Androgen Receptor
LCIS	Lobular Carcinoma In Suti
M	Mesenchymal
M-Dup	Medium Duplications
MMCTM	Multi-Model Correlated Topic Model
MMRD	Mismatch Repair Deficiency
MSL	Mesenchymal Stem-Like
mTNBC	Metastatic TNBC
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
NLOH	Copy Neutral LOH
NMF	Non-Negative Matrix Factorization
OS	Overall Survival
PARP	Poly(ADP-ribose) Polymerase
PARPi	Poly(ADP-ribose) Polymerase Inhibitors
pCR	Pathologic Complete Response
PDB	Protein Data Bank

PR	Progesterone Receptor
QC	Quality Control
RFS	Relapse-Free Survival
S-Del	Small Deletions
S-Dup	Small Duplications
SAM	Sequence Alignment Map
SMGs	Significantly Mutated Genes
SNV	Single Nucleotide Variant
SQL	Structured Query Language
SV	Structural Variant
TCGA	The Cancer Genome Atlas
TNBC	Triple Negative Breast Cancer
TNM	Tumor Node Metastasis
Tr	Translocations
UBCNA	Unbalanced Copy Number Amplification
VCF	Variant Call Format
VEP	Variant Effect Predictor
WGS	Whole Genome Sequencing

Acknowledgements

I would like to thank my supervisor, Dr. Sohrab Shah, without whom my research accomplishments wouldn't have been realized. Thank you Sohrab for such a great opportunity to work and learn from you and the lab in general. Your guidance, encouragement, thoughtfulness and support have been immeasurable. I would also like to thank you for the great environment and establishments put in place in the Shah Lab to enable students work effectively.

I would also like to thank members of my supervisory and defense committee, Dr. David Huntsman, Dr. William Hsiao and Dr. Martin Hirst for their continued support, guidance, keenness and key insights provided during my research journey.

To the faculty, staff and students of the department of Bioinformatics (UBC), thank you for the great insights, ideas and fun-filled events shared. I owe my special thanks to Dr. Steven Jones and Sharon Ruschkowski for the tireless efforts, guidance, advice and help rendered that made my graduate journey at UBC a success.

My appreciation also goes out to members of the Shah lab especially: Yikan Wang, Ali Bashashati, Tyler Funnell, Allen Zhang, Diljot Grewal, Daniel Lai, Andrew McPherson, Sohrab Salehi, Fatemeh Dorri, Kieran Campbell, Camila de Souza, Saeed Saberi, Oleg Golovko and Mirela Andronescu for their support, constructive criticisms and helpful insights that went a long way into shaping my research and its implementations.

Last but not least, I would like to thank my family and friends. Dad and mum, you have been awesome! Thank you for being their for me, encouraging and praying for me. Words are not enough to express my gratitude towards the love, support and care you have provided. To my sisters, brothers and to all my friends, you rock! Thank you for the encouragement and support when most needed.

Dedication

To My Family

Especially, Dad and Mum

Chapter 1

Introduction

1.1 Breast cancer

Worldwide, breast cancer is reported as one of the most common cancers with more than 1,300,000 cases and 450,000 deaths each year [1]. The multiple subtypes of breast cancer portray its heterogeneity that has proven an invaluable asset in understanding differences in patient outcomes and responses to therapy. Clinically, there are three therapeutic categories of breast cancer established by the presence or lack of three hormone receptors: oestrogen receptor *(ER)*, human epidermal growth factor receptor 2 *(HER2)* - also called *ERBB2* and progesterone receptor *(PR)*. However, various studies [1–3] have demonstrated that the heterogeneity of breast cancer extends far beyond these immunohistochemical (IHC) classifications. Intrinsic molecular breast cancer can also be classified as either luminal or basal-like dependent on the expression of different cytokeratins (basal-like cytokeratins: *KRT5*, *KRT6A*, *KRT6B*, *KRT14*, *KRT16*, *KRT17*, *KRT23*, and *KRT81*; luminal cytokeratins: *KRT7*, *KRT8*, *KRT18*, and *KRT19*)[4, 5] with the basal-like subtype accounting for 10-25% of all invasive breast cancers [6].

In addition to cytokeratin expression, breast cancers have further been classified as basal-like, HER2-like, normal breast–like, luminal A, and luminal B based on an “intrinsic/UNC” 306-gene set [2, 3]. This intrinsic subtyping of breast cancer by gene expression analyses was further supported by research done by The Cancer Genome Atlas Network [1] in which various omics data (DNA copy-number arrays, DNA methylation, exome sequencing, messenger-RNA arrays, microRNA sequencing, and reverse-phase protein arrays) were integrated to report four molecular breast cancer subtypes: luminal/*ER+*, *HER2* and basal-like. Each of the identified subtypes exhibited molecular heterogeneity, distinct domination of specific signaling pathways and enrichment for mutations in certain genes like the enrichment of specific mutations in *GATA3*, *PIK3CA* and *MAP3K1* within the luminal A subtype.

Histopathologically, breast cancer can be broadly classified into *in situ* carcinoma or invasive (infiltrating) carcinoma (figure 1.1). *in situ* carcinoma can be further subdivided into ductal carcinoma (DCIS) or lobular carcinoma (LCIS) which arises from multiple foci (10 or more) and therefore regarded as multicentric; bilateral LCIS is also common [7]. LCIS is not a premalignant lesion but is regarded indispensable in identifying women at increased risk of developing succeeding invasive breast cancers (DCIS) and as such mammographies taken regularly could help in early breast cancer detection [7]. DCIS on the other hand are more prevalent and heterogeneous compared to LCIS and have a likelihood of progressing into an invasive cancer. DCIS is therefore characterized as pre-invasive or a precursor lesion and accounts for about 16% of all detected breast cancer malignancies. It is also reported to be multicentric in 40% of breast cancer cases with high rates of local relapse (50% recurrences) that could exceed those of invasive cancer after monotherapy treatment with breast-conserving surgery [8].

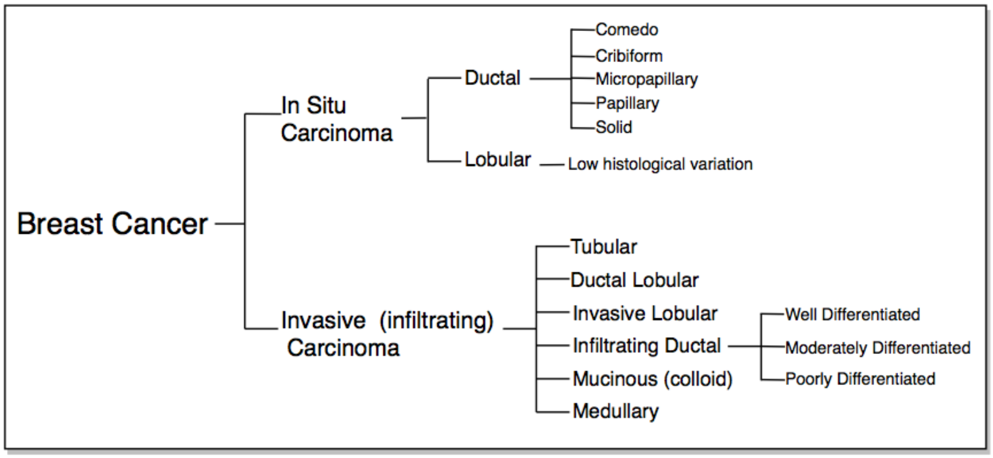


Figure 1.1: Histological classification of breast cancer subtypes. Figure modified from Malhotra et al. [9].

Invasive (infiltrating) carcinomas on the other hand are a heterogeneous group of tumors differentiated into seven histological subtypes: tubular, ductal/lobular, invasive lobular, infiltrating ductal, mucinous (colloid), medullary and papillary carcinomas (Fig. 1.1). Infiltrating ductal carcinoma (IDC) is the most common subtype accounting for 70–80% of all invasive lesions [9] and is further sub-classified by grade as either being well-differentiated, moderately differentiated or poorly differentiated [9].

Classification of breast cancers by histological grade has long been used as an indication of prognosis with a significant bearing on the choice of patient treatment. Grading is done based on cell morphology, similarity of cancerous cells to non-cancerous cells and the nuclear grade which shades light on the size and shape of the nucleous and proliferative index (NCI, 2013). Histological grades range from grade 1 to grade 3. In Grade 1, the cancer cells look like normal cells with a high homology to the normal breast terminal duct lobular unit. They are small and uniform with a mild degree of pleomorphism and are usually slow-growing compared to other breast cancer grades. Grade 1 is therefore regarded well-differentiated with a low proliferative index. Grade 2 breast cancer has cells slightly bigger than normal cells. They vary in shape, grow faster than normal cells and are moderately differentiated while Grade 3 cells look more abnormal compared to normal cells and are poorly differentiated or undifferentiated highly proliferative tumours (Fig. 1.2).

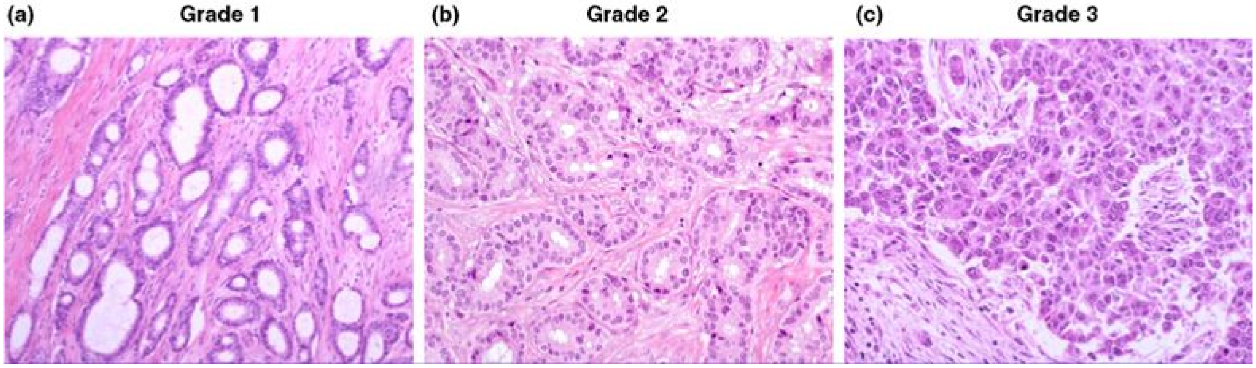


Figure 1.2: Histological grades of breast cancer obtained using the Nottingham Grading System: **(a)** Grade 1 - well differentiated tumors that exhibit high homology to the normal breast terminal duct lobular unit, low mitotic rates, a low incidence of nuclear polymorphism and are arranged in small tubes (tubule formation $>$ *75%*). **(b)** Grade 2 - moderately differentiated tumor. **(c)** Grade 3 - poorly or undifferentiated tumor - lacks normal features (no tubule formation $<$ *10%*), higher incidence of nuclear polymorphism, tends to grow and spread faster. Source: Rakha et.al [10].

Besides using breast cancer grades to classify patients, the tumor node metastasis (TNM) system was developed by the American Joint Committee on Cancer to stratify patients based on prognosis. Characteristics of a patient's primary tumour such as the size, lymph node status, invasiveness and existence of metastasis (local or distant) are among the key features incorporated into this system.

Treatment of breast cancer patients is currently informed by hormone receptor status (ER, PR and HER2), tumour size, lymph node status, cancer stage and the general health condition of a patient. Local, non-invasive breast cancers are treated with surgery as a mono-therapy or in combination with radiation. Towards effective surgery, neoadjuvant therapy is administered before surgery to reduce the size of a patient's tumour. In patients whose lymph node status is positive, adjuvant

Stage	**Tumour Size**	**Node Involvement**
I	< 2cm	No axillary lymph node involvement
	> 5cm	No node involvement
III		Extensive ipsilateral axillary lymph node positivity or supraclavicular lymph node involvement. Inflammatory carcinoma. Tumour extension into the chest wall or skin in the form of ulceration
IV		Distant metastasis

Table 1.1: Breast cancer stages, corresponding tumour sizes and node involvement. Patients with the poorest prognosis commonly present with stage III or IV breast cancer, tumour sizes *>5cm* and/ with a node positive status (node positivity indicates the likelihood of cancer spread to other tissues).

therapy is administered after surgery to reduce the risk of disease recurrence. More systematic approaches that have been applied in the treatment of breast cancer include the administration of chemotherapy, and targeted therapies that putitively reduce toxicity to normal cells. Among the current targeted therapies and standard of care for patients with breast cancer is tamoxifen, an anti-hormonal endocrine compound used to treat patients with ER and PR positive cancers. Trastuzumab a monoclonal antibody has also been used in the treatment of HER2 positive breast cancer.

1.2 Triple Negative Breast Cancer (TNBC)

1.2.1 Immunohistochemical classification and clinical characteristics of TNBC

Triple Negative Breast Cancer (TNBC) is a distinct subtype of breast cancer that represents 10% - 20% of all breast cancers worldwide [1, 4]. Immunohistochemically, TNBC is a breast cancer phenotype whose tumors are a subtype of exclusion, characterized by the lack of expression of biomarkers: estrogen receptor *(ER)* and progesterone receptor *(PR)* and for which the human epidermal growth factor receptor 2 *(HER-2)* is not over expressed or its gene not amplified as assessed by fluorescence in situ hybridization. TNBCs are also classified as basal-like based on the PAM50 classification in which 80.6% of TNBCs classified as basal-like. The notion that basal-like breast cancers account for the highest proportion of TNBCs is also supported by studies conducted by the TCGA [1], Lehman et al. [4] and Curtis et al. [11]. Synonymous to basal-like cancers, TNBCs exhibit high proliferative indices, mutations and genomic deletions in *TP53* and *RB1* [12]. They are also closely associated with the expression of high-molecular-weight basal cytokeratins

5/6, 14, and 17, P-cadherin, *p53*, and *EGFR* [1, 13, 14].

Clinically, TNBC is the most aggressive form of breast cancer [2] with the majority TNBCs histologically classified as being of higher grade compared to other types of breast cancer. They are invasive ductal carcinoma, usually found at a late stage [15, 16]. TNBCs are also characterized with large tumors whose size incongruity does not correlate with node status in women whose *tumours* < *5cm*. In a study conducted by Dent et al., even small tumours in TNBC had a high rate of node positivity with 55% of women with *tumours* < *1cm* having at least one positive lymph node [17] indicating an increased risk of their cancer spreading. TNBCs are also reported to be more common in young women ($age < 50 years$) [4, 16, 18] with a higher incidence among African-American and Hispanic women [19, 20].

Compared to hormone receptor positive invasive ductal carcinomas, TNBCs exhibit poorer prognosis with patients exhibiting a shorter time to relapse, metastatic disease and overall survival [15, 19, 21]. TNBC metastases also distinctively and predominantly affect the central nervous system, lymph nodes and visceral organs (especially the lungs) [22, 23] compared to other breast cancers whose relapses are commonly in bone and skin [22, 24, 25]. The high proliferative index and median survival of TNBC metastases ($\sim$ 12 months) are both reported much higher compared to other breast cancer types. Treatment of TNBCs with presurgical (neoadjuvant) chemotherapy has reported higher clinical response rates in some patients compared to response rates in other beast cancer types [1, 4, 18] (clinical response rates of up to 85% and pathologic complete response (pCR) rates of 30 - 40% [26]), however, despite these encouraging response rates, the vast majority of TNBC patients have very poor outcomes and are still at a greater risk of distant disease recurrence and rapid disease progression within 3 - 5 years of recurrence (Fig. 1.3) [17, 26, 27]. All patients with metastatic TNBC eventually die of the disease, despite having had adjuvant chemotherapy [17, 18].

1.2.2 Molecular and genomic stratification of TNBC

1.2.2.1 Molecular heterogeneity of TNBC

Current efforts towards elucidating TNBC and the stratification of patient groups that may elicit different biology and treatment response have been underway. In a study conducted by Lehmann et al. [4], the results from an aggregate analysis of 21 publicly available expression data sets: 3,247 primary human breast cancers and 587 TNBC gene expression profiles identified six distinct molecular

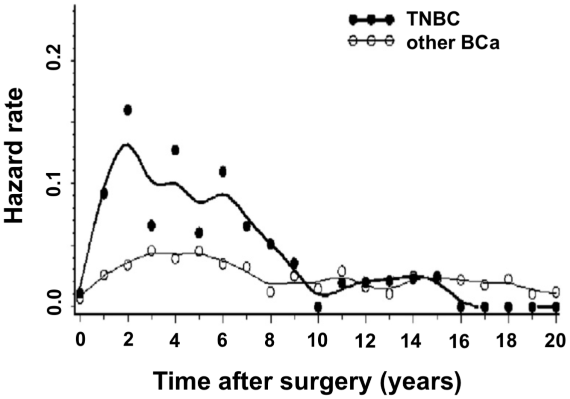


Figure 1.3: Rates of distance recurrence following surgery in a cohort of TNBC patients compared to other breast cancer patients. The hazard ratio for distant recurrence within the first 5 years post-surgery in TNBC compared to other breast cancers was 2.6; 95% Confidence interval (CI) 2.0 – 3.5. Source: Lee et.al [27].

TNBC subtypes: ((two basal-like (BL1 and BL2), two Mesenchymal subtypes (Mesenchymal (M) and Mesenchymal Stem-Like (MSL), Immunomodulatory (IM) and Luminal Androgen Receptor (LAR))), each showing distinctive biological phenotypes, gene ontologies, gene expression patterns and clinical outcomes. Pharmacological targeting of predicted driver signaling pathways in cell line models representative of each of the six subtypes revealed sensitivity to targeted therapeutic agents in the different subtypes.

The BL1 subtype was found to be enriched in cell cycle and DNA damage response gene expression signatures. Patients with tumors in this subtype putatively benefit from agents that preferentially target highly proliferative tumors (e.g., use of proliferation biomarkers such as *Ki-67*, anti-mitotic and DNA-damaging agents); cisplatin and *PARP* inhibitors. BL2 on the other hand was found to be enriched in growth factor signaling and myoepithelial markers and preferentially responded to *mTOR* and growth factor inhibitors. The IM subtype enriched in immune cell signaling pathways responded to cisplatin and PARP inhibitors. The M and MSL subtypes- (with the MSL subtype displaying low expression levels of claudins), were found to be characterized with high expression of genes involved in differentiation and growth factor pathways. The mesenchymal subtypes responded to dasatinib, an *SRC* inhibitor. The LAR subtype on the other hand was driven by androgen receptor signaling and exhibited a high expression of luminal markers, *FOXA1* and *XBP1* and benefited from targeting both the AR antagonist bicalutamide and *PI3K* inhibitors (*PI3K/mTOR* inhibitor *NVP-BEZ235*) attributed to the high frequency of *PIK3CA* mutations in this subtype

[1, 4]. Based on this study, 47% of TNBCs were classified as basal-like, 17% luminal A, 12% normal breast-like, 6% luminal B, 6% *HER2*, and 12% were unclassified. These findings further revealed that not all TNBCs are basal like.

Besides exhibiting unique biology, the study conducted by Lehman et al. also portrayed distinct subtype variations in patient relapse-free survival (RFS) and distant-metastasis-free survival despite the administration of subtype preferential treatments. RFS was significantly decreased in the LAR subtype compared to other non-luminal subtypes. RFS was significantly decreased in the M subtype compared with BL1 and IM subtypes, while the MSL subtype had higher RFS than the M subtype. Distant-metastasis-free survival (DMFS) did not vary between TNBC subtypes. The M and MSL subtypes differed clinically, with patients in the M subtype presenting with shorter RFS. These findings suggest that patient outcomes are strongly correlated with their tumor composition or subtype.

Despite these salient findings on TNBC and its heterogeniety, the analysis of IHC-confirmed *ER*, *PR* and *HER2* expression tumors in Lehman et al.'s study led to the observation of only 5 of the 6 gene expression subtypes. Hinged on this limitation, Burstein and colleagues [28], conducted RNA and DNA profiling analyses on 198 tumors, revealing four distinct and stable TNBC subtypes: (1) Luminal-AR (LAR); 2) Mesenchymal (MES); 3) Basal-Like Immune-Suppressed (BLIS), and 4) Basal-Like Immune-Activated (BLIA). Like the 6 gene expression subtypes from Lehman et al.'s study, each of the 4 subtypes identified in Burstein et al.'s study showed distinct molecular profiles with distinct prognoses, with the BLIS tumors having the worst prognosis while the BLIA tumors had the best. Subtype-specific targets included androgen receptor and the cell surface mucin MUC1 in the LAR subtype; growth factor receptors (*PDGF* receptor A; *c-Kit*) in the MES subtype; an immune suppressing molecule (VTCN1) in BLIS; and Stat signal transduction molecules and cytokines in the BLIA subtype [28].

In contrast to the therapies identified for the 6 gene expression subtypes, Burstein et al.'s study [28] suggests the application of *MUC1* and *AR* antagonists in the treatment of *AR*- and *MUC1*-overexpressing LAR tumors; MES tumors would preferentially respond to beta blockers, *IGF inhibitors*, or *PDGFR* inhibitors. BLIS tumors would benefit from immune-based strategies (e.g., PD1 or VTCN1 antibodies) while STAT inhibitors, cytokine or cytokine receptor antibodies, or ipilumimab a CTLA4 inhibitor [29] may be effective treatments for BLIA tumors. The findings of this study suggest that analysis of TNBCs beyond gene expression profiles reveals novel TNBC subtype-specific markers that could be targeted for more effective treatment of TNBCs.

In a more recent study conducted by Lehmann et al. to further elucidate triple-negative breast cancer subtypes, the original six molecular classifications were refined into four: basal- like 1 (BL1), basal-like 2 (BL2), luminal androgen receptor (LAR) and mesenchymal (M) [30] and are shown to co-occur within given tumours when analyzed using single-cell genomics [31]. These studies show for and confirm the heterogeneity of TNBC and the treatment complications associated with it.

1.2.2.2 Driver mutations in TNBCs

Various studies have been conducted to identify the molecular portraits and sub-type specific mutations that provide a selective growth advantage and thus promote cancer development in TNBC to better understand the disease [1, 2, 4, 28, 32].

In a study conducted by The Cancer Genome Atlas (TCGA), somatic mutations in *TP53*, *PIK3CA*, and *GATA3* were identified occurring at a frequency higher than 10% [1] in primary breast cancers. *TP53* mutations (mostly nonsense and frameshift) were identified to be most prevalent in basal-like breast cancers exhibiting a *TP53* loss of function. Specific to TNBC, Shah et al. identified *TP53*, *PIK3CA*, *USH2A*, *MYO3A*, *PTEN* and *RBI* as the most frequently mutated genes in TNBC [32]. Most of the loss-of-function and gain-of-function alterations in TNBC involve genes associated with DNA damage repair and phosphatidylinositol 3-kinase *(PI3K)* signalling pathways, respectively [1]. Apart from loss of *TP53*, other alterations in DNA damage repair genes include loss of *RB1* and loss of *BRCA1* function [18]. Low PTEN protein levels have also been reported in TNBCs [4]. *FGFR2*, *MAPK13*, *SRC* family, *MUC* family, and the *BCL2* family are another set of hyper-activated genes identified from the exploration of TNBC genomic profiles in which they are revealed to exhibit higher expression levels, more copy number changes (most characterized by loss of 5q and 10q) [1, 11], lower DNA methylation levels (also in concordance with the TCGA study [1]), or seen as targets of miRNAs with lower expression in TNBC than in normal samples. *EGFR* is also reported to be upregulated in approximately 60% of basal-like TNBCs [13]. A further review of the 6 TNBC subtypes identified by Lehman et al.[4] revealed higher mutation rates in basal-like cancers however with less diversity [18]. This finding suggests that the high mutation rate of a gene significantly contributes to fueling a disease as opposed to the diversity and recurrence of mutated genes in the genome. Combining CNA, and mutation data with expression data also implicated well known oncogenes and tumour suppressors: *TP53*, *PIK3CA*, *NRAS*, *EGFR*, *RB1*, *ATM*. *PARK2*, *RB1*, *PTEN* and *EGFR* were the most frequently observed copy number events that mostly belonged to the BL2 subtype that is heavily enriched for growth factor signaling pathways [32]. These genes are suggested to be potential targets for TNBC treatment [33].

1.2.2.3 "BRCAness" in TNBC

Approximately 10 – 20% of TNBC patients harbour germline BRCA mutations. Even in wild-type BRCA patients, somatic mutations of the homologous recombination (HR) pathway can produce a similar phenotype termed "BRCAness" [27, 34].

Tumours with known *BRCA1* and *BRCA2* mutations display phenotypes that correlate with the basal-like subtype [4, 35], a subtype that is also characterized by genomic instability [36]. In particular though, TNBCs exhibit gene expression profiles similar to those of BRCA1-deficient tumors [15], inheriting the increased sensitivity to genotoxic agents exhibited by BRCA [37]. Sporadic basal-like breast tumors and tumors arising in BRCA1 carriers possess a similar etiology, they are both likely to be of high grade, both express basal keratins, they are ER/PR-negative, HER2-negative and have a high frequency of *TP53* mutations [18]. Other hallmarks of "BRCAness" include, *EGFR* expression, *c-MYC* amplification, loss of *RAD51*-focus formation, and sensitivity to DNA-crosslinking agents [38]. To note is that the genomic instability reported in TNBCs and BRCA associated breast cancers could be as a result of deficient DNA repair and may lead to the success of some chemotherapy regimens [39]. A study conducted by Jiang et al.[40] on the predictors of chemosensitivity in TNBC revealed an RNA-based BRCA-deficient subtype that included up to 50% of TNBC tumors that appeared immune primed. It was also found that mutations that lowered the levels of functional *BRCA1* or *BRCA2* RNA were associated with significantly better survival outcomes [40].

1.2.2.4 The clonal spectrum of TNBCs

TNBCs exhibit a wide and continuous spectrum of genomic evolution that portrays a continuously varying distribution of mutation abundance among tumors [32].

By analyzing somatic mutations, copy number aberrations (CNA), gene fusions, and gene expression patterns of 104 primary TNBCs, Shah et al. revealed a mismatch in the proportion of somatic mutation abundance relative to the proportion of the genome altered by CNAs in TNBC cases, with some cases having numerous mutations but only close to 1% alterations in the genome, while other cases presented with few mutations but with notably high numbers of genomic alterations. A significant variation in the clonal composition of TNBCs was also found, with some cases presenting with few genotypes while others presented with multiple genotypes. We would expect that an increase in mutations would increase clonal frequency and that mutations in driver genes occur in the highest frequency groups, however, this was not evidenced in this study as some cases were found to have driver genes in low clonal frequency groups. 12% of cases did not have mutations in any known driver genes, further suggesting that TNBCs are mutationally heterogeneous from the outset with variations in early clonal expansion drivers. Basal-like TNBCs were also reported to present with higher clonality at diagnosis compared to non-basal TNBCs [32]. Jiang et al. further confirmed this notion in their study that also revealed an increased clonal mutational burden (more clonal tumors with a higher number of mutations per clone) in TNBC tumors that are *BRCA* deficient [40]. The pathways of the most frequently mutated genes (*TP53*, *PIK3CA*, *PTEN* - basal-like and luminal) as analysed in Shah et al.'s study were seen in high clonal frequency groups while genes in cell motility and ECM pathways (mesenchymal-like) were seen in lower clonal frequency groups and are believed to have mutations that were acquired much later [32]. The key findings of this study suggest that TNBC tumours are unique, with varying mutational content in particular pathways; they have varying numbers of implicated molecular pathways and are shaped by distinctive mutagens and biological processes that drive mutations, clonal evolution and expansion.

1.2.2.5 TNBC mutational signatures

Somatic mutations in genes that control cellular growth and division are a consequence of mutational processes such as exogenous (ultra-violet radiation and tobacco) or endogenous (age, DNA repair deficiencies) mutagenic processes that offer insights into tumour causative events. These mutational processes are linked to specific molecular lesions and subsequent repair mechanisms initiated by a cell to mitigate the damage which in-turn generates unique combinations of mutation types (signatures) that change DNA in a specific way [41, 42]. For example, endogenous processes like DNA repair deficiencies initiate point mutations and structural variations [43]; APOBEC dysregulation results in C→T substitutions [44] while C→A substitutions are reported induced by tobacco smoke [41]. These signatures can indicate which causative mechanisms are active in a patient's tumor and can reveal clinically actionable events and key features for patient stratification

[42, 45, 46].

In an effort to determine the role of genomic rearrangements as driver mutations in breast cancer, Nik-Zainal et al. identified six rearrangement signatures, 2 of which are associated with TNBC [47]. All rearrangements in Signature 1 were characterized by tandem duplications ($>$ *100kb*), evenly distributed across the genome. Cancers exhibiting this phenomenon are frequently *TP53* mutated. Signature 3 was characterized by tandem duplications ($<10kb$) and most of the cancers (91%) with *BRCA1* mutations or promoter hypermethylation were found in this group, a group also enriched for basal-like TNBCs. Signature 5, characterized by deletions ($<100kb$) was strongly associated with the presence of *BRCA1* mutations or promoter hypermethylation, *BRCA2* mutations and with rearrangement signature 1 large tandem deletions. These events were also revealed to be evenly distributed across the genome. Signature 2 on the other hand was characterized by deletions ($>100kb$), inversions and interchromosamal translocations and contains components implicated in kataegis-focal base substitutions and APOBEC DNA-editing proteins. Signature 4 was characterized by interchromosomal translocations while signature 6 was characterized by inversions and deletions.

In Nik-Zainal et al.'s study, cancers without identifiable mutations of *BRCA1/2* or *BRCA1* promoter methylation showed similar features with those of *BRCA1/2*. This implies that either the *BRCA1* mutations might have been missed or other mutated or promoter methylated genes may be exerting similar effects [47]. Based on this observation, combinations of base substitutions, indel and rearrangement mutational signatures may be better biomarkers of defective homologous recombination of DNA double strand break repair and better biomarkers of responsiveness to cisplatin and *PARP* inhibitors other than relying on *BRCA1/2* mutations/promoter methylation alone.

In a more recent and generalized study on breast and ovary somatic mutations conducted by Funell et.al to better understand mutation signatures from the perspective of DNA repair deficiency, both SNVs and SVs were used for mutation signature inference in which an age-related signature (SNV), APOBEC signature (SNV), deletion (SV), tandem duplication (SV) and HRD (SNV) signatures were identified associated with breast cancer. Unsupervised clustering of tumours revealed subgroups with mutations in *BRCA1/BRCA2* that were associated with an HRD signature. This study also revealed the salient role of mutation signatures and their application in prognostic, patient and therapeutic subgroup discovery[42].

1.2.3 Treatment of TNBC

Hormone receptors *ER*, *PR* and *HER2* (also called *ERBB2*) are known to fuel most breast cancers [4, 32] for which intense efforts have been made to identify druggable targets [4, 28]. To date, the most successful therapies for breast cancer are those that target these receptors, with the most successful being the targeting of *HER2* and *ER* in *HER2+* and *ER+* patients respectively [1]. In contrast, due to the lack of targetable receptors, TNBC patients do not benefit from hormonal therapies. Further, the lack of identification of significant genomic driver alterations in TNBC, and the degree of tumor cell heterogeneity, has limited a targeted approach to the management of TNBC. This has left surgery, radiation and chemotherapy or a combination of these therapies as the first line of treatment for TNBC patients [48]. However, more recently, research has shown the benefit for and identified certain receptors as putative targets for new therapeutic drugs as will be discussed in subsequent sections.

1.2.3.1 Surgery, radiotherapy and chemotherapy treatment in TNBC

Predominantly, local and non-invasive TNBCs are treated with surgery. This is done with or without radiation to eliminate residual disease and reduce recurrence. Studies have shown that the younger age, higher grade or biological aggressiveness of a patient's disease does not impact surgical treatment choices; that is mastectomy vs lumpectomy, with the surgical choice mostly done based on clinicopathological variables and patient preferences [49]. Current surgical approaches however advocate for breast-conservative surgery (BCS) followed by radiation as opposed to mastectomy - a more radical procedure given that both are associated with equivalent survival rates with CBS further reducing surgical complications [50]. TNBCs are reported to be appropriate candidates for breast-conservative surgery as the local recurrence rate after surgery is not as high as that of other breast cancer subtypes [51]. However, this remains controversial as some research teams suggest that BCS followed by radiation therapy in early stage TNBC is not equivalent to mastectomy given the rapid growth and locally aggressive nature of TNBCs [52]. Secondly, given that some TNBCs harbour mutations in *BRCA1*, these tumors are deficient in double-strand DNA break repair by homologous recombination and are potentially highly radiosensitive[53]. Given the complex nature of TNBC and most cancers in general, more systematic treatment options that go beyond surgery and radiotherapy have been applied towards effective treatment of patients.

Among the systematic approaches applied in the treatment of TNBC has been the application of chemotherapy that combines the use of drugs with surgery and radiotherapy. Currently, the most common chemotherapeutic regimens include anthracyclineetaxane chemotherapy (either in the neo-adjuvant or adjuvant setting)[54]. Compared to estrogen receptor positive tumors, TNBCs have shown a higher response rate to neoadjuvant therapy [55], however, there is a higher risk of recurrence in patients who do not achieve pathological complete response. In such cases with metastatic TNBC (mTNBC)), the only available strategy is the re-administration of systemic chemotherapy; unfortunately, this approach is limited by poor response, toxicity and eventual multi-drug resistance. Chemotherapy works by impairing proliferation. It elicits a selective effect on cells that divide rapidly. As chemotherapeutic cytotoxicity is highly proliferative and non-exclusive to cancer cells, normal cells are affected too. This results in undesirable side effects such as anemia, alopecia, fever, mucositis, myelosuppression and immunosuppression [56]. Residual disease is also associated with a poorer prognosis compared to other types of residual breast cancer.

Our evolving and improved understanding of the underpinnings and molecular biology of TNBC is beginning to shed more light on possible and effective theraputic modalities and has led to the discovery of new agents that target specific pathways in TNBC as will be discussed in the subsequent section.

1.2.3.2 Emerging therapeutic modalities in TNBC

Recently, the application of massively parallel sequencing and other ‘omics’ technologies for genomic analysis has begun to reveal molecular alterations and potentially actionable features such as *BRCA1/2* mutations ("BRCAness") and the presence of the androgen receptor in some TNBC subtypes. This has allowed for the discovery of targeted therapies that could be included into clinical trials to improve patient outomces. Various therapeutic modalities (agents targeting some component(s) of the signalling cascades) active in TNBC like *PARP*, *Src*, *EGFR* and *VEGF* inhibitors, have been proposed and identified to putatively benefit TNBC patients.

Poly (ADP-ribose) polymerase inhibitors (PARPi): As earlier mentioned, about 10 - 20% of TNBCs harbor mutations in *BRCA1/2*; genes that are pivotal for genomic stability and regulation of DNA damage repair and maintenance. Cells in tumors with loss of *BRCA1* or *BRCA2* function are deficient for homologous recombination DNA repair mechanisms, hence TNBCs preferentially respond to DNA damaging agents such as *PARP* inhibitors that catalyze the fusion of components needed for alternative pathways of DNA repair (Fig. 1.4), recognize DNA damage and facilitate DNA repair to maintain genomic stability [57, 58].

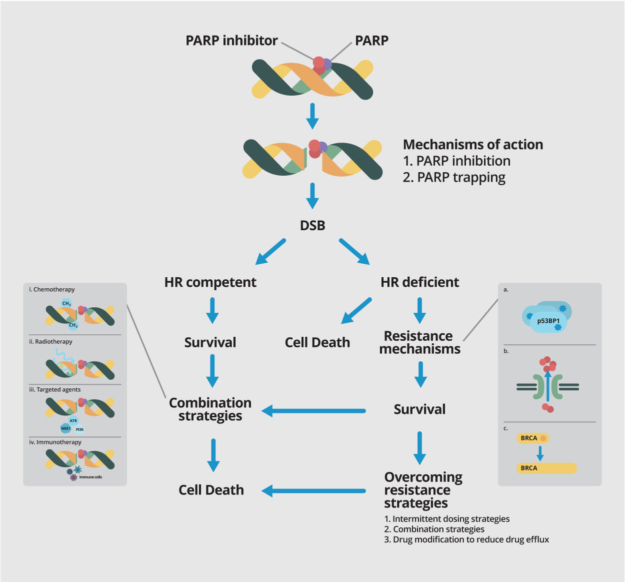


Figure 1.4: *PARP* inhibitors: DNA double strand break caused by *PARPi* via 2 mechanisms of action: inhibition of *PARP* enzyme activity and *PARP* trapping. In HR competent tumors, tumor cells with homologous recombination repair survive while in HR deficient cancers, blockade of this pathway by *PARP* inhibition leads to synthetic lethality and cell death. Source: Lim et.al [59].

In research done by Carey et al., it was hypothesized that "*PARP* inhibition, in conjunction with the loss of DNA repair via *BRCA*-dependent mechanisms, would result in synthetic lethality and augmented cell death" however, identifying patients most likely to respond to *PARP* inhibitors is still a challenge [60]. Poly-ADP ribose polymerase inhibitors – (mono-therapies: veliparib and olaparib) are currently in clinical trials and have shown improved overall response rates when combined with chemotherapy [27].

Other emerging therapeutic modalities in TNBC: Somatic mutations involving the **Epidermal Growth Factor Receptor *(EGFR)*** lead to its activation which subsequently produces uncontrolled cell division, proliferation, epithelial-mesenchymal transition (EMT), migration, invasion and angiogenesis [61] which in-turn promote primary tumorigenesis and metastasis. *EGFR* is a promising therapeutic target for TNBC given that it has been reported expressed in approximately 89% of TNBCs [62]. **Angiogenesis inhibition** using *VEGF* (Vascular Endothelial Growth Factor) pathway inhibitors have also been involved in TNBC clinical trials due to the poorer prognosis associated with *VEGF* and its expression that is reported significantly higher in TNBCs compared to other breast cancer subtypes [63]. ***Src*** an oncoprotein and member of the family of nonreceptor tyrosine kinases is reported expressed higher in TNBC cells than in *ER+* cancer cells regulating a number of signaling pathways including but not limited to proliferation, metastasis, survival, migration, invasion, and angiogenesis [64, 65]. The combination of dasatinib, cetuximab and cisplatin have provided therapeutic promise by enhancing the inhibition of cell growth, migration and invasion [66]. A number of clinical trials using **androgen receptor** targeting therapy for the treatment of *AR*-positive TNBC are also underway as Androgen Receptor *(AR)* is reported expressed in 12 – 60% of TNBCs, particularly in the LAR-subtype [67]. *AR* inhibitors such as Tamoxifen have shown to reduce disease recurrence in *AR*-positive TNBCs [68]. Other promising treatments are a combination of *AR* inhibition + radiotherapy and the combination of antiandrogens + immunotherapy for TNBC patients that co-express *AR* and *PD-L1* [69].

Some of the above mentioned therapeutic modalities in TNBC have undergone multiple clinical trials whereas others have only been investigated in early-phase trials or tested preclinically using TNBC cell lines. Despite these efforts, no targeted therapies have been approved for TNBC. Current research has shown promising prospects in effectively treating TNBC patients, however, some clinical trials have reported no objective response to therapies such as those involving EGFR inhibitors like erlotinib and lapatinib and the monoclonal antibodies (mAbs) cetuximab and panitumumab [70, 71]. Other clinical trials such as those involving the administration of anti-angiogenic therapies, bevacizumab in combination with chemotherapy to improve pCR in a neoadjuavent setting have resulted in undesirable side effects like hypertension and cardiotoxicity [26] and high rates of high grade (3 and 4) toxicities when tyrosine kinase inhibitors (RTKIs) like Sorafenib in combination with chemotherapy are administered [72]. Secondly, the development of these targeted therapies has been hindered by the inability to define patient groups that would preferentially benefit from a particular targeted agent. There is still a dire need to understand TNBC and develop new strategies for classifying and treating TNBC patients, including recurrent and metastatic cases.

1.2.4 Whole genome profiling as a stratification tool in cancer

Recent cancer studies have implemented the use of Next Generation Sequencing (NGS) to reveal variants (SNVs and idels) that have provided insights into the genomic landscapes, molecular and genomic underpinnings and the heterogeneity existing in different cancer tumours. However, the variants revealed in these studies reside in protein-coding regions that comprise $< 1\%$ of the human genome [1, 11, 32]. Furthermore, in complex diseases like cancers, genotypic biomarkers do not provide a comprehensive representation of the biological nature of a cancer [73]. The breadth and significance of various mutation types across multiple genes affecting biological pathways relevant to cancer and their potential clinical significance remain largely unexplored.

Whole-genome sequencing (WGS) has been proven as a useful approach for in-depth analysis of the landscape of mutations occurring across the genome [73, 74] and has over the years elucidated complex mutational processes at all scales through the exploration of genomic aberrations like copy number aberrations (CNAs), structural variants (SVs), small insertions and deletions (indels), single nucleotide variants (SNVs) including the identification of intricate events like SV patterns (tandem duplications, interchromosomal translocations, foldback inversions and interstitial deletions) that represent double-strand break repair mechanisms in tumors characterized with genomic instability [73, 75].

In concurrence with the above notions, Wang et al. analyzed whole-genome point mutations and structural variation patterns of 133 ovary tumors to reveal seven subgroups within the studied ovarian cancer cohort. In this study, somatic alterations in the tumor genome of each patient were identified to include SVs, indels, CNAs, and SNVs. They then conducted hierarchical clustering of the 133 patients based on selected genomic features like the aforementioned identified genomic aberrations and mutation signatures revealing 7 distinct subgroups "(G-BC: GCT tumors with mutation signature S.BC (associated with breast cancer and medulloblastoma); E-MSI: MSI ENOC tumors characterized by mutation signature S.MMR (reflective of mismatch-repair deficiency); Mixture: HGSC, CCOC, and ENOC cases without obvious discriminant features; C-APOBEC: CCOC cases characterized by mutation signature S.APOBEC (attributed to activity of the AID/APOBEC family of cytidine deaminases); C-AGE: CCOC cases characterized by mutation signature S.AGE (associated with age at diagnosis); H-FBI: HGSC cases with high prevalence of foldback inversion SVs; H-HRD: HGSCs with prevalence of duplications or deletion rearrangements and mutation signature S.HRD (reflective of homologous recombination deficiency)"[73].

In Wang et al.'s study. structural variants and point mutations in the somatic genome were seen to provide solid, discriminant biomarkers for subgroup discovery in ovarian cancer. This ground breaking research established whole genome profiling as a useful design for patient stratification and further highlights specific genomic events as putative targetable vulnerabilities suggestive of better treatment strategies for patients. Based on the findings of this study and the notion that TNBCs are genomically similar to high-grade serous carcinoma (HGSC) - both well characterized by genomic instability, heterogeneity and mutations in BRCA1/2; can TNBCs similarly be stratified based on their whole genome profiles? We answer this question in chapter 3.

1.2.5 Databases for large scale and integrated genomic data mining and analysis

Databases have long been used as an indispensable tool in modelling and organizing vast amounts of data, not excluding biological data. Currently, though, profiling of patient genomes to infer patterns of mutations and genomic events underpinning a patient's disease heavily relies on data stored in flat files. Flat files such as bam (Binary Alignment Map), vcf (Variant Call Format), txt (text) or csv (comma-separated values) are structured to contain a collection of singular records each having atomic data with record specific fields. Each of the fields in these files is separated by delimiters such as commas, white space or tabs. The structure of these files makes them relatively quick, easy to set up and use, however, this very structure complicates tasks required to query and analyze highly relational and complexly structured genomic data coupled with data redundancy and high efforts required to access data in flat files.

The explosive growth of biological data such as data from sequencing, gene expression, annotation of features and genomic events, protein structures, and data from alignment has continuously seen the need for storing, managing and accessing data using data structures that supersede those of flat files. This rapid explosion of sequencing data is attributed to the decreasing cost to sequence whole genomes and is evidenced by the exponential increase in entries from commonly used systems such as the Catalogue of Somatic Mutations in Cancer (COSMIC), dbSNP, the Cancer Genome Atlas (TCGA), European Molecular Biology Laboratory (EMBL) and the Protein Data Bank (PDB) [76]. To this effect, data generation is no longer a bottle neck but the management and analysis of these large volumes of data is [77], which has made database management systems an attractive solution for the storage and management of genomic data. In existence are different types of structural database management systems that include types such as hierarchical, network, object-oriented and relational databases. Given the highly relational data under study, we opted for relational databases that support relational data structures between objects and are more reliable than hierarchical or network database structures.

1.2.5.1 Why database management systems?

The concept of database management systems (DMBS) and in particular relational (DBMSs) as a solution to data management started as far back as 1970 [78] and has since matured to provide faster and more accurate access to large amounts of data, excellent data integration capabilities and effective and efficient data management, sharing, storage and analysis functions. Relational databases collect data in multiple tables linked together by a common piece of data and can be arranged to support ad hoc queries. Relational databases are able to capture data of various types (numbers, strings, images, booleans, dates and time, arrays, integers, floats, characters etc.) and provide advanced data structuring capabilities that support the creation of more complex relationships between data. This has further supported storage, organization, retrieval and sharing of large data sets with the ability to facilitate data visualization. Extending these functionalities to support analysis of genome data directly within databases has allowed for reliable management of genome data and the analysis of genomic variants [79].

The application of databases for genomic data analysis also supports flexible user defined parameterization and analysis of data on the fly - directly from a database using query and programming languages such as SQL without incurring costly data exports or having to rely on precomputed results. Secondly, database query optimizers have the ability to analyse queries using statistical data characteristics in a database to determine the most efficient execution mechanism for a query to improve query performance and execution runtime. Queries used also provide the execution description that can be used for documentation purposes and reproducibility of the analysis process.

The DBMS data architecture supports for data independence where changes in the application layer or user view are immune to changes in the physical (storage) or logical (conceptual) schemas and vice versa. For instance the addition or removal of new entities, attributes, or relationships is possible without having to rewrite existing application programs or changing the inter file organization system or storage structures, storage devices or indexing strategy. DBMSs also enforce both user and system defined constraints to support data consistency, integrity and security. They possess excellent concurrency control mechanisms like Strict-Two-Phase locking (Strict 2PL) where shared locks are acquired to read a data object in a database and exclusive locks acquired when an object needs to be modified. This prevents update anomalies such as a database user process reading data that is still in the process of being updated by another concurrent user process. Defined in the ARIES recovery algorithm to avoid data loss in event of a crash, DBMs execute crash recovery protocols such as write-ahead logging where all updates must be written to stable storage before they are written to disk. The recovery algorithm also retraces all actions of a database before a crash and restores it to the state it was before a crash.

With the vast amounts of data from sequencing has come the need for better management and access methods for data from sequencing. To this effect, publicly available databases have been widely established and used to aid the access, querying and visualization of data in these repositories for example the cBioPortal [80] that fuses cancer genomics data at gene level from multiple and various studies, platforms and other databases such as the NCBI Gene database and the Human Reference Protein Database (HPRD). The cBioPortal was established to support interactive exploration, visualizing, and analysis of clinical outcomes and molecular profiling data (e.g gene expression, genetic and proteomic events) across multiple samples, genes and pathways. The portal in itself is a web service interface that supports database access and querying for the presence of specific biological events in each sample such as, gene homozygous deletions, amplifications and increased or decreased mRNA or miRNA expression) as a means to accelerate the translation of genomic data into new biological insights and therapies [80].

Another similar database is the Database of Genomic Variants that [81] consists of a front-end web application that facilitates data analysis and a back-end relational database (implemented in PostgreSQL) that supports flexible and interactive database querying for structural variations within or across multiple studies.

1.2.6 Research aims, rationale and hypotheses

Currently, TNBC clinical trials use a similar patient selection criteria, however, these trials often display surprising heterogeneity in response to treatment, survival rates, and the likelihood of recurrence and metastasis. This is attributed to the differences in prognostic factors and patient characteristics like mutations and molecular signatures, gene expression profiles and tissue and organ morphologies. To improve our understanding of TNBC and to identify potential clinically actionable events, better characterization of the genetic, molecular and clinical biomarkers of TNBC is still urgently needed. Whole genome sequencing approaches have shown to reflect specific mutational processes as targetable vulnerabilities in human cancers. However, a whole genome sequencing study in TNBC at scale to investigate genomic properties as a stratification tool has not been undertaken. Secondly, data from whole genome sequencing is often stored and management in flat file format. This format is very cumbersome and ineffective for the optimal exploration, analysis and visualization of clinical outcomes and vast amounts of genomic data from which novel insights into complex diseases such as TNBC can be generated. Hinged on the **hypotheses** that (1) TNBC patients can be stratified into distinct subgroups based on their whole genome profiles and (2) the identified TNBC subgroups exhibit distinct clinical, molecular and genomic characteristics; the **main objective** of this study was to design and develop a relational database of highly structured clinical and mutation data of a cohort of TNBC and implement the developed database to support the exploration of the genomic landscape and mutational characteristics underpinning TNBCs. The developed clinical and genomic variants database was further applied to support the comprehensive analysis of clinical and whole genome profiles of 88 TNBC patients, with a novel aim of stratifying TNBC patients into distinct genomic subgroups to improve our understanding of the disease and provide valuable insights into options for novel therapeutic modalities and the identification of patients most likely to respond to specific modalities. The results of this study will also go a long way in identifying subgroup-specific clinically actionable events, clarifying uncertain histopathological diagnosis, informing prognosis, guiding treatment options for patients and supporting the use of the genome as a potential biomarker in patient treatment.

Towards testing our hypotheses and achieving the main goal of this research, the following specific objectives were established:

Objective 1: Design and develop an object-relational database of clinical outcomes and genome-wide somatic variants extracted from whole genome sequencing data of 88 TNBCs

Objective 2: Structure output data from variant calling and analytics pipelines and implement data loaders to bulk-load the structured variants and clinical outcomes into the database

Objective 3: Apply the developed database for the exploration and analysis of the clinical data and genomic variants in the developed database with specific focus on the following:

- Conduct quality control checks and analyses on the data from whole genome sequencing
- Identify and analyse all genome-wide somatic mutations (copy number aberrations (CNAs), structural variants (SVs), insertions/deletions (indels) and single-nucleotide variants (SNVs)) in a cohort of TNBC of 88 cases and extract genomic features for patient stratification
- Identify the significantly mutated genes (SMG) in the TNBC cohort

- Identify TNBC genomic subgroups and conduct comparative subgroup analyses:
  - Compute the prevalence of mutations in each subgroup, specifically on the alterations (SNVs, CNAs, and SVs) in DNA damage repair genes and the identified SMGs
  - Investigate the association between SMGs and the genomic subgroups
  - Examine the association between the identified subgroups and clinical outcomes
  - Identify mutations that appear mutually exclusive between the identified genomic subgroups
  - Investigate the association between driver mutations and mutation signatures which stratified the genomic subgroups of the TNBC cohort

Objective 4: Build a database user interface to support interactive data access, exploration, user defined querying and analysis, interpretation and sharing of the stored genomic variants and clinical outcomes among various research groups.

1.2.6.1 Research questions the database infrastructure is intended to support:

1. Can we stratify TNBC patients using their whole genome profiles?
2. Can we identify fold-back inversion events in TNBC tumours?
3. Do mutational signatures associate with specific driver mutations?
4. Are the segregated TNBC subgroups associated with distinct clinical outcomes?

1.2.6.2 Research methods and workflow

Samples from 88 TNBC cases were collected from various facilities across Canada (British Columbia, Montreal, Alberta) and tumor/normal sample pairs subjected to whole genome sequencing using Illumina HiSeq2500 (Fig. 1.5). Patient clinical data was also collected to include but not limited to: the date of diagnosis, age at diagnosis, tumour grade, tumour size (in centimeters), node status, patient status, HER2, ER2 and PR status and survival and recurrence status. However, due to the premature data collected on the overall survival status of patients, comprehensive analyses involving overall survival were not included in this study.

Aim 1: The exponential growth of data generated from DNA sequencing has continually seen the need for optimal data management, access, analysis and visualization methods. Currently data

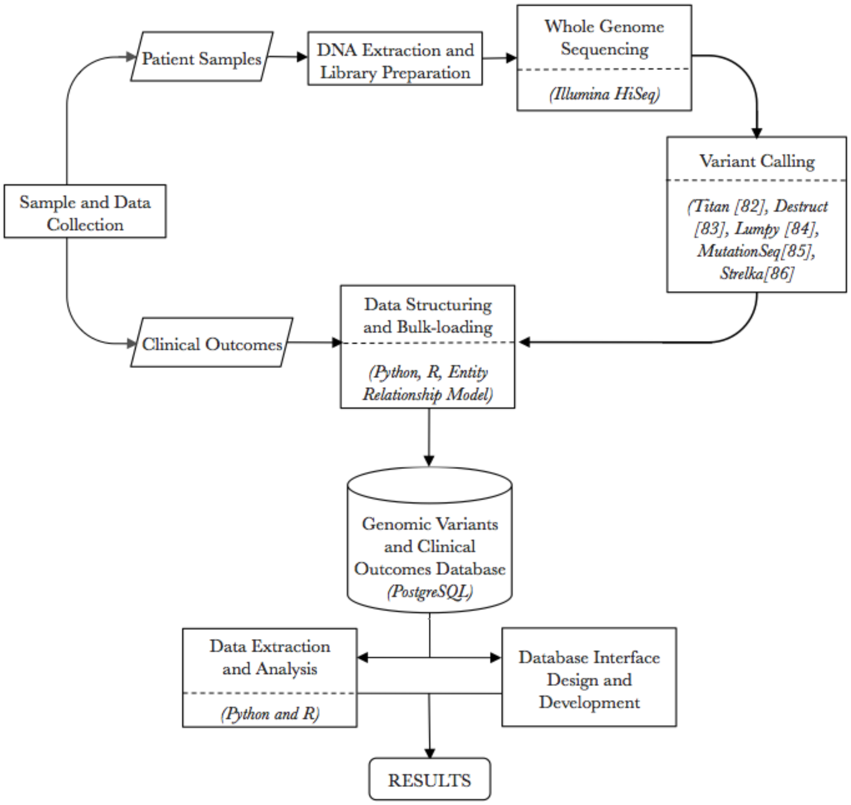

Patient Samples
DNA Extraction and Library Preparation
Whole Genome Sequencing
(Illumina HiSeq)
Variant Calling
(Titan [82], Destruct [83], Lumpy [84], MutationSeq[85], Strelka[86]
Sample and Data Collection
Clinical Outcomes
Data Structuring and Bulk-loading
(Python, R, Entity Relationship Model)
Genomic Variants and Clinical Outcomes Database
(PostgreSQL)
Data Extraction and Analysis
(Python and R)
Database Interface Design and Development
RESULTS

Figure 1.5: Research workflow.

from sequencing is often stored in flat files which are inefficient for optimal storage, querying and analysis of orthogonally collected data. To overcome these challenges, we designed and developed a relational database structure to support optimal storage, access, querying, exploration and analysis of clinical outcomes and whole genome profiling data at the level of genome-wide individual variants from the 88 TNBCs in this study cohort. Entity relationship modeling using Crow's Foot Notation was used to design the database that was implemented using PostgreSQL (psql version 10.5, server 9.4.8), an object-relational database management system (DBMS). The developed database was hosted, run and managed on a CentOS 6.5 server, with an Intel(R) Xeon(R) CPU(E5-2660v2) with a 2.20GHz base frequency (2 CPUs, 10 physical cores per CPU, 20 logical CPU units in total), 126GB of RAM and a 40GB InfiniBand connection. The choice of this DBMS (PostgreSQL) stems from its ability to hold highly relational and large datasets which are characteristic of genomic data. PostgreSQL also supports parameterized and user defined queries, custom data-types and indexes for query optimization; it is an open source DBMS that supports ACID (Atomicity, Consistency, Isolation, Durability) properties and stored procedures/SQL functions. The choice of this DBMS was also based on its interoperability and ability to support other languages such as pgSQL, python and R that were largely used in this study.

Aim 2: Genomic alterations have over time been shown to have predictive and prognostic implications in cancer patients. The discovery of all genome-wide somatic mutations was done to support the identification of putative molecular underpinnings of patients with TNBC and the potentially actionable molecular events that could provide insights into treatment options for TNBC patients. Applied were a number of various bioinformatics tools developed and assembled into an analytics pipelines to support variant calling and analyses of data from whole genome sequencing.

TITAN [82], an R Bioconductor package was used to compute cellularity and identify regions (clonal and subclonal) of copy number alterations within patient samples. To further support our analysis, gene annotations for each copy number segment was performed using pygenes a python library based on the human genome reference Homo sapiens GRCh37.73.gtf. Structural variants (SVs) including rearrangement breakpoints were predicted using deStruct [83], a tool that identifies breakpoints and assigns read alignments to the identified breakpoints. Deletions, duplications, inversions, translocations and foldback inversions were identified based on the relative position and orientation of the break-ends in the genome. Breakpoints detected by an alternative variant calling tool - Lumpy [84] were used to filter results from deStruct and to remove low mapability regions. Single nucleotide variants (SNVs) were predicted using mutationSeq [85] while the variant calling analysis for somatic SNVs and insertion/deletions (indels) was performed using Strelka [86]. The SnpEff tool was used to annotate the identified SNVs and indels for variant effects and gene-coding status. All put together, the variants identified in this cohort shed more light on the mutation patterns and signatures exhibited by different patients and patient subgroups.

Given the nature of the various tools used for variant calling, the data output from the variant calling pipelines was in disparate formats and in flat files. As earlier mentioned, this complicates data querying and processes that involve comparative data analyses. To solve this problem, the data was structured using python and R scripts that were also used to load all the structured data into the database. Also loaded in the database were statistics derived from bam files using the *Flagstat* software tool to extract bamstats and *mpileup* to extract average read coverage. These statistics were also structured and loaded into the database for further exploration.

Aim 3: We then applied the developed database to support optimal access, exploration, analysis and visualization of the mutation contents and clinical outcomes in the developed database towards answering our research questions and providing insights and a better understanding of TNBCs.

First, we used the database to conduct quality control checks and analyses on data from whole genome sequencing. Of interest was the average read coverage of tumour samples for which samples that did not meet the set threshold (60X) were excluded. We then used the database to identify and explore somatic mutations (copy number aberrations (CNAs), structural variants (SVs), insertions/deletions (indels) and single-nucleotide variants (SNVs)) by analysing mutation loads and patters across the cohort and per case. All analyses were completed using R which was both locally and remotely linked to the developed database.

MutSigCV [87] was used to identify the significantly mutated genes (SMGs) across the TNBC cohort as it has the ability to discover unexpected variations in the mutation frequency and spectrum across the genome with a unique ability to incorporate mutational heterogeneity to eliminate most of the artifactual significantly mutated genes. This enables the identification of genes truly associated with a cancer type. In this study, only genes whose false discovery rate < 0.1 were regarded as most significantly muted in this TNBC cohort. The identification of the significantly mutated genes in this cohort shed light on the subgroup putative drivers and implicated pathways that could further be probed for druggable targets. This analysis also shed light on defects co-occurring in certain pathways that may be of benefit in patient treatment for example a combination of defects in two DDR pathways leads to synthetic lethality that may be an effective therapeutic strategy for patients with such defects.

Stratification of patients is key in providing effective treatment options. To identify the genomic subgroups in this TNBC cohort, patient stratification was done based on the integration of the identified genomic features: CNAs, SVs, indels, SNVs and mutation signatures discovered using the multi-modal correlated topic model (MMCTM) [42]. Non-negative matrix factorization (NMF) approaches have been used extensively to study point mutation and structural variation signatures, however, NMF does not effectively support joint inference of signatures. MMCTM on the other hand provides an integrative approach that infers signatures using joint statistical inference from multiple mutation types like point mutations and structural variants. This further supports discovery of signatures active among patient groups as seen in the case of homologous recombination deficiency that induces patterns of both SNVs and SVs in breast and high grade serous ovarian cancers [42]. It's because of the aforementioned attributes that MMCTM was preferred in this study to support signature inference for the discovery of genomic subgroups in this TNBC cohort. All the identified stratification features were used for integrative hierarchical clustering analysis using the R package pheatmap and the manhattan distance measure to determine patient subgroups and to support the discovery of prognostic and therapeutic stratification, driver-gene associations and clinical predictions.

To further our understanding on the identified genomic subgroups, a number of comparative analyses were done. The overall mutation loads were computed and the prevalence of mutations in SMGs and DNA damage repair genes identified per subgroup. Chi-square tests were run to identify mutations that appear mutually exclusive between the subgroups. Also conducted were investigations on the association between SMGs and the genomic subgroups and the association between driver mutations and mutation signatures to provide insights into the identified subgroups and their genomic and clinical characteristics.

In most studies, mutation profiles and signatures are not routinely investigated in the clinical setting despite their salient benefit in detecting subtypes implicated in pathways that are associated with favourable prognosis [47] like those with defective mismatch repair that may benefit from immune checkpoint inhibition. In this study, the integrative analysis of the various mutation types (CNAs, SVs, SNVs and indels) with clinical data shed more light on the correlation between mutation profiles and clinical outcomes.

Aim 4: Finally, we developed a database user interface using Shiny, Plotly and JavaScript to extend the database functionality to various research groups. The developed back-end PostgresQL database was linked with the data analysis module (R) to support both local and remote data extraction, exploration, analysis and visualisation, results of which are rendered dynamically into the front-end web application for utilization by researchers, biologists and clinicians using intuitive and interactive plots and data tables. The data can be shared across individuals and research groups that also have the ability to upload files for data analysis and visualization without having to need any programming knowledge. This establishment will go a long way in helping researchers generate novel insights and hypotheses by triggering analyses and visualizations of the clinical outcomes and genomic variants data in the database.

Chapter 2

Database Design, Implementation and Optimization

The research presented herein was hinged on the application of relational databases as an indispensable tool for the exploration and analysis of tumour contents of patients in cancer studies. This chapter presents work done on the design, development and optimization of the database of clinical outcomes and genomic variants of TNBC cases in this study. Section 2.1 starts with a preliminary overview of the data structuring processes to suit database storage and downstream analysis. Section 2.2 presents the design of the variants database followed by the physical database implementation to meet data mining and data analysis functions and the methods deployed to optimize the developed database in Section 2.3.

2.1 Data structuring

The large volumes of data generated by genomic pipelines like variant calling pipelines is produced in formats such as the Variant Call Format (VCF), tsv (Tab Separated Values) or text files. As earlier mentioned in section 1.2.5, these output data files take on formats that do not support effective and efficient data mining processes. To prepare pipeline output data for database storage and further downstream analysis, the data was structured before bulk loading into the database as will be presented in the following sections.

VCF files: The Variant Call Format (VCF) is a file specification format used to store genetic variation data obtained from genomic sequencing and large-scale genotyping. It specifies a text format that contains three main sections: (1) metadata lines prefixed by "##" that describe the data values in the body of a file (Fig. 2.1). These lines describe the INFO (information), FILTER and FORMAT fields used in the body of a VCF file. (2) The header line prefixed by "#" contains 8 fixed and mandatory fields: "#CHROM POS ID REF ALT QUAL FILTER INFO". The header line also contains a "FORMAT" field and an arbitrary number of "sample ID" fields if genotype data is present in a file and finally (3) the data section that contains the variants per chromosome position for each field (Fig. 2.2).

INFO fields in the metadata lines are described as follows:

```
##INFO=<ID=PR,Number=1,Type=Float,Description="Probability of somatic mutation">
```

The above line shows the data value being captured, "PR" (probability of somatic mutation) and also shows the number of expected PR values. In this example, the number is equal to 1 (Number=1) and implies that we can only have one value for the probability of a variant call being a somatic mutation. Also included is the data type and in this case PR is of type float (Type=Float). Other data types captured in INFO fields include: integer, flag, character and string.

```
##INFO=<ID=PR,Number=1,Type=Float,Description="Probability of somatic mutation">
##INFO=<ID=TC,Number=1,Type=String,Description="Tri-nucleotide context">
##INFO=<ID=TR,Number=1,Type=String,Description="Count of tumour with reference to REF">
##INFO=<ID=TA,Number=1,Type=String,Description="Count of tumour with reference to ALT">
##INFO=<ID=NR,Number=1,Type=String,Description="Count of normal with reference to REF">
##INFO=<ID=NA,Number=1,Type=String,Description="Count of normal with reference to ALT">
##INFO=<ID=ND,Number=1,Type=String,Description="Number of Deletions">
##INFO=<ID=NI,Number=1,Type=String,Description="Number of Insertions">
##FILTER=<ID=threshold,Description="Threshold on probability of positive call">
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##SnpEffCmd="SnpEff  GRCh37.75 -noStats /shahlab/archive/sochan_tmp/jobs/SA681/temp/wgs_SA681/mutationseq/
museq.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation |
Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank |
HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance |
ERRORS / WARNINGS / INFO' ">
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this
variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_affected'">
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for
this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_gene | Percent_of_transcripts_
affected'">
##INFO=<ID=MA,Number=.,Type=String,Description="Predicted functional impact of amino-acid substitutions in
proteins.Format: (Mutation|RefGenome variant|Gene|Uniprot|Info|Uniprot variant|Func. Impact|FI score) ">
##DBSNP_DB=/shahlab/pipelines/reference/dbsnp_142.human_9606.all.vcf.gz
##INFO=<ID=DBSNP,Number=.,Type=String,Description="DBSNP flag">
##1000Gen_DB=/shahlab/pipelines/reference/1000G_release_20130502_genotypes.vcf.gz
##INFO=<ID=1000Gen,Number=.,Type=String,Description="1000Gen flag">
##Cosmic_DB=/shahlab/dgrewal/cosmic/CosmicMutantExport.sorted.vcf.gz
##INFO=<ID=Cosmic,Number=.,Type=String,Description="Cosmic flag">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
20      64871   .       C       A       7.36    PASS    PR=0.82;TR=40;TA=4;NR=28;NA=0;TC=ACA;NI=0;ND=32;
ANN=A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000382410|protein_coding|
|c.-3480C>A||||3480|,A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000608838|
processed_transcript||n.-3020C>A||||3020|,A|intergenic_region|MODIFIER|CHR_START-DEFB125|
CHR_START-ENSG00000178591|intergenic_region|CHR_START-ENSG00000178591|||n.64871C>A||||||;
MA=();DBSNP=F;1000Gen=F;Cosmic=F
20      139915  .       T       A       5.73    INDL PR=0.73;TR=53;TA=5;NR=33;NA=0;TC=CTA;NI=31;ND=1;
ANN=A|downstream_gene_variant|MODIFIER|DEFB127|ENSG00000088782|transcript|ENST00000382388|protein_coding|
|c.*250T>A||||111|,A|intergenic_region|MODIFIER|DEFB127-DEFB128|ENSG00000088782-ENSG00000185982|
intergenic_region|ENSG00000088782-ENSG00000185982|||n.139915T>A||||||;
MA=();DBSNP=[rs11471580,rs386393059,rs386393060,rs397947941];1000Gen=F;Cosmic=F
20      351395  .       G       A       17.61   PASS    PR=0.98;TR=58;TA=14;NR=38;NA=0;TC=TGT;NI=0;ND=0;
ANN=A|intergenic_region|MODIFIER|NRSN2-TRIB3|ENSG00000125841-ENSG00000101255|intergenic_region|
ENSG00000125841-ENSG00000101255|||n.351395G>A||||||;MA=();DBSNP=F;1000Gen=F;Cosmic=F
```

Figure 2.1: Variant Call Format (VCF) file structure.

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
20      64871   .       C       A       7.36    PASS    PR=0.82;TR=40;TA=4;NR=28;NA=0;TC=ACA;NI=0;ND=32;
ANN=A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000382410|protein_coding|
|c.-3480C>A||||3480|,A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000608838|
processed_transcript||n.-3020C>A||||3020|,A|intergenic_region|MODIFIER|CHR_START-DEFB125|
CHR_START-ENSG00000178591|intergenic_region|CHR_START-ENSG00000178591|||n.64871C>A||||||;
MA=();DBSNP=F;1000Gen=F;Cosmic=F
```

Figure 2.2: VCF data line.

Another key field captured in VCF files is the annotation (ANN) field (Fig. 2.3) described as shown in the INFO field in Fig. 2.4:

```
ANN=A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000382410|protein_coding|
|c.-3480C>A||||3480|,A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000608838|
processed_transcript||n.-3020C>A||||3020|,A|intergenic_region|MODIFIER|CHR_START-DEFB125|
CHR_START-ENSG00000178591|intergenic_region|CHR_START-ENSG00000178591||n.64871C>A|||||;
MA=();DBSNP=F;1000Gen=F;Cosmic=F
```

Figure 2.3: VCF annotation (ANN) field and corresponding data values.

```
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation |
Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Feature_ID | Transcript_BioType | Rank |
HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance |
ERRORS / WARNINGS / INFO' ">
```

Figure 2.4: VCF INFO field describing the annotation (ANN) field.

The metadata section also contains filters that have been applied to the data and are described as follows:

```
##FILTER=<ID=threshold,Description="Threshold on probability of positive call">
```

Fig. 2.1 shows a snippet from a VCF file of one of the TNBC samples in this study cohort. Of particular interest to the data structuring process was the decomposition of multi-valued fields like the functional annotation (ANN) field into atomic values. The annotation field contains multiple data fields for one genomic position, encoded separated by a pipe sign "|" and each annotation delimited by ";". Multiple effects (consequences) are separated by a comma as shown in Fig. 2.3. Atomizing mutli-valued fields involved decomposing and mapping each distinct functional annotation with each genomic position providing distinct variant tupples with a one-to-one mapping (one record/tupple for each REF/ALT combination) to support relational and downstream data analysis (Fig. 2.5).

Example data line extract: chr20 64871 . C A,A . . ANN=A|... , A|...

Structured output of the above line:

chr20 64871 . C A . . ANN=A|...
chr20 64871 . C A . . ANN=A|...

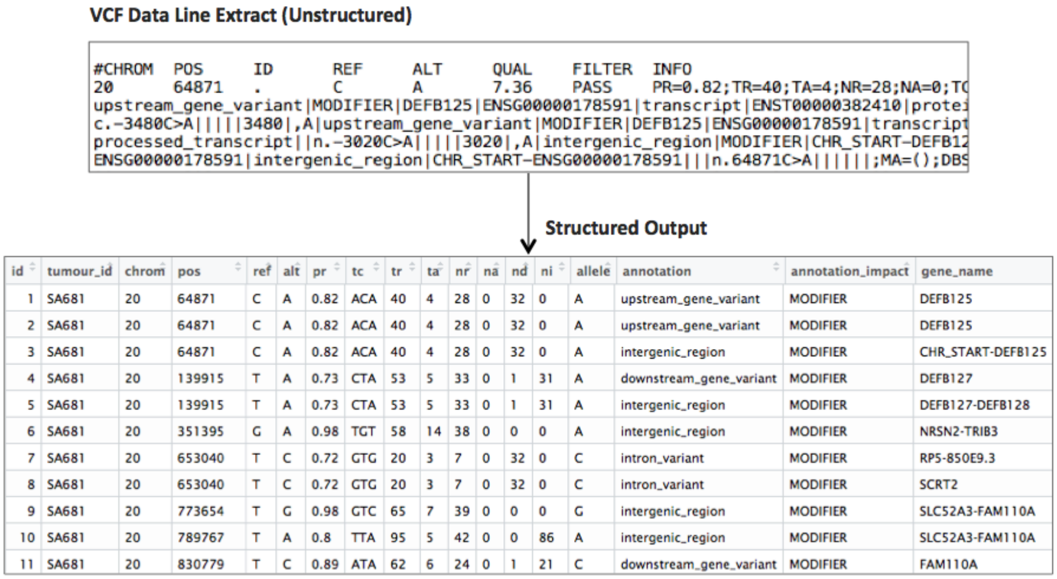
Data structuring also involved breaking down INFO fields into atomic variables.

Example line: PR=0.82;TR=40;TA=4;NR=28;NA=0;TC=ACA;NI=0;ND=32;

Structured output of the above line:

chr	pos	ref	alt	pr	tr	ta	nr	na	tc	ni	nd
20	64871	C	A	0.82	40	4	28	0	ACA	0	32

The structured data from all VCF files as called by respective variant callers (MutationSeq, Strelka and Lumpy) was then loaded into respective database tables (Fig. 2.5 and Fig. 2.6) in which fields are denoted by: tumour_id , chrom , pos , ref , alt, pr , tc , tr , ta , nr , na , nd , ni , annotation , annotation_impact , gene_name , gene_id , feature_type , feature_id , transcript_biotype , rank , hgvs_c , hgvs_p , cdna_pos_cdna_length , cds_pos_cds_length , aa_pos_aa_length , distance , errors_warnings_info , lof , nmd , ma , dbsnp , x1000gen and cosmic with each record containing one genomic event at a particular chromosomal position. Having an atomic value for each data field enabled effective data mining, querying, manipulation and analysis.



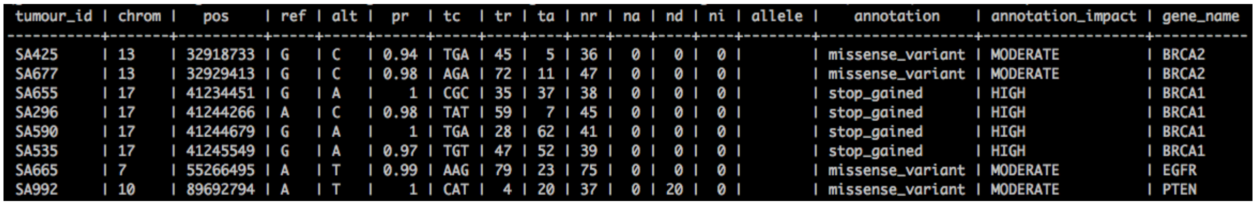
VCF Data Line Extract (Unstructured)

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
20      64871   .       C       A       7.36    PASS    PR=0.82;TR=40;TA=4;NR=28;NA=0;TC
upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript|ENST00000382410|protei
c.-3480C>A||||3480|,A|upstream_gene_variant|MODIFIER|DEFB125|ENSG00000178591|transcript
processed_transcript||n.-3020C>A||||3020|,A|intergenic_region|MODIFIER|CHR_START-DEFB12
ENSG00000178591|intergenic_region|CHR_START-ENSG00000178591|||n.64871C>A||||||;MA=();DBS
```

Structured Output

id	tumour_id	chrom	pos	ref	alt	pr	tc	tr	ta	nr	na	nd	ni	allele	annotation	annotation_impact	gene_name
1	SA681	20	64871	C	A	0.82	ACA	40	4	28	0	32	0	A	upstream_gene_variant	MODIFIER	DEFB125
2	SA681	20	64871	C	A	0.82	ACA	40	4	28	0	32	0	A	upstream_gene_variant	MODIFIER	DEFB125
3	SA681	20	64871	C	A	0.82	ACA	40	4	28	0	32	0	A	intergenic_region	MODIFIER	CHR_START-DEFB125
4	SA681	20	139915	T	A	0.73	CTA	53	5	33	0	1	31	A	downstream_gene_variant	MODIFIER	DEFB127
5	SA681	20	139915	T	A	0.73	CTA	53	5	33	0	1	31	A	intergenic_region	MODIFIER	DEFB127-DEFB128
6	SA681	20	351395	G	A	0.98	TGT	58	14	38	0	0	0	A	intergenic_region	MODIFIER	NRSN2-TRIB3
7	SA681	20	653040	T	C	0.72	GTG	20	3	7	0	32	0	C	intron_variant	MODIFIER	RP5-850E9.3
8	SA681	20	653040	T	C	0.72	GTG	20	3	7	0	32	0	C	intron_variant	MODIFIER	SCRT2
9	SA681	20	773654	T	G	0.98	GTC	65	7	39	0	0	0	G	intergenic_region	MODIFIER	SLC52A3-FAM110A
10	SA681	20	789767	T	A	0.8	TTA	95	5	42	0	0	86	A	intergenic_region	MODIFIER	SLC52A3-FAM110A
11	SA681	20	830779	T	C	0.89	ATA	62	6	24	0	1	21	C	downstream_gene_variant	MODIFIER	FAM110A

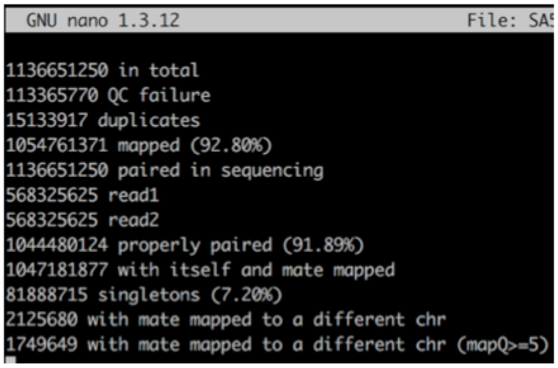
Figure 2.5: Extract of a structured VCF file: Rows denote data values captured for each variable/ field(columns).



```
 tumour_id | chrom |   pos    | ref | alt |  pr  | tc  | tr | ta | nr | na | nd | ni | allele |    annotation    | annotation_impact | gene_name
-----------+-------+----------+-----+-----+------+-----+----+----+----+----+----+----+--------+------------------+-------------------+-----------
 SA425     | 13    | 32918733 | G   | C   | 0.94 | TGA | 45 |  5 | 36 |  0 |  0 |  0 |        | missense_variant | MODERATE          | BRCA2
 SA677     | 13    | 32929413 | G   | C   | 0.98 | AGA | 72 | 11 | 47 |  0 |  0 |  0 |        | missense_variant | MODERATE          | BRCA2
 SA655     | 17    | 41234451 | G   | A   |    1 | CGC | 35 | 37 | 38 |  0 |  0 |  0 |        | stop_gained      | HIGH              | BRCA1
 SA296     | 17    | 41244266 | A   | C   | 0.98 | TAT | 59 |  7 | 45 |  0 |  0 |  0 |        | stop_gained      | HIGH              | BRCA1
 SA590     | 17    | 41244679 | G   | A   |    1 | TGA | 28 | 62 | 41 |  0 |  0 |  0 |        | stop_gained      | HIGH              | BRCA1
 SA535     | 17    | 41245549 | G   | A   | 0.97 | TGT | 47 | 52 | 39 |  0 |  0 |  0 |        | stop_gained      | HIGH              | BRCA1
 SA665     | 7     | 55266495 | A   | T   | 0.99 | AAG | 79 | 23 | 75 |  0 |  0 |  0 |        | missense_variant | MODERATE          | EGFR
 SA992     | 10    | 89692794 | A   | T   |    1 | CAT |  4 | 20 | 37 |  0 | 20 |  0 |        | missense_variant | MODERATE          | PTEN
```

Figure 2.6: Structured VCF file - sample database extract.

BAMStats Output Files: Sequence Alignment Map (SAM) and Binary Alignment Map (BAM) file formats have long been used as a standard of storage for large sequence alignments generated from genome mapping. The BAMstats software tool (Flagstat) was used to generate mapping statistics on BAM files containing sequence data to provide statistics on the total__reads, qc__failure, number of duplicate reads, number of mapped reads, mapped__percentage, paired__in__sequencing, reads (1 and 2), properly__paired, properly__paired__percentage, self__and__mate__mapped, singletons, singletons__percentage, MAPQ values, and the avg__read__coverage (Fig. 2.7).



```
1136651250 in total
113365770 QC failure
15133917 duplicates
1054761371 mapped (92.80%)
1136651250 paired in sequencing
568325625 read1
568325625 read2
1044480124 properly paired (91.89%)
1047181877 with itself and mate mapped
81888715 singletons (7.20%)
2125680 with mate mapped to a different chr
1749649 with mate mapped to a different chr (mapQ>=5)
```

Figure 2.7: Sample BAMStats output file showing row-wise data fields and values.

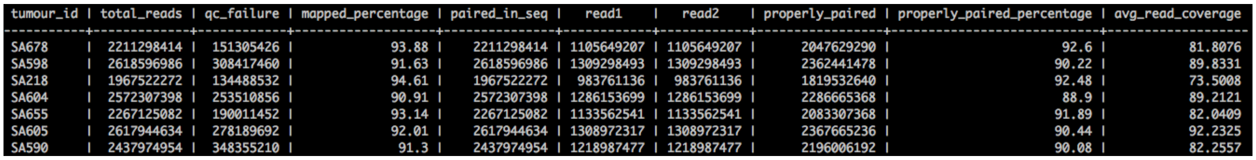
BAMStats output file data was then transformed and structured into specific fields and their respective values that were later loaded into database tables (bamstats__tumour and bamstats__normal for tumour and normal bam files respectively) for further downstream analysis (Fig. 2.8). Below are examples of data lines from BAMstats output files structured to suit database storage.

Example Line1: 1136651250 in total

Example Line2: 1044480124 properly paired (91.89%)

Structured output:

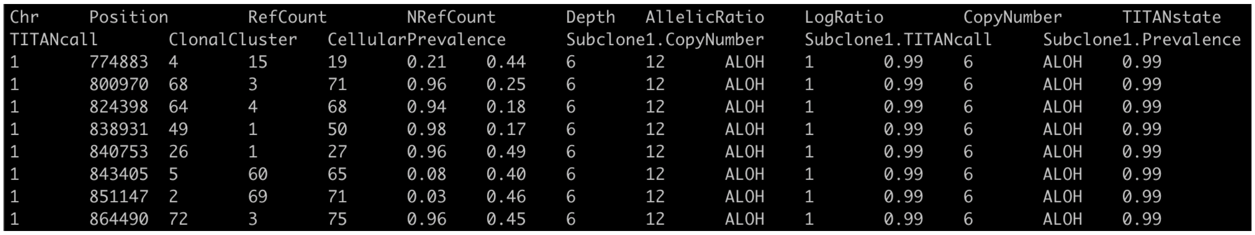
total_reads	properly_paired	properly_paired_percentage
1136651250	1044480124	91.89



```
tumour_id | total_reads | qc_failure | mapped_percentage | paired_in_seq |    read1    |    read2    | properly_paired | properly_paired_percentage | avg_read_coverage
----------+-------------+------------+-------------------+---------------+-------------+-------------+-----------------+----------------------------+-------------------
SA678     |  2211298414 |  151305426 |             93.88 |    2211298414 | 1105649207 | 1105649207 |      2047629290 |                       92.6 |           81.8076
SA598     |  2618596986 |  308417460 |             91.63 |    2618596986 | 1309298493 | 1309298493 |      2362441478 |                      90.22 |           89.8331
SA218     |  1967522272 |  134488532 |             94.61 |    1967522272 |  983761136 |  983761136 |      1819532640 |                      92.48 |           73.5008
SA604     |  2572307398 |  253510856 |             90.91 |    2572307398 | 1286153699 | 1286153699 |      2286665368 |                       88.9 |           89.2121
SA655     |  2267125082 |  190011452 |             93.14 |    2267125082 | 1133562541 | 1133562541 |      2083307368 |                      91.89 |           82.0409
SA605     |  2617944634 |  278189692 |             92.01 |    2617944634 | 1308972317 | 1308972317 |      2367665236 |                      90.44 |           92.2325
SA590     |  2437974954 |  348355210 |              91.3 |    2437974954 | 1218987477 | 1218987477 |      2196006192 |                      90.08 |           82.2557
```

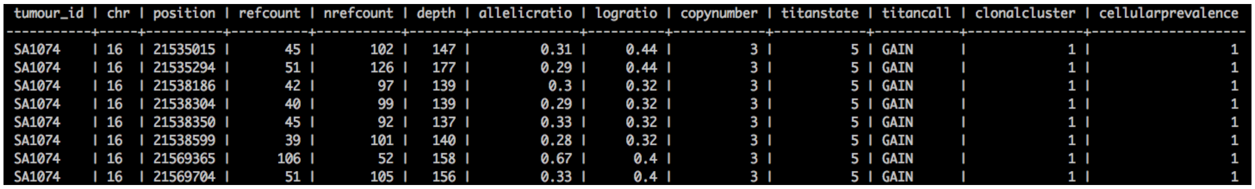
Figure 2.8: Structured BAM statistics output file - database extract of selected columns.

Text Files: Variant callers like TITAN, and deStruct provide output in form of text files (Fig. 2.9) most of which conform to a "unique_field - singular_value" data structure. Files such as these were loaded into the database as is with a few changes made to support database storage (Fig. 2.10).



```
Chr     Position        RefCount        NRefCount       Depth   AllelicRatio    LogRatio        CopyNumber      TITANstate
TITANcall       ClonalCluster   CellularPrevalence      Subclone1.CopyNumber    Subclone1.TITANcall     Subclone1.Prevalence
1       774883  4       15      19      0.21    0.44    6       12      ALOH    1       0.99    6       ALOH    0.99
1       800970  68      3       71      0.96    0.25    6       12      ALOH    1       0.99    6       ALOH    0.99
1       824398  64      4       68      0.94    0.18    6       12      ALOH    1       0.99    6       ALOH    0.99
1       838931  49      1       50      0.98    0.17    6       12      ALOH    1       0.99    6       ALOH    0.99
1       840753  26      1       27      0.96    0.49    6       12      ALOH    1       0.99    6       ALOH    0.99
1       843405  5       60      65      0.08    0.40    6       12      ALOH    1       0.99    6       ALOH    0.99
1       851147  2       69      71      0.03    0.46    6       12      ALOH    1       0.99    6       ALOH    0.99
1       864490  72      3       75      0.96    0.45    6       12      ALOH    1       0.99    6       ALOH    0.99
```

Figure 2.9: TITAN pipeline output in text file format specifying atomic values for each data field (tab delimited).



```
tumour_id | chr | position | refcount | nrefcount | depth | allelicratio | logratio | copynumber | titanstate | titancall | clonalcluster | cellularprevalence
----------+-----+----------+----------+-----------+-------+--------------+----------+------------+------------+-----------+---------------+--------------------
SA1074    | 16  | 21535015 |       45 |       102 |   147 |         0.31 |     0.44 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21535294 |       51 |       126 |   177 |         0.29 |     0.44 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21538186 |       42 |        97 |   139 |          0.3 |     0.32 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21538304 |       40 |        99 |   139 |         0.29 |     0.32 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21538350 |       45 |        92 |   137 |         0.33 |     0.32 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21538599 |       39 |       101 |   140 |         0.28 |     0.32 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21569365 |      106 |        52 |   158 |         0.67 |      0.4 |          3 |          5 | GAIN      |             1 |                  1
SA1074    | 16  | 21569704 |       51 |       105 |   156 |         0.33 |      0.4 |          3 |          5 | GAIN      |             1 |                  1
```

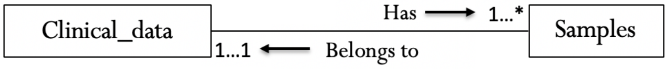
Figure 2.10: Titan output file - database extract.

2.2 Database design and development

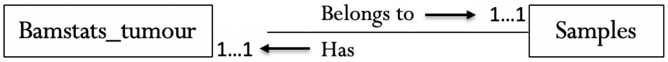
A good and well thought-out database design is a prerequisite for the development of effective and high performance databases that address efficient data manipulation, mining and analysis processes by minimizing data redundancy and the cost of running a query in terms of total execution/run time. They also enforce referential integrity and alleviate the need for data restructuring.

The key tasks undertaken to design the clinical outcomes and genomic variants database involved identifying data objects (entities represented as a logical collection of items and correspond to a table in the database), their attributes that correspond to the columns of a particular table and the relationships between the identified objects. Entity relationship diagrams (ERDs) have long been used as data models for relational databases to map out and show database entities, attributes, constraints and relationships between them. Fig.2.11 shows the created data model upon which the database was built. The entities and their description, attributes and constraints of the developed database as shown in the data model are extensively presented in a data dictionary in Appendix C.

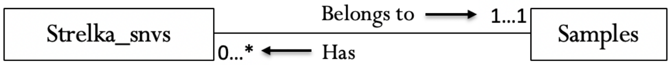
2.2.1 Relationships between entities and data constraints



The entity clinical_data is an embodiment of the cases/ patients in the TNBC cohort: one patient has one or more samples (denoted by —< in Crow's Foot Notation); a sample belongs to one and only patient (denoted by —+)



One tumour bamstats record is captured for each sample (—+) and one sample is mapped to one tumour bamstats record (—+). The relationship described here applies to the relationship between the bamstats_normal entity and the samples entity



Every single nucleotide variant (SNV) call generated by Strelka belongs to one sample (—+) and one sample can have zero to many SNV calls (—o<). The relationship described here applies to the relationship between all variants identified by other variant callers

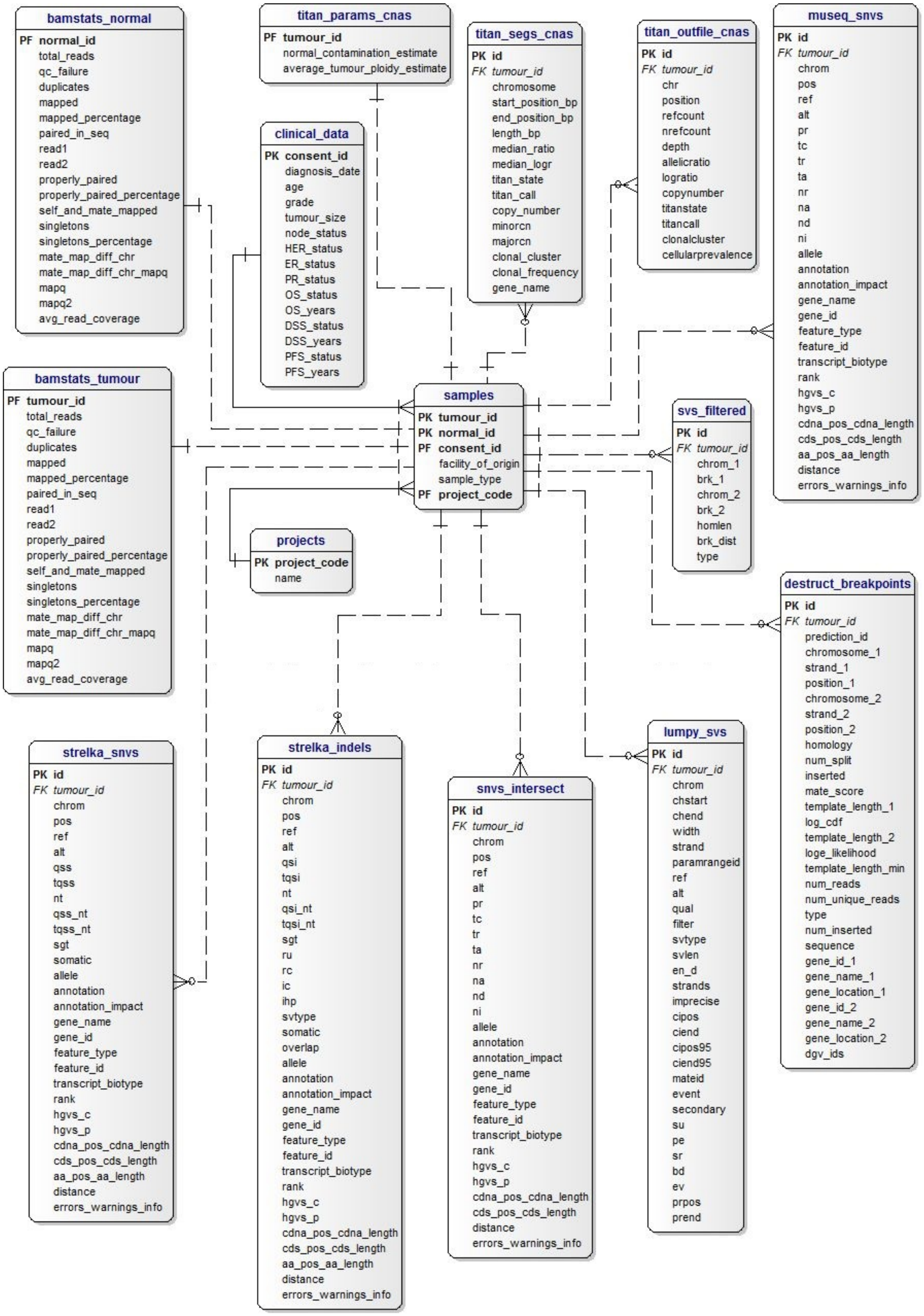


Figure 2.11: Database model (Entity Relationship Diagram (ERD)) developed using Crow's Foot Notation: Database entities are represented as boxes while relationships between entities are represented as lines. The cardinality of a relationship is represented by symbols |, -0<-, |<- that denote "one and only one", "zero to many", or "one to many" relationships respectively.

Towards creating an effective database, we imposed a number of constraints in the database design:

1. Primary keys are record identifiers that support efficient database querying. Primary key constraints were supplied to a column or group of columns to uniquely identify the rows in each database table. In cases where tables had no candidate primary keys, a sequential key was supplied. This was common for entities that contain variant data. In such entities one sample, identified by the tumour_id can have multiple genomic variant entries. This implies that in any single table that captures variants, the tumour_id is duplicated across rows making it ineligible for primary key candidacy. Apart from being unique, fields chosen as primary keys were also required not to be null.

2. Fields expected to have a data value were specified with the "NOT NULL" constraint so that a null value is not assumed. A case in point is, if data at a genomic position has been captured, we expect entries of the chromosome and position not to be null. On the other hand, a patients' ER status or tumour grade may be unknown. Such fields were left to default to "NULL" in case data on a subject was not available.

3. A unique constraint was supplied for fields whose data is expected to be unique.

4. A data type for each field was specified to validate the kind of data that can be stored in a field. As examples, date_of_diagnosis of a patient was stored with a data type 'date', the field that captures a patient's age was constrained to store integers, chromosome was constrained to store character data, position was constrained to store big integers (8 Bytes), pr (probability of somatic mutation) was constrained to store floating-point data values and annotation and annotation impact were constrained to store variable character data.

5. Referential integrity is a key feature in the design of relational databases. This constraint ensures that implied relationships between database entities are enforced hence the notion "relational database". To implement referential integrity, foreign key constraints were supplied specifying which values in a column (or a group of columns) must match the values appearing in a row of another table. The enforcement of this and other constraints mentioned in 1 - 4 above is shown in the example query below that was used to create the samples database object. The query specifies the following constraints:- "primary key" (unique by default), "not null", "foreign keys" and the "data variable types":

   ```
   cur.execute("
   ```

```
CREATE TABLE samples(tumour_id VARCHAR,
normal_id VARCHAR NOT NULL,
consent_id VARCHAR REFERENCES clinical (consent_id) ON DELETE CASCADE,
facility_of_origin VARCHAR NOT NULL,
sample_type VARCHAR NOT NULL,
project_code VARCHAR REFERENCES projects (project_code)
ON DELETE CASCADE, PRIMARY KEY (tumour_id))")
```

In the above query, the samples table is being created with 6 fields or variables whose data types such as "VARCHAR" (variable character) are shown: tumour_id, normal_id , consent_id, facility_of_origin, sample_type and project_code. The consent_id field references the consent_id field (primary key) in the clinical table and the project_code field references the project_code field in the project table. This further implies that the clinical and project tables (parent tables) be created before the samples table (child) as some of its fields reference fields in other tables. Also, in such cases, an update or deletion of a patient by consent_id in the clinical table would require a cascade update or delete of all corresponding patient records in child tables. Thus a row in a parent table cannot be deleted until all referenced rows in the child tables are deleted towards enforcing database integrity.

2.3 Database optimization

Databases have the ability to store enormous amounts of data and support the storage of millions of data entries. In this study, we opted for the deployment of a PostreSQL database that unlike other database types like MySQL or MongoDB has an unlimited database size that supports storage of as much data as required with the main constraint being system based storage constraints. PostreSQL databases also support a 32 terabyte (TB) maximum table size. With such large databases that have the ability to contain large table sizes comes the salient issue of database performance - the larger the table, the higher the cost of running a table scan in terms of total execution/ run-time and page I/Os (reads and writes of blocks containing data records to and from disk into main memory) as shown in the query plan extracts below:

Smaller table with 91 rows:

```
Aggregate   (cost=2.14..2.15 rows=1 width=0)
            (actual time=0.047..0.048 rows=1 loops=1)
->  Seq Scan on samples (cost=0.00..1.91 rows=91 width=0)
            (actual time=0.016..0.023 rows=91 loops=1)
Planning time: 0.393 ms
Execution time: 0.132 ms
```

Larger table with 157,070,329 rows:

```
Aggregate   (cost=3919120.84..3919120.85 rows=1 width=0)
            (actual time=27677.309..27677.309 rows=1 loops=1)
->  Seq Scan on titan_outfile_cnas
            (cost=0.00..3526380.07 rows=157096307 width=0)
            (actual time=0.020..17813.013 rows=157070329 loops=1)
Planning time: 0.070 ms
Execution time: 27677.332 ms
```

Given the high cost associated with working with large databases and in particular the database created herein to store mutation data, database optimization was imperative to reduce the system response time by maximizing the speed and efficiency with which data is retrieved. To optimize database performance, various optimization strategies were implemented to include: indexing, query optimization, vacuuming, partition large tables and bulk loading as will be presented in the following sections.

2.3.1 Indexing

Indexing has long been proven to be one of the most beneficial methods for optimizing database and query performance by supporting fast access to data records in associated database tables and minimizing the overall cost required to process a user query. These data structures are created using column(s) from a database table and contain a search key value and a pointer that holds the address of the disk block where a particular key value can be found.

In existence are B+ tree indexes that allow both range (e.g. 50 > age > 80) and equality (e.g. age = 60) searches, and Hash indexes that only support equality searches, the most common being B+

tree indexes as they support both search types. Well constructed indexes have a huge bearing on query optimization as these could avoid scanning an entire table for results by opting for a more efficient query plan such as an index scan that involves iterating over most or all index items when an index item meets a search condition. The required records as specified by a query are retrieved through an index whose entries (Fig. 2.12) are read left to right. In cases where all required data can be accessed through an index, there is no need for the database query optimizer to visit the much larger data table whose access consequently amounts to more page I/Os and greater run-time. Another alternative access path to a table scan is an index seek/ probe that requires searching an index for a specific value or a small set of values (fewer than those required in an index scan).

B+ tree indexes on data tables in the developed databases were created using queries like the below:

```
CREATE INDEX dest_idx1 ON destruct_breakpoints (tumour_id);
```

Where "dest_idx1" is the name of the index being created on column "tumour_id" in the database table "destruct_breakpoints".

When a new index is created, the database server automatically updates database statistics from which a query optimizer can discover the distribution of values in a column to determine the optimal execution plan for a query. The rough estimate of the number of elements within a specific range in a histogram of the query optimizer helps the optimizer decide on whether to use an index scan or a table scan for query execution.

There are two types of B+ tree indexes: clustered and non-clustered indexes, each with unique benefits depending on the data or query in question. Clustered indexes sort the data and dictate the storage order of the data records in a table (Fig. 2.12) - the order of data records is the same (or close) to the order of data entries in an index. As an example, given a B+ tree index on a column with patient ages, the ages will be ordered in ascending order, in that, ages 20 - 30 could be on one page while ages 30 - 60 on another index page. Running a query that requires retrieving ages between 50 and 60 would require reading one page into memory. This type of index is more efficient if build on columns of data that are most often accessed for ranges of values. Given that data entries are arranged in sorted order, this index type also excels at finding a specific row when the indexed value is unique.

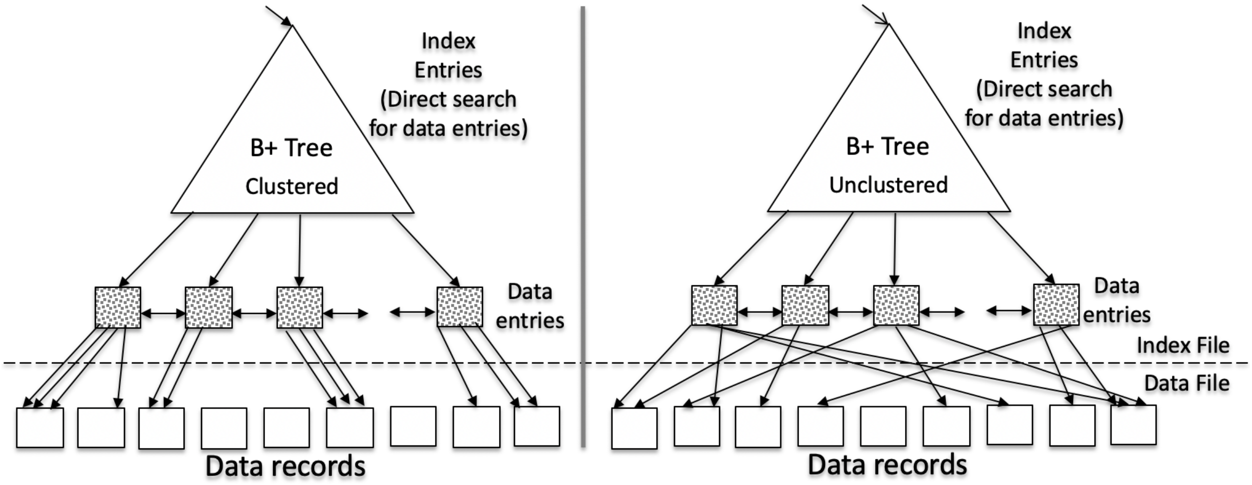


Figure 2.12: Clustered and unclusterd B+ tree index structure: A hierarchical data search structure is maintained with all searches beginning at the root of the tree to the lowest level of the tree (leaf level containing data entries). Using node pointers (separated by search key values), index entries direct searches to the correct leaf page. In clustered B+ tree indexes, node pointers to the left of a key value k point to a subtree that contains only data entries less than k and the node pointer to the right of a key value k points to a subtree that contains only data entries greater than or equal to k while unclustered indexes do no maintain this order.

In contrast, with non clustered indexes, two records that are close to each other as defined by the index might not appear on the same data page or adjacent data pages. With such indexes, there is no defined order as seen in Fig. 2.12. This implies that if we have patient ages scattered across multiple pages and we have a query that searches for patients aged between 30 and 60, we can read as many as 30+ pages into memory instead of 1!! as is the case in clustered indexes. This is because all records are on different pages that all need to be fetched into memory. Because of the high costs accrued with unclusted indexes, clustered indexes were applied in this study.

Application of clustered B+ tree indexes for optimization

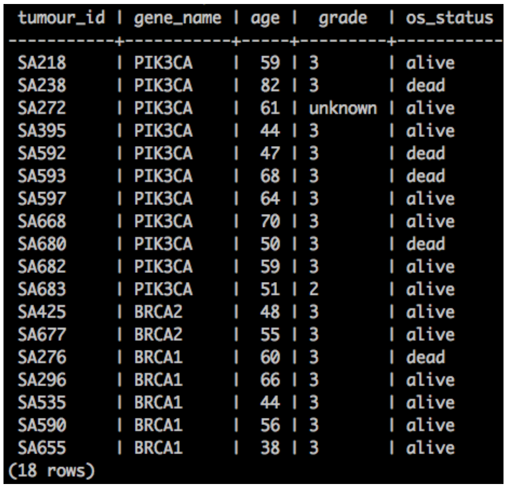
Given a database query for the tumour_id, gene_name, age, grade and overall survival status for all patients with a high impact mutation in PIK3CA, BRCA1 and BRCA2 and for which the variant was called at a probability $> = 0.9$, we could run the below query that produces the database output shown in Fig. 2.13:

```
SELECT DISTINCT s.tumour_id, i.gene_name, c.age, c.grade, c.os_status
```

```
FROM clinical_data c, snvs_intersect i, samples s
WHERE c.tumour_id = s.tumour_id and s.tumour_id = i.tumour_id
AND (i.gene_name = ‘PIK3CA’ or i.gene_name = ‘BRCA2’ or i.gene_name = ‘BRCA1’)
AND i.pr > =  0.9
AND (i.annotation_impact = ‘HIGH’ or i.annotation_impact = ‘MODERATE’) ORDER BY 2;
```

Relational Algebraic Notation of the above query:

$\pi_{\text{tumour\_id, gene\_name, age, grade, os\_status}}((\sigma_{(\text{gene\_name} = \text{‘PIK3CA’} \vee \text{gene\_name} = \text{‘BRCA2’} \vee \text{gene\_name} = \text{‘BRCA1’}) \wedge \text{pr} >=}$ snvs_intersect) $\bowtie$ samples $\bowtie$ clinical_data)



```
 tumour_id | gene_name | age |  grade  | os_status
-----------+-----------+-----+---------+-----------
 SA218     | PIK3CA    |  59 | 3       | alive
 SA238     | PIK3CA    |  82 | 3       | dead
 SA272     | PIK3CA    |  61 | unknown | alive
 SA395     | PIK3CA    |  44 | 3       | alive
 SA592     | PIK3CA    |  47 | 3       | dead
 SA593     | PIK3CA    |  68 | 3       | dead
 SA597     | PIK3CA    |  64 | 3       | alive
 SA668     | PIK3CA    |  70 | 3       | alive
 SA680     | PIK3CA    |  50 | 3       | dead
 SA682     | PIK3CA    |  59 | 3       | alive
 SA683     | PIK3CA    |  51 | 2       | alive
 SA425     | BRCA2     |  48 | 3       | alive
 SA677     | BRCA2     |  55 | 3       | alive
 SA276     | BRCA1     |  60 | 3       | dead
 SA296     | BRCA1     |  66 | 3       | alive
 SA535     | BRCA1     |  44 | 3       | alive
 SA590     | BRCA1     |  56 | 3       | alive
 SA655     | BRCA1     |  38 | 3       | alive
(18 rows)
```

Figure 2.13: Database query output.

Query tree without indexing:

The relational algebra tree in Fig. 2.14 shows the query evaluation plan of the query in question and consists of annotations at each tree node indicating the data access methods for the query. Query execution starts with a full table/file scan of the snvs_intersect table for gene_name = (‘PIK3CA’ or ‘BRCA2’ or ‘BRCA1’) and pr > = 0.9 and annotation_impact = ‘HIGH’ or ‘MODERATE’. Records that satisfy the query conditions are selected (selection denoted by sigma (σ)) and the results of this subtree query are joined by tumour_id (using a Nested Loop Join (joins denoted by a bowtie $\bowtie$) to the samples table. Using a Merge Join, the resultant subquery results are joined to the clinical_data table by tumour_id from which an overall projection (denoted by pi (π)) of the requested queried data preceded by a Hash Aggregate to select distinct records is returned. In the query tree, '$\wedge$' denotes an intersection (or 'AND') while '$\vee$' denotes union (or 'OR'). The query plan of the query and corresponding tree in Fig. 2.14 is shown in (Fig. 2.15). The total total execution time for this query is 19423.822ms.

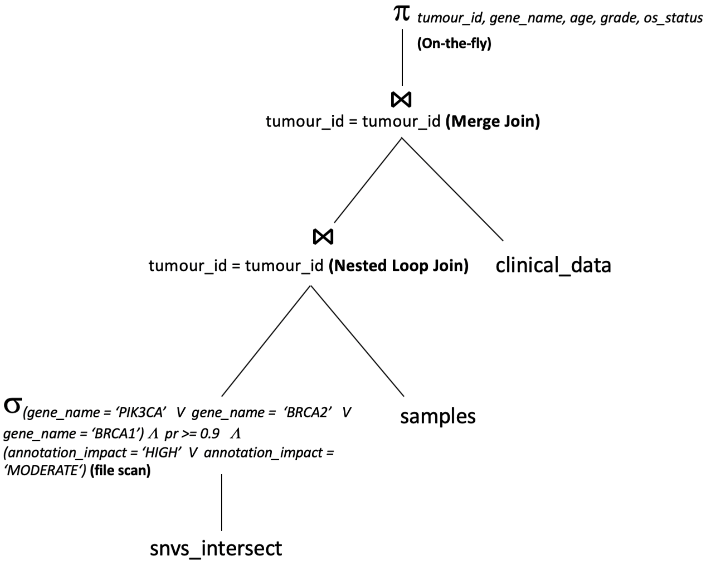
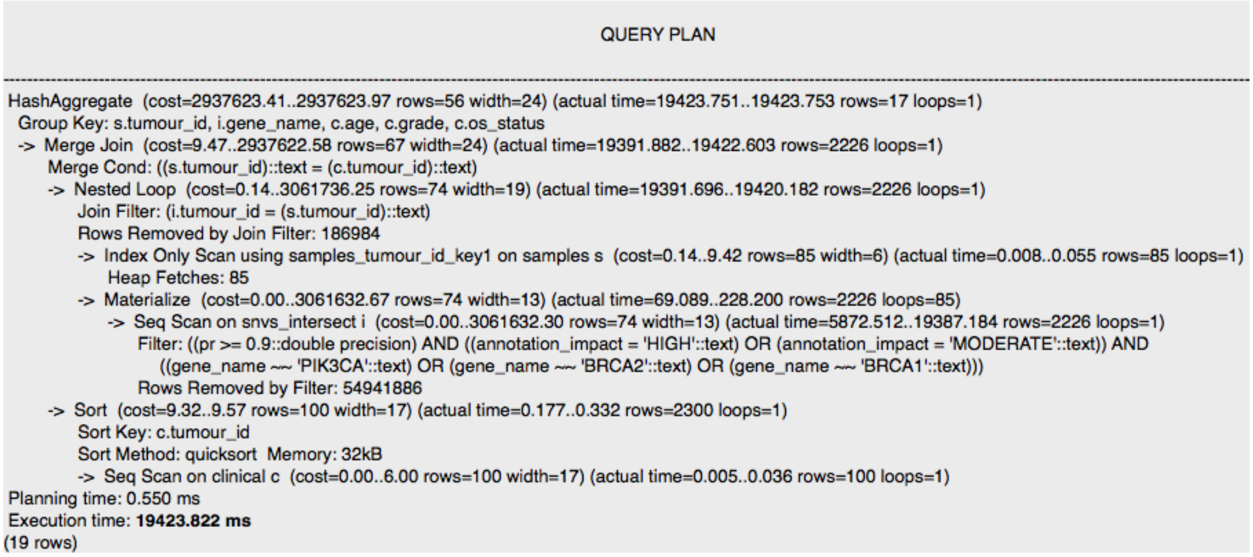

π tumour_id, gene_name, age, grade, os_status
(On-the-fly)
⋈
tumour_id = tumour_id (Merge Join)
⋈
tumour_id = tumour_id (Nested Loop Join)
clinical_data
σ(gene_name = 'PIK3CA' V gene_name = 'BRCA2' V gene_name = 'BRCA1') Λ pr >= 0.9 Λ (annotation_impact = 'HIGH' V annotation_impact = 'MODERATE') (file scan)
samples
snvs_intersect

Figure 2.14: Query tree without indexing.

Query plan without indexing:



```
                                   QUERY PLAN

-------------------------------------------------------------------------------------------------------------------------------
HashAggregate  (cost=2937623.41..2937623.97 rows=56 width=24) (actual time=19423.751..19423.753 rows=17 loops=1)
  Group Key: s.tumour_id, i.gene_name, c.age, c.grade, c.os_status
  ->  Merge Join  (cost=9.47..2937622.58 rows=67 width=24) (actual time=19391.882..19422.603 rows=2226 loops=1)
        Merge Cond: ((s.tumour_id)::text = (c.tumour_id)::text)
        ->  Nested Loop  (cost=0.14..3061736.25 rows=74 width=19) (actual time=19391.696..19420.182 rows=2226 loops=1)
              Join Filter: (i.tumour_id = (s.tumour_id)::text)
              Rows Removed by Join Filter: 186984
              ->  Index Only Scan using samples_tumour_id_key1 on samples s  (cost=0.14..9.42 rows=85 width=6) (actual time=0.008..0.055 rows=85 loops=1)
                    Heap Fetches: 85
              ->  Materialize  (cost=0.00..3061632.67 rows=74 width=13) (actual time=69.089..228.200 rows=2226 loops=85)
                    ->  Seq Scan on snvs_intersect i  (cost=0.00..3061632.30 rows=74 width=13) (actual time=5872.512..19387.184 rows=2226 loops=1)
                          Filter: ((pr >= 0.9::double precision) AND ((annotation_impact = 'HIGH'::text) OR (annotation_impact = 'MODERATE'::text)) AND
                            ((gene_name ~~ 'PIK3CA'::text) OR (gene_name ~~ 'BRCA2'::text) OR (gene_name ~~ 'BRCA1'::text)))
                          Rows Removed by Filter: 54941886
        ->  Sort  (cost=9.32..9.57 rows=100 width=17) (actual time=0.177..0.332 rows=2300 loops=1)
              Sort Key: c.tumour_id
              Sort Method: quicksort  Memory: 32kB
              ->  Seq Scan on clinical c  (cost=0.00..6.00 rows=100 width=17) (actual time=0.005..0.036 rows=100 loops=1)
Planning time: 0.550 ms
Execution time: 19423.822 ms
(19 rows)
```

Figure 2.15: Query plan without indexing (Run-time = 19423.822ms).

Applying a clustered B+ tree:

Below is a query used to create a clustered B+ tree index on columns (pr and gene_name) in the snvs_intersect database table:

```
CREATE INDEX gene_pr_idx ON snvs_intersect (pr, gene_name);
CLUSTER snvs_intersect USING gene_pr_idx;
```

With the application of a clustered B+ tree index on columns (pr and gene_name) of the snvs_intersect table, a scan on the index is done for only data entries whose values (pr and gene_name) satisfy the search conditions in the query (Fig. 2.16 and Fig. 2.17). These (fewer) data entries are then used to return only the required data as specified by the query conditions. This decreases the cost required to execute the query by avoiding a full table scan. The total execution time for this query is 876.877ms compared to 19423.822ms without an index.

Query tree with indexing:

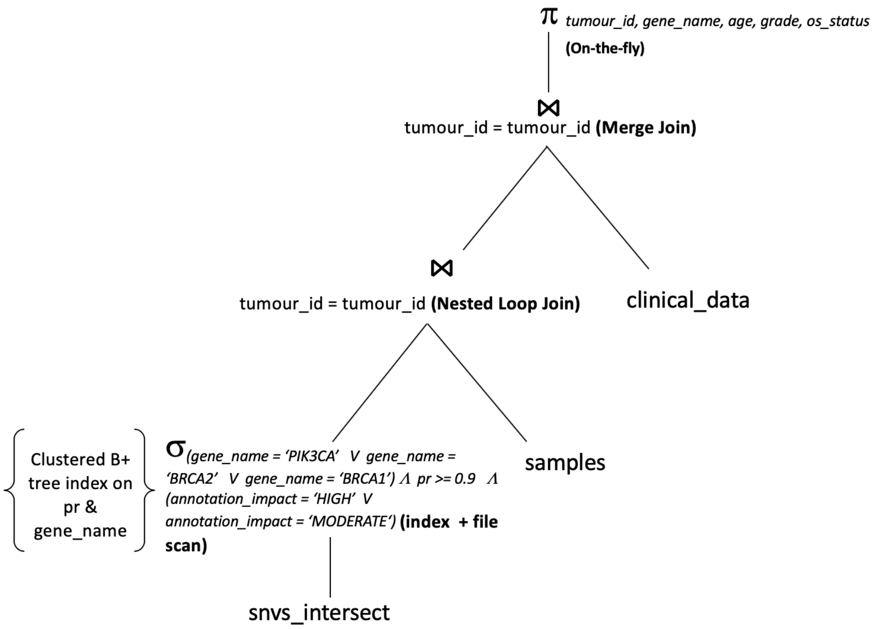
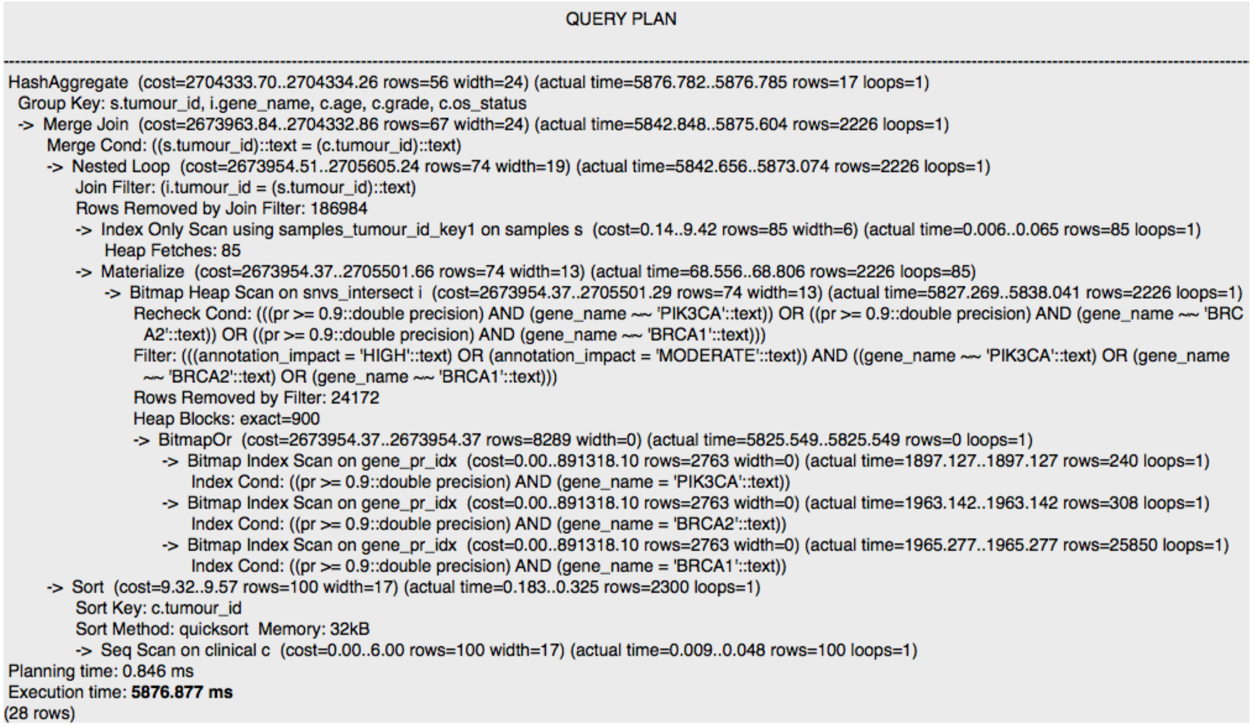

π tumour_id, gene_name, age, grade, os_status
(On-the-fly)
⋈
tumour_id = tumour_id (Merge Join)
⋈
tumour_id = tumour_id (Nested Loop Join)
clinical_data
Clustered B+ tree index on pr & gene_name
σ(gene_name = 'PIK3CA' V gene_name = 'BRCA2' V gene_name = 'BRCA1') Λ pr >= 0.9 Λ (annotation_impact = 'HIGH' V annotation_impact = 'MODERATE') (index + file scan)
samples
snvs_intersect

Figure 2.16: Query tree with indexing (Clustered B+ Tree).

Query plan with indexing:



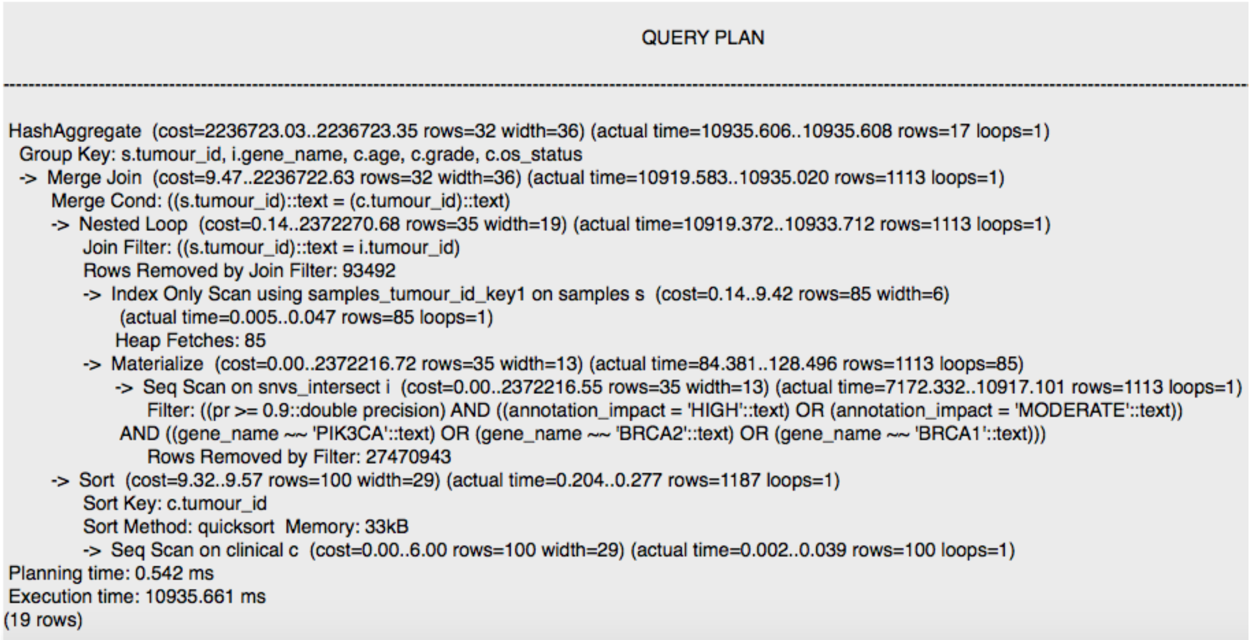
```
                                                  QUERY PLAN

------------------------------------------------------------------------------------------------------------------------------------------------
HashAggregate  (cost=2704333.70..2704334.26 rows=56 width=24) (actual time=5876.782..5876.785 rows=17 loops=1)
  Group Key: s.tumour_id, i.gene_name, c.age, c.grade, c.os_status
  ->  Merge Join  (cost=2673963.84..2704332.86 rows=67 width=24) (actual time=5842.848..5875.604 rows=2226 loops=1)
        Merge Cond: ((s.tumour_id)::text = (c.tumour_id)::text)
        ->  Nested Loop  (cost=2673954.51..2705605.24 rows=74 width=19) (actual time=5842.656..5873.074 rows=2226 loops=1)
              Join Filter: (i.tumour_id = (s.tumour_id)::text)
              Rows Removed by Join Filter: 186984
              ->  Index Only Scan using samples_tumour_id_key1 on samples s  (cost=0.14..9.42 rows=85 width=6) (actual time=0.006..0.065 rows=85 loops=1)
                    Heap Fetches: 85
              ->  Materialize  (cost=2673954.37..2705501.66 rows=74 width=13) (actual time=68.556..68.806 rows=2226 loops=85)
                    ->  Bitmap Heap Scan on snvs_intersect i  (cost=2673954.37..2705501.29 rows=74 width=13) (actual time=5827.269..5838.041 rows=2226 loops=1)
                          Recheck Cond: (((pr >= 0.9::double precision) AND (gene_name ~~ 'PIK3CA'::text)) OR ((pr >= 0.9::double precision) AND (gene_name ~~ 'BRC
A2'::text)) OR ((pr >= 0.9::double precision) AND (gene_name ~~ 'BRCA1'::text)))
                          Filter: (((annotation_impact = 'HIGH'::text) OR (annotation_impact = 'MODERATE'::text)) AND ((gene_name ~~ 'PIK3CA'::text) OR (gene_name
 ~~ 'BRCA2'::text) OR (gene_name ~~ 'BRCA1'::text)))
                          Rows Removed by Filter: 24172
                          Heap Blocks: exact=900
                          ->  BitmapOr  (cost=2673954.37..2673954.37 rows=8289 width=0) (actual time=5825.549..5825.549 rows=0 loops=1)
                                ->  Bitmap Index Scan on gene_pr_idx  (cost=0.00..891318.10 rows=2763 width=0) (actual time=1897.127..1897.127 rows=240 loops=1)
                                      Index Cond: ((pr >= 0.9::double precision) AND (gene_name = 'PIK3CA'::text))
                                ->  Bitmap Index Scan on gene_pr_idx  (cost=0.00..891318.10 rows=2763 width=0) (actual time=1963.142..1963.142 rows=308 loops=1)
                                      Index Cond: ((pr >= 0.9::double precision) AND (gene_name = 'BRCA2'::text))
                                ->  Bitmap Index Scan on gene_pr_idx  (cost=0.00..891318.10 rows=2763 width=0) (actual time=1965.277..1965.277 rows=25850 loops=1)
                                      Index Cond: ((pr >= 0.9::double precision) AND (gene_name = 'BRCA1'::text))
        ->  Sort  (cost=9.32..9.57 rows=100 width=17) (actual time=0.183..0.325 rows=2300 loops=1)
              Sort Key: c.tumour_id
              Sort Method: quicksort  Memory: 32kB
              ->  Seq Scan on clinical c  (cost=0.00..6.00 rows=100 width=17) (actual time=0.009..0.048 rows=100 loops=1)
Planning time: 0.846 ms
Execution time: 5876.877 ms
(28 rows)
```

Figure 2.17: Query plan with index (Run-time = 5876.877ms).

2.3.2 Query optimization

Besides the application of indexes to enhance performance, the construction of smart queries that leverage knowledge on database tables can also yield faster data access. Below is a differently structured query that provides the same output as seen in Fig. 2.13, however, this query has a longer execution time (10935.661 ms, Fig. 2.18) despite the created clustered index.

```
SELECT DISTINCT s.tumour_id, i.gene_name, c.age, c.grade, c.os_status
FROM clinical c
JOIN samples s on c.tumour_id = s.tumour_id
JOIN snvs_intersect i on s.tumour_id = i.tumour_id
WHERE (i.gene_name like ‘PIK3CA’ or i.gene_name like ‘BRCA2’ or
       i.gene_name like ‘BRCA1’)
AND i.pr > = 0.9
AND (i.annotation_impact = ‘HIGH’ or i.annotation_impact = ‘MODERATE’) ORDER BY 2;
```



```
                                                              QUERY PLAN
-----------------------------------------------------------------------------------------------------------------------------------------------------------
HashAggregate  (cost=2236723.03..2236723.35 rows=32 width=36) (actual time=10935.606..10935.608 rows=17 loops=1)
  Group Key: s.tumour_id, i.gene_name, c.age, c.grade, c.os_status
  ->  Merge Join  (cost=9.47..2236722.63 rows=32 width=36) (actual time=10919.583..10935.020 rows=1113 loops=1)
        Merge Cond: ((s.tumour_id)::text = (c.tumour_id)::text)
        ->  Nested Loop  (cost=0.14..2372270.68 rows=35 width=19) (actual time=10919.372..10933.712 rows=1113 loops=1)
              Join Filter: ((s.tumour_id)::text = i.tumour_id)
              Rows Removed by Join Filter: 93492
              ->  Index Only Scan using samples_tumour_id_key1 on samples s  (cost=0.14..9.42 rows=85 width=6)
                    (actual time=0.005..0.047 rows=85 loops=1)
                    Heap Fetches: 85
              ->  Materialize  (cost=0.00..2372216.72 rows=35 width=13) (actual time=84.381..128.496 rows=1113 loops=85)
                    ->  Seq Scan on snvs_intersect i  (cost=0.00..2372216.55 rows=35 width=13) (actual time=7172.332..10917.101 rows=1113 loops=1)
                          Filter: ((pr >= 0.9::double precision) AND ((annotation_impact = 'HIGH'::text) OR (annotation_impact = 'MODERATE'::text))
                    AND ((gene_name ~~ 'PIK3CA'::text) OR (gene_name ~~ 'BRCA2'::text) OR (gene_name ~~ 'BRCA1'::text)))
                          Rows Removed by Filter: 27470943
        ->  Sort  (cost=9.32..9.57 rows=100 width=29) (actual time=0.204..0.277 rows=1187 loops=1)
              Sort Key: c.tumour_id
              Sort Method: quicksort  Memory: 33kB
              ->  Seq Scan on clinical c  (cost=0.00..6.00 rows=100 width=29) (actual time=0.002..0.039 rows=100 loops=1)
 Planning time: 0.542 ms
 Execution time: 10935.661 ms
(19 rows)
```

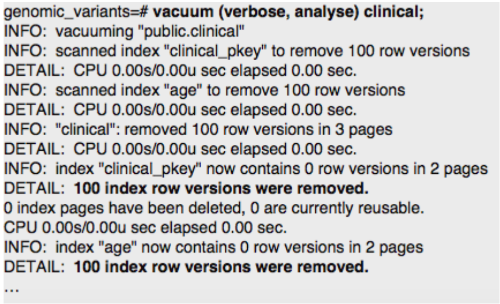
Figure 2.18: Query plan of poor performance query: Increased run-time regardless of applied indexes.

2.3.3 Re-clustering

Clustered tables are physically ordered based on the order of created clustered indexes, however, as clustering is a one time operation, subsequent table updates or inserts are not clustered. For example, if data records are ordered by probability of somatic mutation (pr), new table entries may be inserted at the end of a file whereby a new record with pr = 0.5 may be found on a page whose pr range was originally 0.8 - 1.0. With time, a table tends to be unclustered which ends up affecting performance by increasing the cost of executing a query. To avoid this, occasional reclustering was done by reissuing the same clustering command especially on updated tables.

2.3.4 Vacuuming

Vacuuming is another optimization mechanism that was used to reclaim storage occupied by dead tuples in database tables. In normal database operations, records that are deleted or obsolete by an update are not physically/completely removed from a table and keep occupying storage space until a "VACUUM" is done. Fig. 2.19 shows an example of vacuuming done on a sample table (clinical) to reclaim storage space from dead tuples not removed by "autovacuum".



```
genomic_variants=# vacuum (verbose, analyse) clinical;
INFO:  vacuuming "public.clinical"
INFO:  scanned index "clinical_pkey" to remove 100 row versions
DETAIL:  CPU 0.00s/0.00u sec elapsed 0.00 sec.
INFO:  scanned index "age" to remove 100 row versions
DETAIL:  CPU 0.00s/0.00u sec elapsed 0.00 sec.
INFO:  "clinical": removed 100 row versions in 3 pages
DETAIL:  CPU 0.00s/0.00u sec elapsed 0.00 sec.
INFO:  index "clinical_pkey" now contains 0 row versions in 2 pages
DETAIL:  100 index row versions were removed.
0 index pages have been deleted, 0 are currently reusable.
CPU 0.00s/0.00u sec elapsed 0.00 sec.
INFO:  index "age" now contains 0 row versions in 2 pages
DETAIL:  100 index row versions were removed.
...
```

Figure 2.19: Vacuuming for database optimization.

2.3.5 Bulk-loading

All data in the developed database was bulk-loaded using developed data loading scripts. This approach significantly improved performance as it is much faster than repeated inserts. Secondly records are sorted before bulk-loading. All scripts that performed data structuring of pipeline data had a database loading function to pass structured data instantly to the database as shown below.

```
#Database connection
pw <- { " "}
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "genomic_variants", host = "",
user = "", password = pw)
rm(pw)
                        .
                        .
            data structuring script
                        .
                        .
#Writing structured data to the database into table "museq_unfiltered"
dbWriteTable(con, "museq_unfiltered", museq_unfiltered, append=TRUE,
row.names=0)
```

Unlike pipeline output data that contained genomic variants, the clinical data used in this study was loaded from a .csv file as shown in the abstract script below.

```
#!/home/rasiimwe/miniconda3/bin/python
import psycopg2
import sys
import csv
import os

con = None
try:
        con = psycopg2.connect("host=‘localhost’ dbname=‘genomic_variants’
        user=‘ ’ password=‘ ’")

        cur = con.cursor()

        path="path to file"

        ## Creating table clinical_data
        ##-----------------------------------------------------------------
        cur.execute("DROP TABLE IF EXISTS clinical_data")
        cur.execute("CREATE TABLE clinical_data (...)

        ## Data Loading
        ##-----------------------------------------------------------------
        cur.execute("COPY clinical (consent_id, diagnosis_date, age, ...)
        FROM ‘%s’ delimiter ‘,’ csv header" % (path))
        ##-----------------------------------------------------------------



        con.commit()
except psycopg2.DatabaseError as e:
        if con:
                con.rollback()
        print (‘Error %s’) % e
        sys.exit(1)

finally:
        if con:
                con.close()
```

Chapter 3

Database Application to Whole Genome Profiling and Stratification of TNBCs

This chapter presents the utility of the developed database in facilitating the exploration and analysis of mutation contents and patterns in complex diseases with emphasis on understanding the genomic landscape and mutational characteristics underlying TNBCs towards TNBC subgroup discovery. Section 3.1 provides an overview of the utility of the developed database in supporting preliminary quality control (QC) checks and analyses on the data in the database. Section 3.2 presents database mining and exploratory functions to support comprehensive genome and gene-level analyses in the cohort to further support the discovery and subsequent analysis of TNBC genomic subgroups as presented in section 3.3.

3.1 Quality Control (QC)

QC checks for whole genome sequencing:- Before embarking on downstream data analysis, the database was explored for sequencing thresholds that were applied to the data during whole genome sequencing. The sequencing parameters used in this study were derived from the bam file of each sample using the SAMtools-*mpileup* utility and were thereon loaded into the database for storage and subsequent analysis. Among the data variables captured in the bam statistics data tables include ‘total_reads’, ‘qc_failure’, ‘duplicates’, ‘mapped’, ‘mapped_percentage’, ‘paired_in_seq’, ‘read1’, ‘read2’, ‘properly_paired’, ‘properly_paired_percentage’, ‘self_and_mate_mapped’, ‘singletons’, ‘singletons_percentage’ and ‘avg_read_coverage’. Of interest to our study was the average read coverage used for whole genome sequencing (Fig. 3.1) and the percentage of mapped and properly paired reads (Fig. 3.2 and Fig. 3.3 respectively). For effective and high confidence variant discovery, the established average read coverage threshold in this study was 60X. All samples that did not meet this threshold were flagged for higher resequencing coverage and excluded from further downstream analyses. Queries such as the below were used to extract data used to check sequencing parameters:

```
stats.tumour <- dbGetQuery(con,
                           "SELECT tumour_id, mapped_percentage, properly_paired,
                            avg_read_coverage
                            FROM bamstats_tumour
                            ORDER BY 3 DESC)")
```

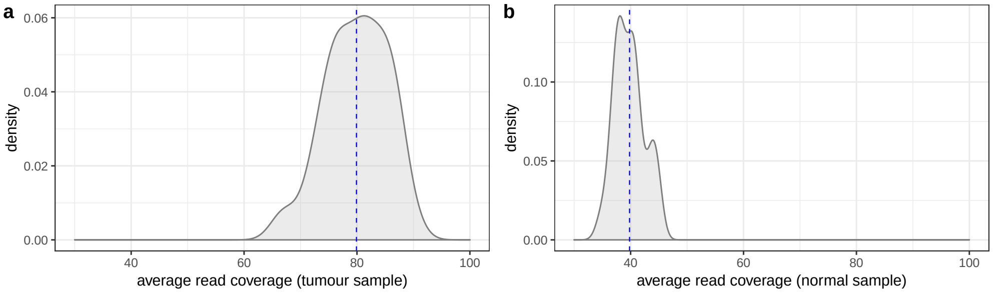


Figure 3.1: Average read coverage: **(a)** Tumour samples: mean = 79.91X, range = 66.88X - 89.83X. **(b)** Normal samples: mean = 39.81X, range = 34.80X - 45.07X.

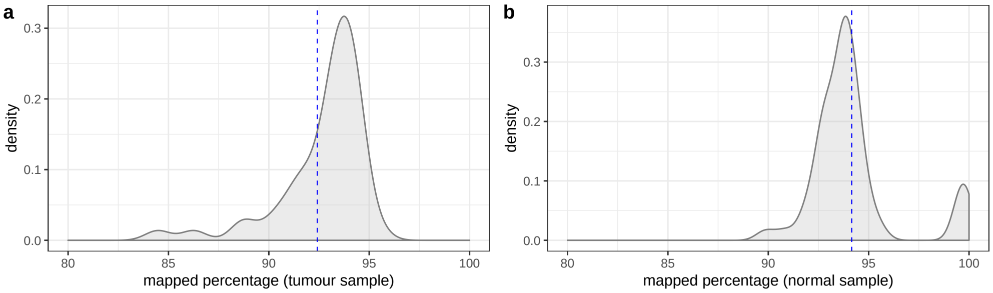


Figure 3.2: Percentage of mapped reads: **(a)** Tumour samples: mean = 92.41%, range = 78.83% - 95.44%. **(b)** Normal samples: mean = 94.16%, range = 89.85% - 99.77%.

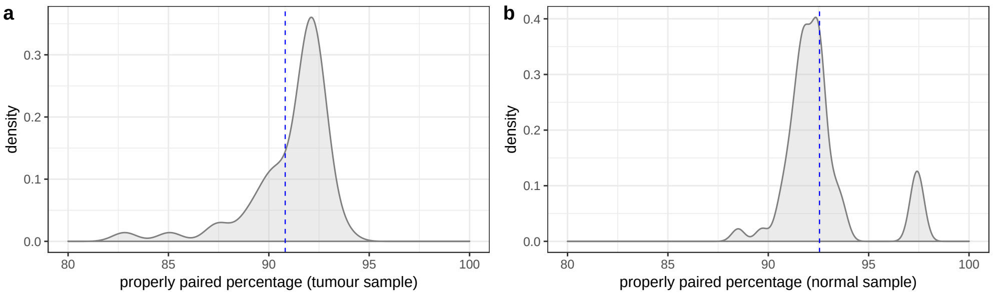


Figure 3.3: Percentage of properly paired reads: **(a)** Tumour samples: mean = 90.81%, range = 77.42% - 93.74%. **(b)** Normal samples: mean = 92.55%, range = 88.50% - 97.58%.

QC checks for normal contamination levels:- In various genomic studies, sequencing of matched tumor and normal samples has become a conventional study design to distinguish between somatic and germline variants towards supporting reliable detection of somatic mutations. Tumor-normal sample contamination causes decreased sensitivity in mutation detection that could result in inaccurate sequencing data [88]. The detection of normal contamination estimates in this study was derived from TITAN output data [82]. From the perspective of copy number inference, the exploration and analysis of genomic allelic imbalances and loss of heterozygosity events as derived from allelic ratio data *(RefCount/Depth)* is significantly influenced by the proportion of the normal content in a tumour sample *(tumour content = 1 - (normal contamination estimate))* [82]. Fig. 3.4 presents database derived normal contamination estimates of the samples in this cohort, all of which were rendered viable for subsequent downstream analysis. To note is that 7 out of 11 samples with no normal contamination are patient-derived xenografts (PDX).

QC implementations for genomic variant data:- In our study, the identification of genome-wide somatic mutations was executed using the Kronos workflow assembler [89] that was used to run TITAN [82] to infer copy number aberrations and loss of heterozygosity (LOH) events in each patient_tumour sample(s), deStruct [83] and Lumpy [84] to infer structural variants (SVs), mutationSeq [85] to infer single nucleotide variants (SNVs) and Strelka [86] to infer both indel and single nucleotide variants. All QC implementations by the various pipeline tools were applied to the TNBC whole genome sequencing data. To maintain high confidence calls, further downstream quality control involved intersecting SVs inferred by deStruct and Lumpy and removing those with breakpoints falling in low-mapability regions. SNVs called by both mutationSeq and Strelka were also intersected and variants for which the probability of somatic mutation *(pr) >= 0.9* were used in all subsequent study analyses. Given that databases support computation of results on demand,

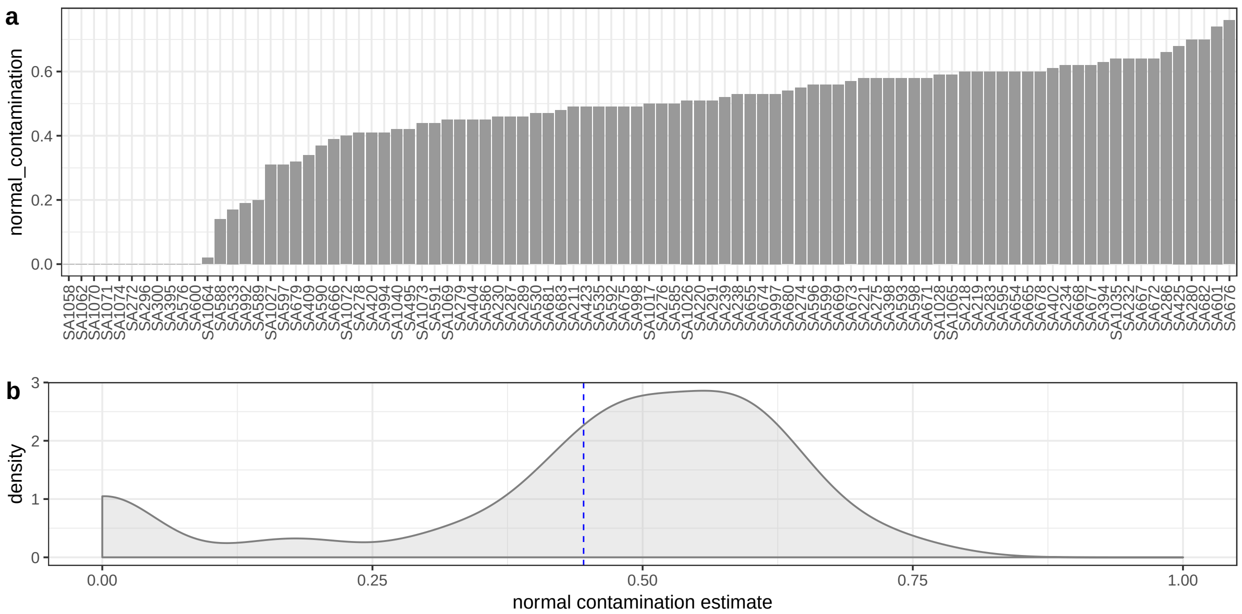


Figure 3.4: Normal contamination estimates: **(a)** Proportion of the normal content in a tumour sample *(tumour content = 1 - (normal contamination estimate))*, mean = 0.45, range = 0 - 0.76. **(b)** Corresponding density plot showing the distribution of the normal contamination estimate in this cohort.

pr thresholds were directly applied during database query time allowing for flexibility in setting thresholds for data analysis. Below is an example query that returns results based on user defined parameters.

```
gene <- "BRCA1"
effect <- "stop_gained"
pr.pass <- 0.9
query <- fn$identity("
                    SELECT DISTINCT tumour_id, gene_name, annotation
                    FROM strelka_indels
                    WHERE gene_name = ‘$gene’ AND annotation like ‘$effect’
                    UNION SELECT distinct tumour_id, gene_name, annotation
                    FROM snvs_intersect
                    WHERE gene_name = ‘$gene’ AND annotation like ‘$effect’
                    AND pr >= $pr.pass ORDER BY 1 ASC")
data <- dbGetQuery(con, query)
```

Output of the above query:

```
 tumour_id | gene_name | annotation
-----------+-----------+-------------
 SA296     | BRCA1     | stop_gained
 SA535     | BRCA1     | stop_gained
 SA590     | BRCA1     | stop_gained
 SA655     | BRCA1     | stop_gained
```

3.2 Somatic aberrations characteristic of TNBC

Somatic aberrations (CNAs, SVs, SNVs and indels) present in the tumor genome of each patient were discovered using the aforementioned variant calling tools:- TITAN, deStruct and Lumpy, mutationSeq and Strelka (snvs) and strelka (indels) respectively. Database driven explorations and analyses conducted on the identified somatic mutations to infer patient specific and cohort-wide mutation loads, patterns and characteristics are presented in the following sections.

3.2.1 Distribution of mutation loads per sample and across the cohort

The distribution of mutation loads in this TNBC cohort as depicted in Fig. 3.5 shows a varying distribution of mutation loads among TNBC cases with variations seen across the cohort and in the mutation-type loads in each sample. Some key questions to ask here would be whether mutation loads have a bearing on survival outcomes and patient stratification and whether cases with higher mutation burdens associate with higher levels of genomic instability. We answer these questions in section 3.3.

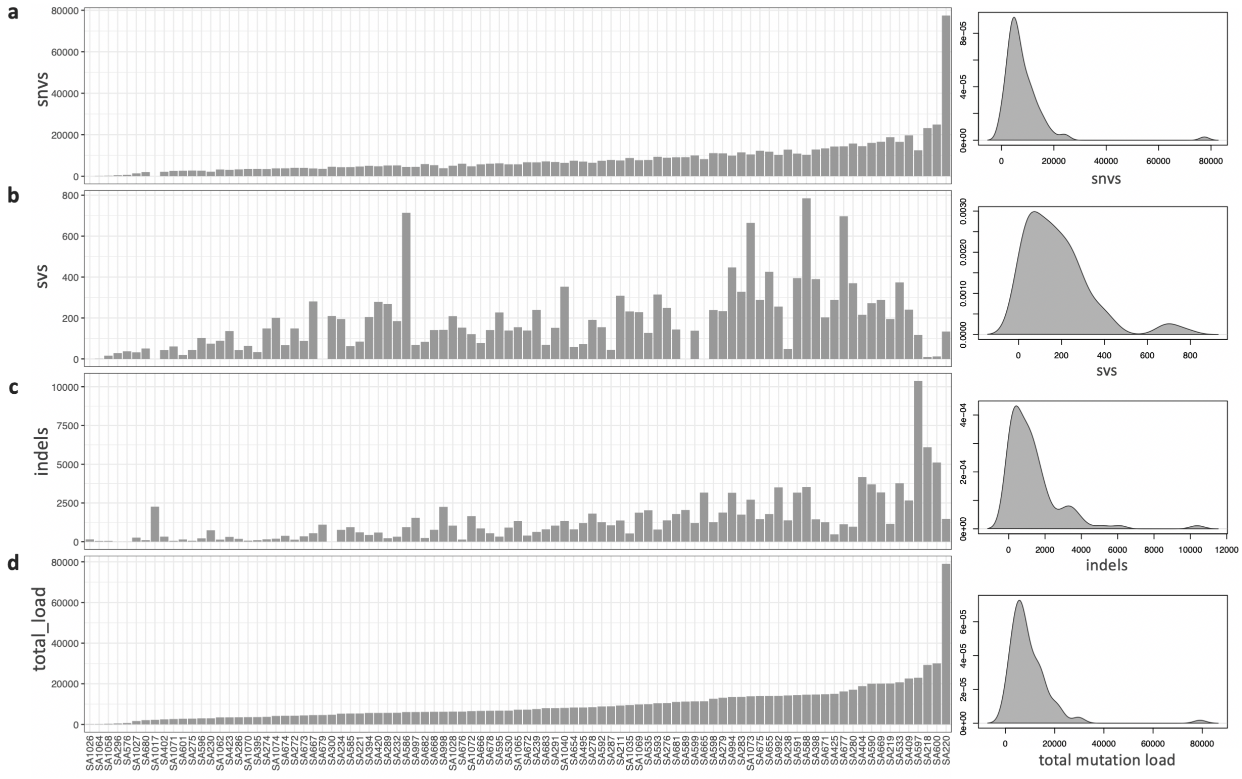


Figure 3.5: Distribution of mutation loads: Track 1 **(a)** shows the number of SNV mutations (y-axis) for each sample (x-axis), *(mean = 8172.62, range = 0 - 77463)*. **b)** Shows the number of SVs for each sample *(mean = 180.70, range = 0 - 785)*, **c)** shows the number of indels for each sample *(mean = 1327.41, range = 1 - 10373)* and **d)** shows the total mutation load for each of the samples *(mean = 9680.74, range = 149 - 79075).* Samples are sorted in ascending order based on the total mutation load.

3.2.2 Structural variants

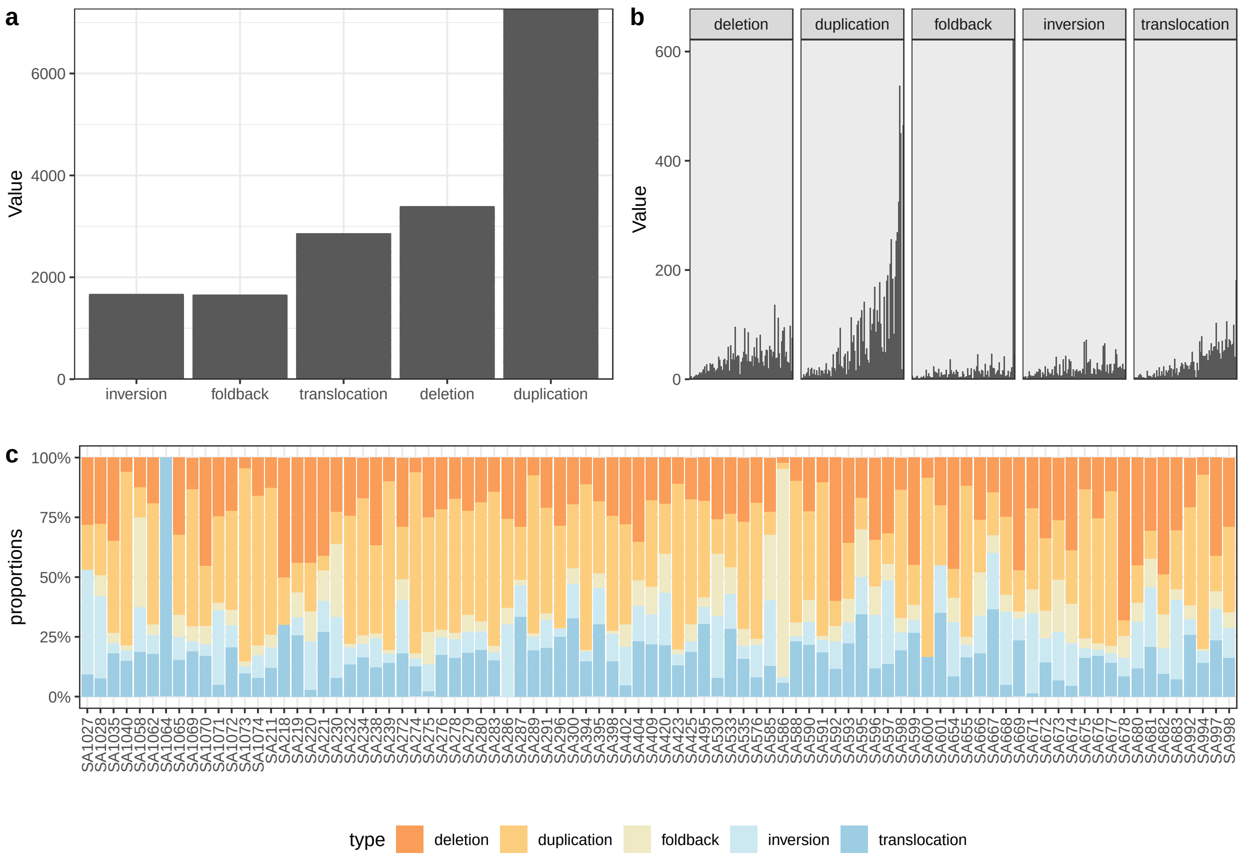


Figure 3.6: Distribution of structural variants (SVs) per sample and across the cohort: **(a)** and **(b)** show the distribution and abundance of SV types across the cohort, sorted in ascending order based on the total number of mutations observed in each SV type (inversions, foldback, translocations, deletions and duplications). **(c)** The proportion of SV types (y-axis) identified in each sample (x-axis).

Overall, we see that TNBCs are enriched for duplications followed by deletions (Fig. 3.6 **a)**) with clear genomic heterogeneity observed between cases in this cohort, some harboring significant structural variations in specific variant types compared to others (Fig. 3.6 **c)**). Specific structural variants disrupt gene structures and consequently promote tumour progression. SVs and mutation signatures derived from specific structural variants have played an important role in patient and prognostic stratification and the identification of potentially actionable events [42, 47, 73]. Detected SVs in this study were used as key features for patient stratification as will be expounded on in section 3.3. The data object 'breakpoints.all' used to store structural variant data extracted from the database and used to generate figures 3.6 **a)**, **b)** and **c)** was created using the following query:

```
breakpoints.all <- dbGetQuery(con,
                              "SELECT DISTINCT tumour_id, type, COUNT(*)
                               FROM svs_filtered
                               WHERE type = ‘foldback’
                               GROUP BY 1, 2
                               UNION SELECT DISTINCT tumour_id, type, COUNT(*)
                               FROM svs_filtered
                               WHERE type = ‘duplication’
                               GROUP BY 1, 2
                               UNION SELECT DISTINCT tumour_id, type, count(*)
                               FROM svs_filtered WHERE type = ‘translocation’
                               GROUP BY 1, 2
                               ... ")
```

Sample data extract:

```
 tumour_id |     type      | count
-----------+---------------+-------
 SA1071    | duplication   |    22
 SA669     | inversion     |    26
 SA673     | deletion      |    23
 SA586     | duplication   |    18
 SA423     | duplication   |    94
 SA287     | foldback      |     1
 SA275     | inversion     |     5
 SA232     | duplication   |    99
 SA997     | duplication   |    10
 SA211     | translocation |    37
```

3.2.3 Copy number aberrations

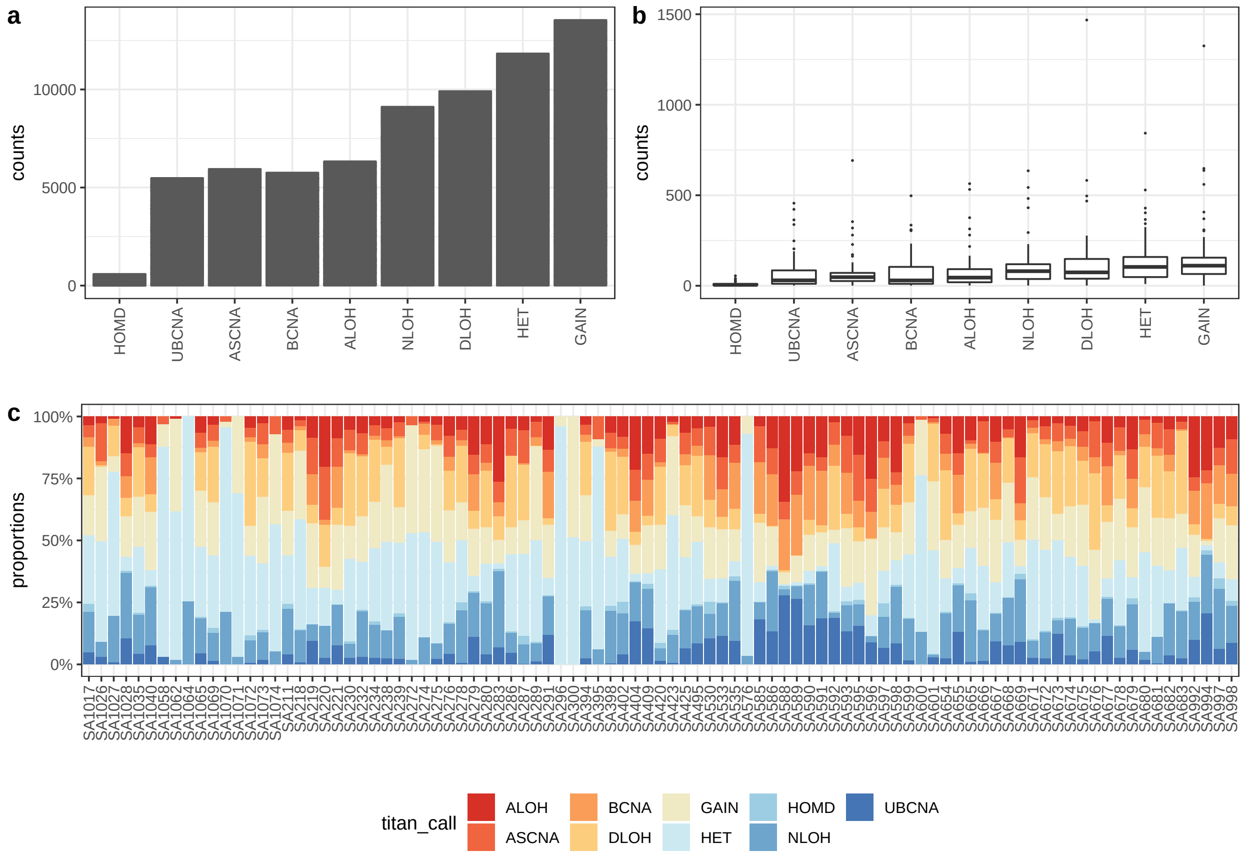


Figure 3.7: Distribution of copy number aberrations (CNAs) per sample and across the cohort: **(a)** and **(b)** The distribution and abundance of CNA types across the cohort, sorted in ascending order based on the total number of mutations observed in each CNA type (Homozygous deletion (HOMD), Unbalanced CNA (UBCNA), Allele-specific CNA (ASCNA), Balanced CNA (BCNA), Amplified LOH (ALOH), Copy-neutral LOH (NLOH), Hemizygous deletion (DLOH), Diploid heterozygous (HET) and copy number GAIN). **(c)** The proportion of CNA types (y-axis) identified in each sample (x-axis).

Fig. 3.7 **(c)** shows the variation in copy number profiles in the TNBC cohort and the heterogeneity of TNBCs at CNA level. The Intra-sample heterogeneity at both CNA and SV level is further depicted in Fig. 3.8 that displays the variations in the genome structure of patient sample SA586 and the corresponding relationships between genomic intervals.

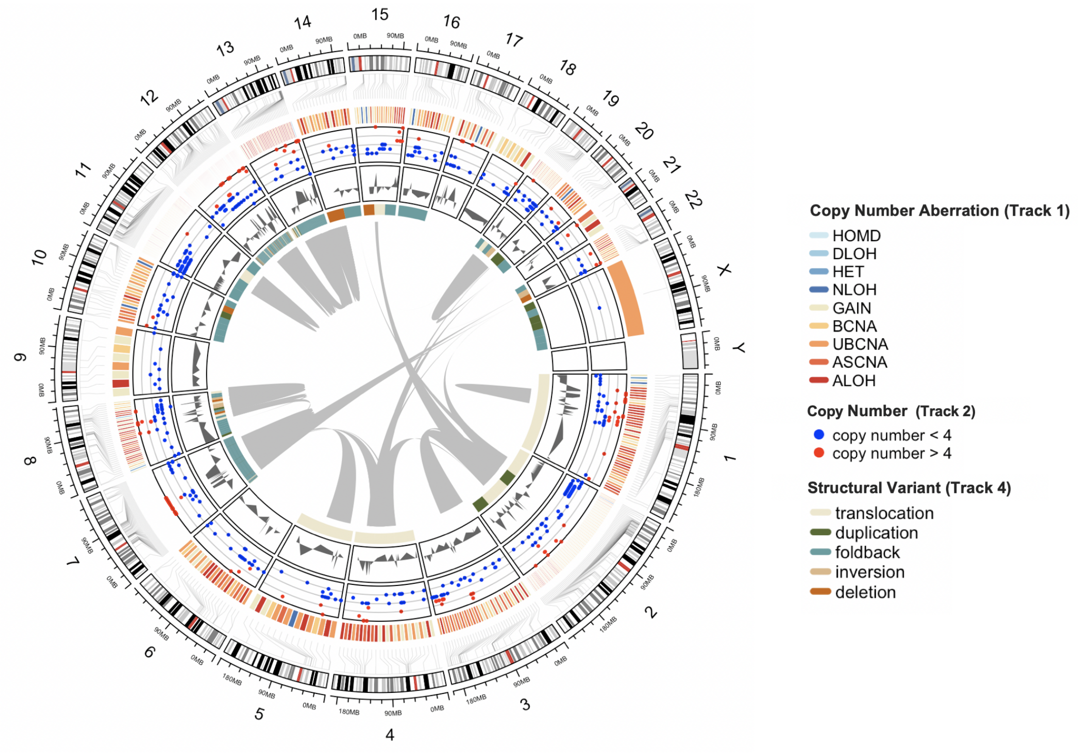


Figure 3.8: Intra-sample heterogeneity at both CNA and SV levels: Circos plot showing the type of copy number aberrations (HET, BCNA, UBCNA, ALOH, ASCNA, NLOH, DLOH, HOMD, GAIN) across the genome (track 1), copy number variations in the genome (track 2 and 3) and the type of structural variations (translocation, duplication, foldback, deletions, and inversions - track 4) followed by corresponding links between genomic positions.

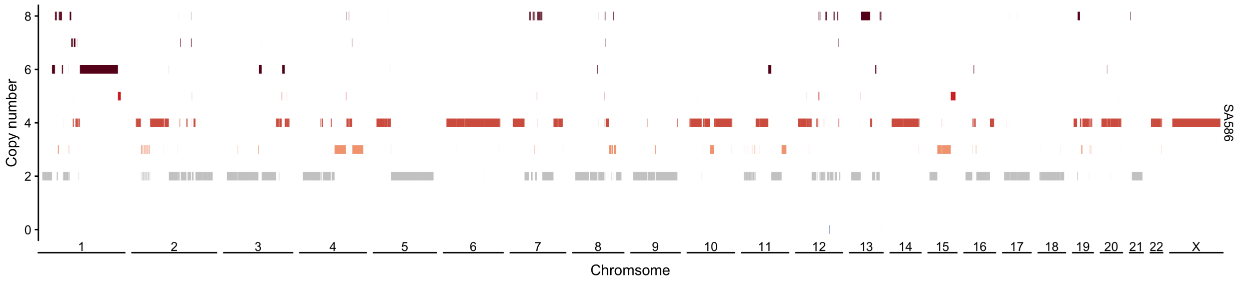


Figure 3.9: Case-based copy number profile: Copy number profile of patient sample SA586 showing copy number variations (y-axis) along the genome denoted by coordinates representing genomic positions (x-axis).

The below snippet shows the database call required to extract the data used to generate Fig. 3.8 and Fig. 3.9 for CNAs and SVs respectively followed by sample data extracts.

CNAs data call:

```
sample <- input$sample_id
query <- fn$identity("SELECT chromosome, start_position_bp, end_position_bp,
                    titan_call, copy_number
                    FROM titan_segs_cnas
                    WHERE tumour_id = ‘$sample’")
cnas <- dbGetQuery(con, query)
```

Sample data extract (CNAs):

```
chromosome | start_position_bp | end_position_bp | titan_call | copy_number
-----------+-------------------+-----------------+------------+-------------
13         |          50046072 |        79052038 | ALOH       |           8
7          |          66617961 |        66628179 | ALOH       |           8
7          |          66591794 |        66594881 | ALOH       |           8
4          |         156518707 |       191043593 | ALOH       |           3
22         |          47415860 |        47415875 | ALOH       |           8
```

SVs data call:

```
sample <- input$sample_id
query <- fn$identity("SELECT chrom_1, brk_1, chrom_2, brk_2, brk_dist, type
                    FROM svs_filtered
                    WHERE tumour_id = ‘$sample’ ORDER BY 1")
svs <- dbGetQuery(con, query)
```

Sample data extract (SVs):

```
 chrom_1 |   brk_1   | chrom_2 |   brk_2   | brk_dist |     type
---------+-----------+---------+-----------+----------+---------------
 1       |  64435479 | 3       | 101565610 | Infinity | translocation
 10      |  61995259 | 10      |  61993238 |     2021 | duplication
```

```
    10       |  34280375 | 10       |  34280333 |       42 | foldback
    10       |  28596747 | 10       |  28596716 |       31 | foldback
    10       |  43885134 | 10       |  43885091 |       43 | foldback
    10       | 124903202 | 10       | 124903123 |       79 | foldback
    10       |  83072235 | 10       |  83106088 |    33853 | deletion
    10       |   2186974 | 10       |   2186941 |       33 | foldback
    11       | 124022349 | 11       | 124022281 |       68 | foldback
    11       | 119800299 | 13       |  68122937 | Infinity | translocation
```

3.2.4 Gene-level analysis

The identification of significantly mutated genes (Appendix B) across this TNBC cohort was accomplished using MutSigCV from which *EMCN*, *TP53*, *MUC21*, *PIK3CA*, *MUC4*, *MB*, *CTU2*, *RAB3IL1*, *PTEN* were identified as the most significantly mutated genes in this cohort *(FDR < 0.1)*. Database derived mutations in each gene per case were visualized using an oncoplot (Fig. 3.10) with each row representing a gene and each column representing a case. As expected, *PIK3CA* mutations appear mutually exclusive with *PTEN* loss. The script written to extract the data used to generate Fig. 3.10 is shown in Appendix A.2.0.2.

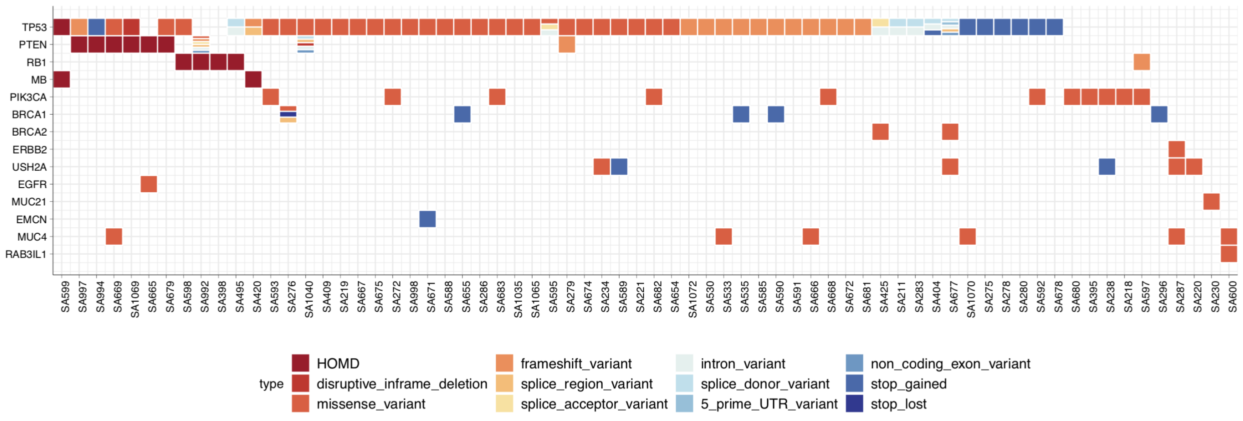


Figure 3.10: Visualizing gene-based mutations: Oncoplot showing high impact mutations in each gene (rows) per sample (columns). Multiple mutations in a gene are represented by multiple colors representative of specific mutation types in a single gene. *TP53* (56.9%) was identified as the most frequently mutated gene in this cohort, followed by *PIK3CA* (8.9%), *PTEN* (7.3%), *BRCA1* (5.7%), *USH2A* (4.9%), *MUC4* (4.9%) and *RB1* (4.1%) respectively.

3.3 TNBC genomic subgroup discovery

One of the main objectives of this study was to stratify TNBCs into distinct subgroups using genomic features extracted from the developed database. This was hinged on our hypothesis that TNBC patients can be stratified into distinct genomic subgroups based on their whole genome profiles. Genomic features integrated for subgroup discovery included CNAs (HET, DLOH, GAIN, NLOH, HOMD, ASCNA, ALOH, BCNA and UBCNA), SNVs (stop_gained, splice_donor, splice_acceptor, start_lost and stop_lost), indels (frameshift_variant, splice_donor, splice_acceptor, stop_gained, bidirectional_gene_fusion, gene_fusion and stop_lost), SVs (duplication, deletion, translocation, inversion and foldback) and mutation signatures (POLE, APOBEC, HRD (Homologous Recombination Deficiency), UNK (Unknown), MMRD (Mismatch Repair Deficiency), T→C, M-Dup (Medium Duplications), S-Del (Small Deletions), Cl-SV (Clustered Structural Variants), FBI (Foldback Inversions), Cl-FBI (Clustered Foldback Inversions), L-Del (Large Deletions), S-Dup (Small Duplications), Tr (Translocations) and L-Dup (Large Duplications)). CNAs, SNVs, SVs, and indels were computed as the proportion of each variant over all domain specific variants while mutation signatures were inferred using the multi-modal correlated topic model (MMCTM) [42]. All the identified stratification features were used for integrative hierarchical clustering analysis using the R package pheatmap and the Manhattan distance measure to support the discovery of patient subgroups and their genomic and clinical characteristics. Figures 3.11, 3.12, 3.13, 3.14, 3.15 and 3.17 show subgroups identified by mutation signatures, CNAs, SNVs, indels, SVs and by multi-feature integration respectively.

3.3.1 TNBC subgroups identified by mutation signatures

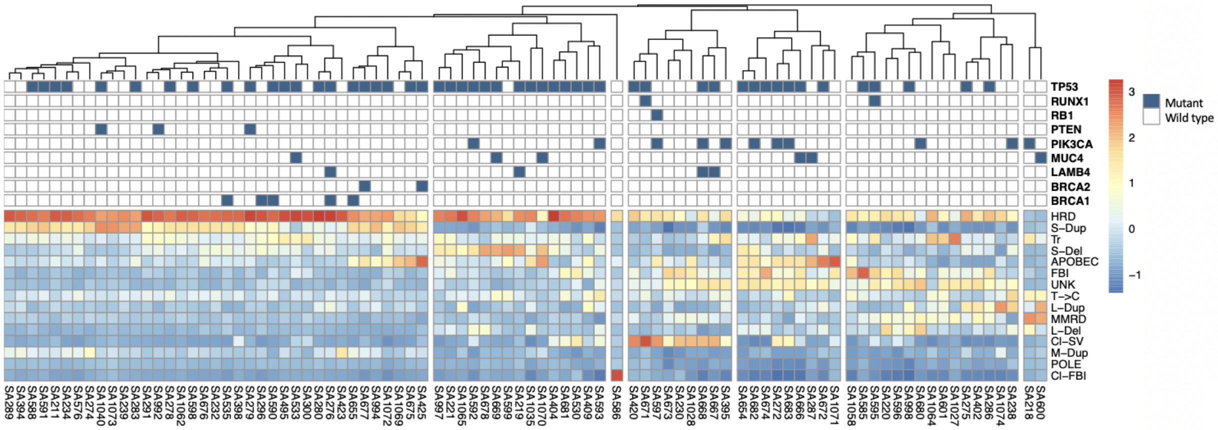


Figure 3.11: TNBC genomic subgroups identified by mutation signatures: Hierarchical clustering of 88 TNBC cases (x-axis) reveals 5 subgroups using scaled values of mutation signatures (POLE, APOBEC, HRD, UNK, MMRD, T→C, M-Dup, S-Del, Cl-SV, FBI, Cl-FBI, L-Del, S-Dup, Tr and L-Dup) (rows in the bottom panel of the heatmap). Color scales range from blue to red to reflect no or low proportions of a variant (blue) relative to high variant proportion levels (red) in each case. Heatmap annotations are shown in rows in the top panel where blue signifies presence of a mutation (mutant) in significantly mutated genes and in DNA damage repair genes while white signifies absence of a mutation in a gene (wild type) for each case.

Stratification of TNBC cases in this cohort by mutation signatures (Fig. 3.11) led to the discovery of 5 main subgroups. The first 2 groups (leftmost) were identified enriched for the HRD signature and further distinguished by S-Dup and S-Del signatures in group 1 and group 2 respectively. ∼1/5 of the samples in group 1 were identified enriched for the APOBEC signature. Group 3 unlike other groups was highly enriched for the Cl-SV signature, group 4 was enriched for APOBEC, FBI and a signature unknown (UNK) while group 5 was enriched for FBI, UNK and MMRD. 3 cases (cluster 3 and 7) did not fall in any of the main clusters and therefore flagged as outliers. Based on this stratification, all patients with a *BRCA1/BRCA* 2 mutation were classified in group 1 which also had no case with a *PIK3CA* mutation. All cases with a *PTEN* mutation were also classified in group 1. A chi-square test was conducted to check for mutual exclusivity among subgroup gene-based mutations, however, this test and all subsequent tests yielded low p-values (> 0.33) due to few observations. There are future prospects of re-testing mutual exclusivity with a larger cohort.

3.3.2 TNBC subgroups identified by CNAS

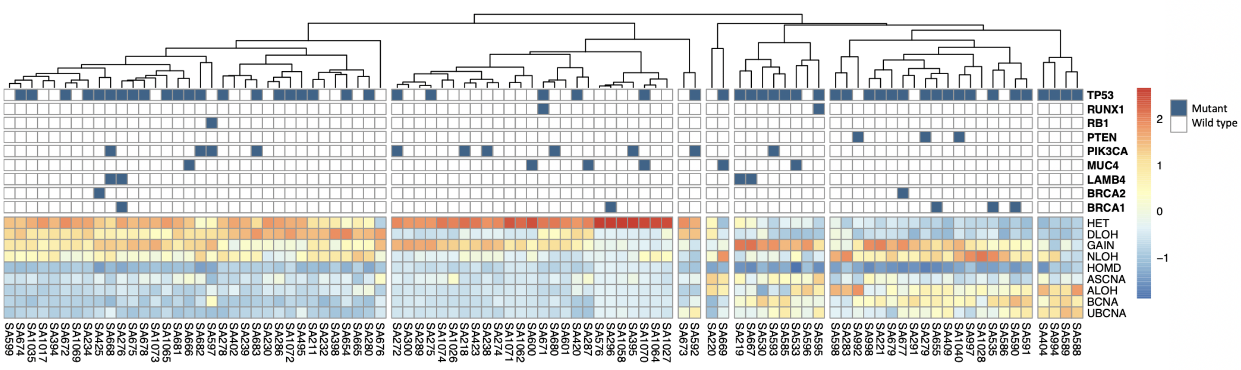


Figure 3.12: Stratification of cases by scaled values of CNA proportions (HET, DLOH, GAIN, NLOH, HOMD, ASCNA, ALOH, BCNA and UBCNA) reveals 5 subgroups.

Stratification of TNBC cases by copy number aberrations revealed 5 subgroups (Fig. 3.12); group 1 (leftmost) was enriched for HET, DLOH, GAIN and NLOH; group 2 was significantly enriched for HET, with ∼3/4 of the cases being enriched for copy number GAIN, the third and 4th clusters containing 2 cases each were flagged as outliers. Group 3 and 4 were identified enriched for copy number GAIN with NLOH being a distinguishing feature found enriched in group 4 while group 5 was identified enriched for UBCNA compared to other subgroups followed by ALOH and BCNA respectively. All cases with a mutation in *PTEN* were found in group 4 which also comprised of most cases with a *BRCA1* mutation and no case with a *PIK3CA* mutation. Group 2 that was enriched for HET, had the fewest cases with a *TP53* mutation and with the highest number of cases with a mutation in *PIK3CA*.

3.3.3 TNBC subgroups identified by SNVs

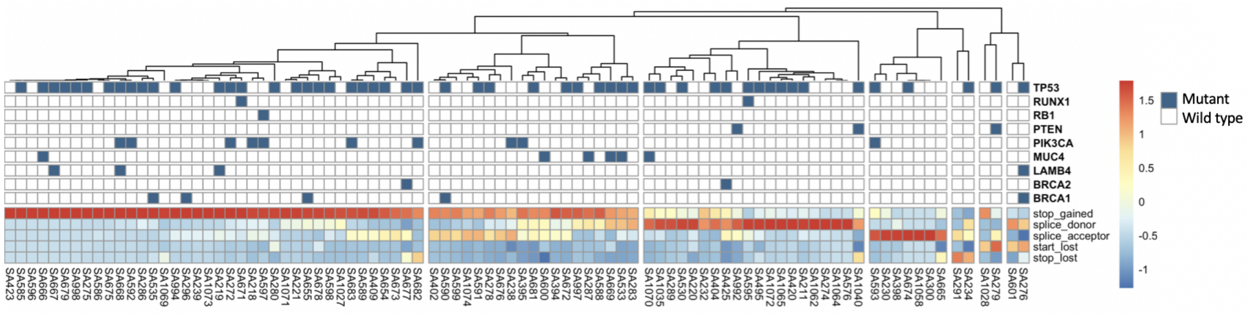


Figure 3.13: Stratification of cases by scaled values of SNV proportions (stop_gained, splice_donor, splice_acceptor, start_lost and stop_lost) reveals 4 subgroups.

Patient stratification by SNVs revealed 4 main subgroups (Fig. 3.13), the first (leftmost) heavily enriched for stop_gained mutations while group 2 was identified enriched for stop_gained with ∼1/2 of the cases in this group being enriched for splice_donor and splice_acceptor mutations (∼1/3). Group 3 was heavily enriched for splice_donor while group 4 was heavily enriched for splice_acceptor. The 3 right-most clusters comprising of 2 cases each were flagged as outliers. Most cases with *BRCA1, PIK3CA* or *LAMB4* mutations were clustered in group 1 which had no case with a *PTEN* mutation. Group 2 had most cases with a mutation in MUC4, group 3 had no cases with either a *BRCA1* or *PIK3CA* mutation while group 4 had cases with the fewest mutations in the genes of interest.

3.3.4 TNBC subgroups identified by indels

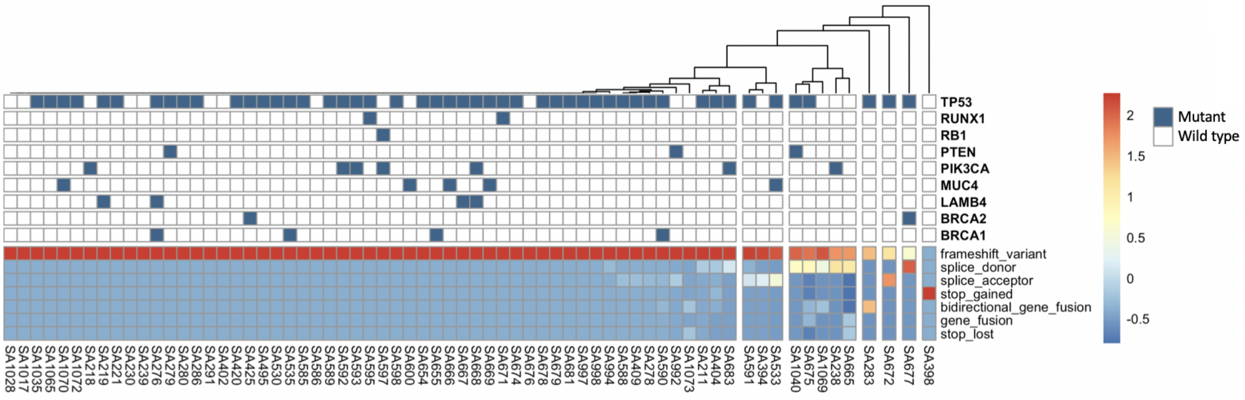


Figure 3.14: Stratification of cases by scaled values of indel proportions (frameshift_variant, splice_donor, splice_acceptor, stop_gained, bidirectional_gene_fusion, gene_fusion and stop_lost) identifies 3 subgroups.

Stratification of TNBC cases by indels (Fig. 3.14) identified 3 groups, all of which were enriched for frame_shift variants with a stronger enrichment in group 1. Compared to other groups, group 2 and 3 had a higher signal for splice_acceptor and splice_donor mutations respectively. Flagged as outliers were the right-most 4 clusters comprising of 1 case each. Whether cases in group 2 and 3 have no *BRCA1/2* mutations remains inconclusive due to the few cases in these groups.

3.3.5 TNBC subgroups identified by SVS

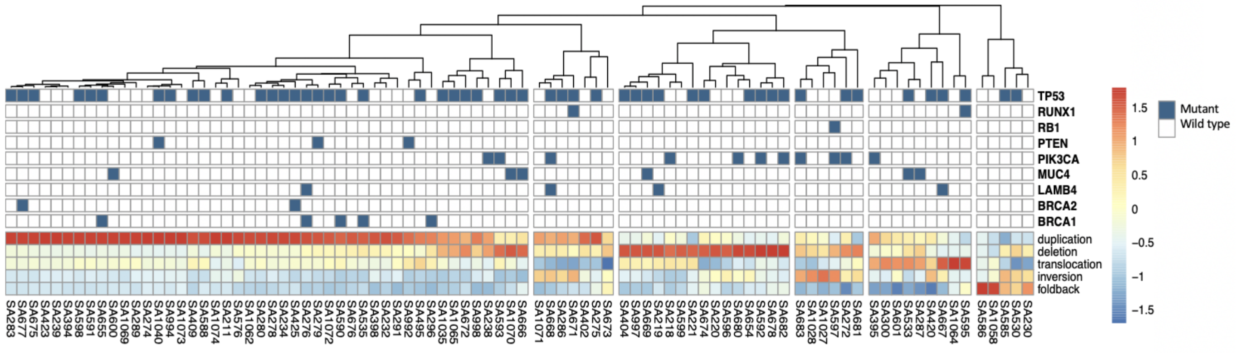


Figure 3.15: Stratification of cases by scaled values of SV proportions (duplication, deletion, translocation, inversion and foldback) reveals 6 subgroups.

Patient stratification was also done based on SVs (Fig. 3.15) from which 6 subgroups were identified. Group 1 (leftmost) in which all cases with a *PTEN* or a *BRCA1/2* mutation were clustered was heavily enriched for duplications. Group 2 was identified enriched for duplications, deletions and inversions while group 3 was heavily enriched for deletions. Group 4 was enriched for inversions (compared to other groups) while group 5 and 6 were enriched for translocations and foldback inversions respectively. No cases in groups 2 - 6 had a mutation in *PTEN*, *BRCA1* and *BRCA2*.

The above database derived analyses for patient stratification show that TNBCs can be stratified based on mutation signatures and individual domain specific somatic variants (CNAs, SVs, SNVs and indels) into subgroups that depict unique patterns and characteristics with mutations in SMGs or in DNA damage repair genes seen enriched in certain groups compared to others. We also see that some driver mutations are associated with mutation signatures that stratified the TNBCs as seen in Fig. 3.11 where all *BRCA1*, *BRCA2* and *PTEN* mutations were seen enriched in the HRD group.

3.3.6 TNBC subgroup discovery by genomic feature integration

The preceding section demonstrates that TNBCs can be stratified into distinct subgroups by mutation signatures and by domain specific genomic variants (CNAS, SNVs, indels and SVs). Integrating all genomic features (CNAs (HET, DLOH, GAIN, NLOH, HOMD, ASCNA, ALOH, BCNA and UBCNA), SNVs (stop_gained, splice_donor, splice_acceptor, start_lost and stop_lost), indels (frameshift_variant, splice_donor, splice_acceptor, stop_gained, bidirectional_gene_fusion, gene_fusion and stop_lost), SVs (duplication, deletion, translocation, inversion and foldback) and mutation signatures (POLE, APOBEC, HRD, UNK, MMRD, T→C, M-Dup, S-Del, Cl-SV, FBI, Cl-FBI, L-Del, S-Dup, Tr and L-Dup)) for patient stratification identified 5 subgroups (Fig. 3.17). The optimal number of clusters was identified using the Elbow method (Fig. 3.16) that suggested 5 optimal clusters.

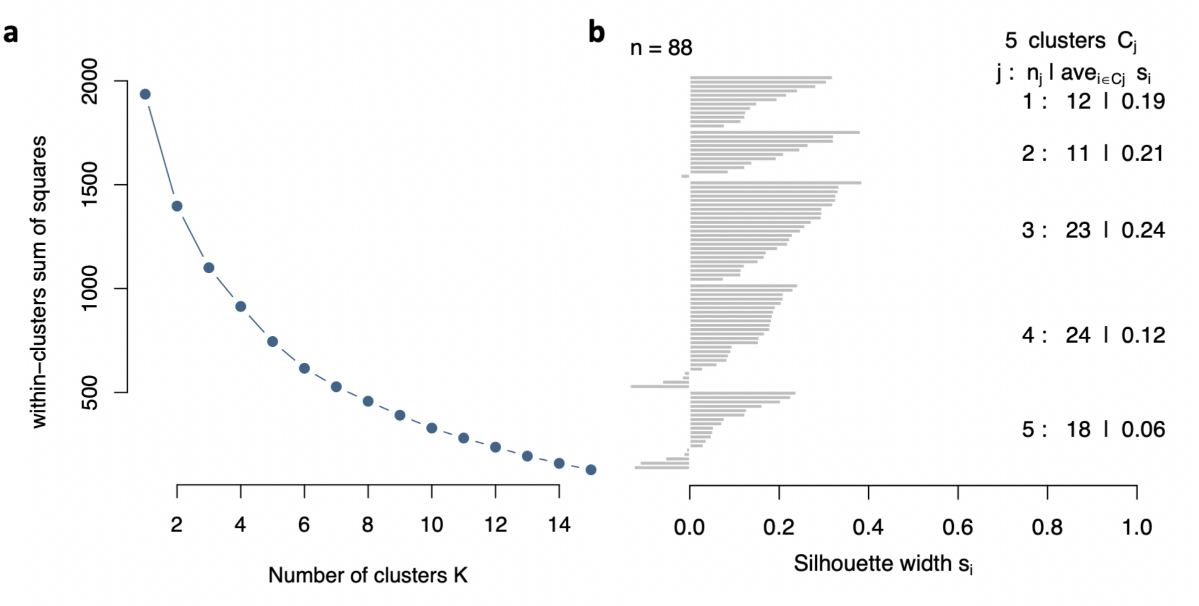


Figure 3.16: TNBC genomic subgroups - Optimal number of clusters: **(a)** Optimal number of clusters = 5 as identified by the Elbow method. **(b)** Silhouette measure (range = -1 to +1) of within-cluster similarity where high values indicate that a case is well matched to its own cluster and poorly matched to neighboring clusters).

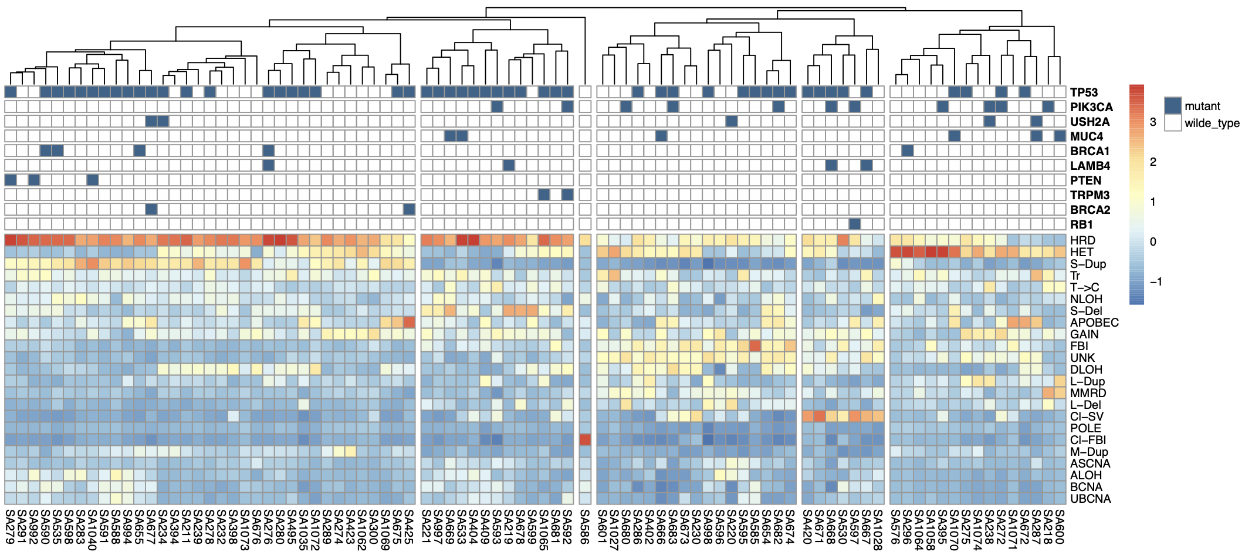


Figure 3.17: TNBC genomic subgroups identified by genomic feature integration (mutation signatures, CNAs, SVs, SNVs and indels): Hierarchical clustering of 88 TNBC cases (x-axis) revealed by scaled values of mutation signatures (HRD, S-Dup, Tr, T→C, S-Del, APOBEC, FBI, UKN, L-Dup, MMRD, L-Del, Cl-SV, POLE, Cl-FBI, M-Dup), CNAs (HET, NLOH, GAIN, DLOH, ASCNA, ALOH, BCNA, UBCNA) (rows in the bottom panel of the heatmap) with a dendrogram of the hierarchical cluster analysis. Color scales range from blue to red to reflect no or low proportions of a variant (blue) relative to high variant proportion levels (red) in each case. Heatmap annotations are shown in rows in the top panel where blue signifies presence of a mutation (mutant) in significantly mutated genes and in DNA damage repair genes while white signifies absence of a mutation in a gene (wild type) for each case.

Stratification of TNBCs in this cohort by genomic feature integration using hierarchical clustering and the Manhattan distance measure, identified 5 novel TNBC subgroups. Group 1 and 2 (leftmost) were found enriched for the HRD signature and were further distinguished by the S-Dup signature that is seen only enriched in group 1. Unlike group 2, group 1 was also enriched for copy number aberration HET. Flagged as an outlier was the third cluster comprising of one case heavily enriched for the Cl-FBI signature compared to other subgroups. Group 3 and 4 were enriched for the FBI signature with a stronger signal found in group 3. These 2 groups were further distinguished by the Cl-SV signature that was enriched in group 4. Group 5 was enriched for HET with ∼1/3 of the cases enriched for APOBEC. All groups except for group 1 had at least one case with a *PIK3CA* mutation. No cases with a mutation in *MUC4* were identified in group 1 and 4 while all groups except for groups 2, 3 and 4 had at least one case with a *BRCA1* mutation. All cases with a mutation in *PTEN* or in *BRCA2* were clustered in group 1. As seen in previous analyses, the identified subgroups are associated with differing characteristics such as the enrichment of mutations in the genes of interest and the association of specific gene mutations with mutation signatures. Almost all *BRCA1* mutant cases are seen associated with the HRD group (group 1) that contains all the cases with a *BRCA2* or *PTEN* mutation. We also see less association of *PIK3CA* and *MUC4* mutations with the HRD signature.

3.3.7 TNBC genomic subgroup analysis

3.3.7.1 Subgroup comparative analyses of mutation loads

To improve our understanding of the identified genomic subgroups beyond their association with SMGs, DNA damage repair genes, mutation signatures and somatic aberrations, comparative analyses were conducted from both genomic and clinical perspectives to compute the prevalence of mutations in each subgroup and to conduct comparative clinical data analyses. Fig. 3.18 shows the mutation loads per subgroup based on **a)** SNVS, **b)** indels, **c)** SVs and **d)** the total mutation load. Based on SNVs, group 2 (HRD) had the highest mutation load followed by group 1 (HRD + S-Dup) while group 3 had the lowest (FBI) and by indels, group 4 (FBI + Cl-SV) had the highest mutation load while group 3 (FBI) had the lowest. Based on SVs, group 1 (HRD + S-Dup) had the highest mutation load while group 5 (HET) had the lowest. Overall, group 2 (HRD) had the highest mutation load followed by group 1 (HRD + S-Dup) while group 3 had the lowest. The summary of these analyses is presented in Fig. 3.19.

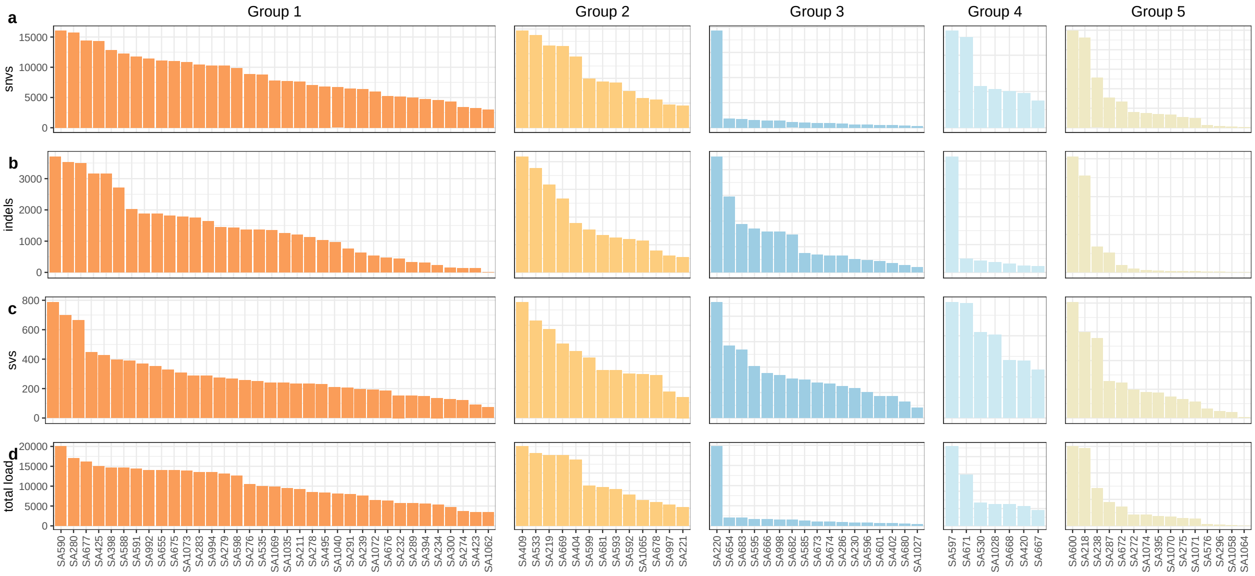


Figure 3.18: Subgroup mutation loads: **(a):** SNV average mutation load per subgroup:- Group 1 = 8,630, Group 2 = 11,001, Group 3 = 4,041, Group 4 = 7,222, Group 5 = 6,463, **(b)**: Indel average mutation load per subgroup:- Group 1 = 1,406, Group 2 = 1,847, Group 3 = 609, Group 4 = 2,211, Group 5 = 982, **(c):** SV average mutation load per subgroup:- Group 1 = 284, Group 2 = 193, Group 3 = 81, Group 4 = 196, Group 5 = 59, **(d):** Total average mutation load per subgroup:- Group 1 = 10,320, Group 2 = 13,041, Group 3 = 4,674, Group 4 = 9,628, Group 5 = 7,504.

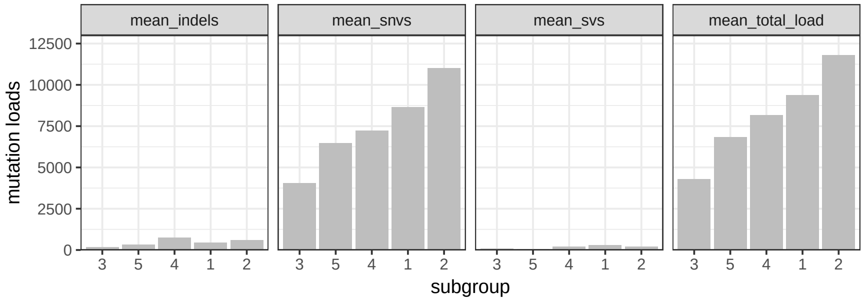


Figure 3.19: Subgroup mean mutation loads

3.3.7.2 Subgroup comparative analyses of the distribution of rearrangements

Comparative analyses for subgroup rearrangement distributions were conducted (Fig. 3.20) and identified a high proportion of translocations, duplications, deletions, inversions (other types of inversions other than foldback inversions), and foldback inversions in group 2, 1, 2, 4 and 3 respectively as shown below.

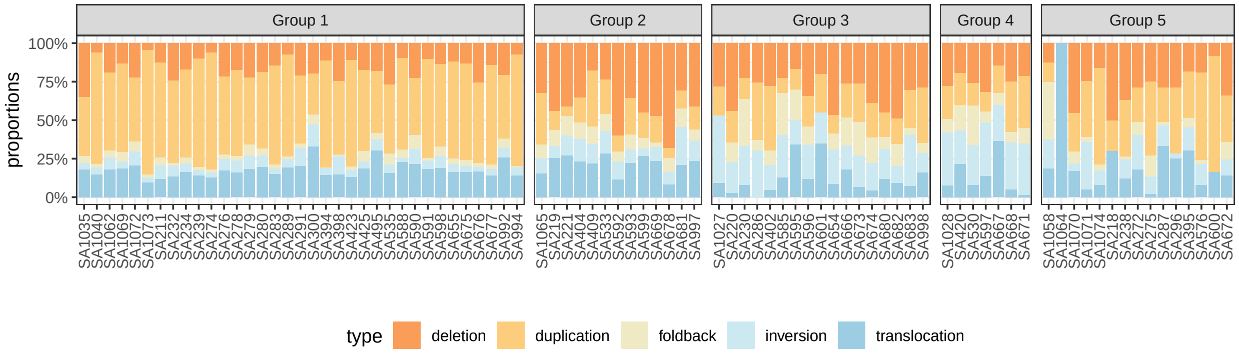


Figure 3.20: Subgroup rearrangement distributions. Group 1 was found enriched for duplications (58.69%), Group 2 enriched for deletions (37.73%), Group 3 enriched for deletions (30.64%), Group 4 enriched for inversions (28.27%) while group 5 was enriched for duplications (38.69%). The group that associated with the highest translocations was Group 2, duplications - Group 1, deletions - Group 2, inversions - Group 4 and foldback - Group 3.

```
               | Group 1     Group 2     Group 3     Group 4     Group 5

---------------|-----------------------------------------------------------

Translocation  |17.26%      22.15%      13.83%      15.78%      14.40%

Duplication    |58.69%      19.24%      21.03%      22.13%      38.70%

Deletion       |14.62%      37.73%      30.64%      22.21%      26.66%

Inversion      |5.69%       11.65%      20.01%      28.27%      13.39%

Foldback       |3.74%       9.22%       14.48%      11.61%      6.86%
```

3.3.7.3 Subgroup comparative analyses of trinucleotide distributions

Trinucleotide distribution analyses (Fig. 3.21) also revealed distinct trinucleotide distributions per subgroup where group 2 (HRD) was enriched for C→A substritutions followed by group 1 (HRD + S-Dup) with the least abundance of C→A substritutions identified in group 3 (FBI). Group 2 (HRD) followed by group 1 (HRD + S-Dup) had the highest abundance of C→G substitutions while group 3 (FBI) had the lowest. C→T substitutions were highest in group 5 (HET) closely followed by group 2 while group 3 had the lowest. The abundance of T→A substritutions was seen enriched in group 2 (HRD) and lowest in group 3 (FBI). T→C substritutions were enriched in group 2 (HRD) compared to group 3 (FBI) that had the lowest while T→G substritutions were most abundant in group 2 (HRD).

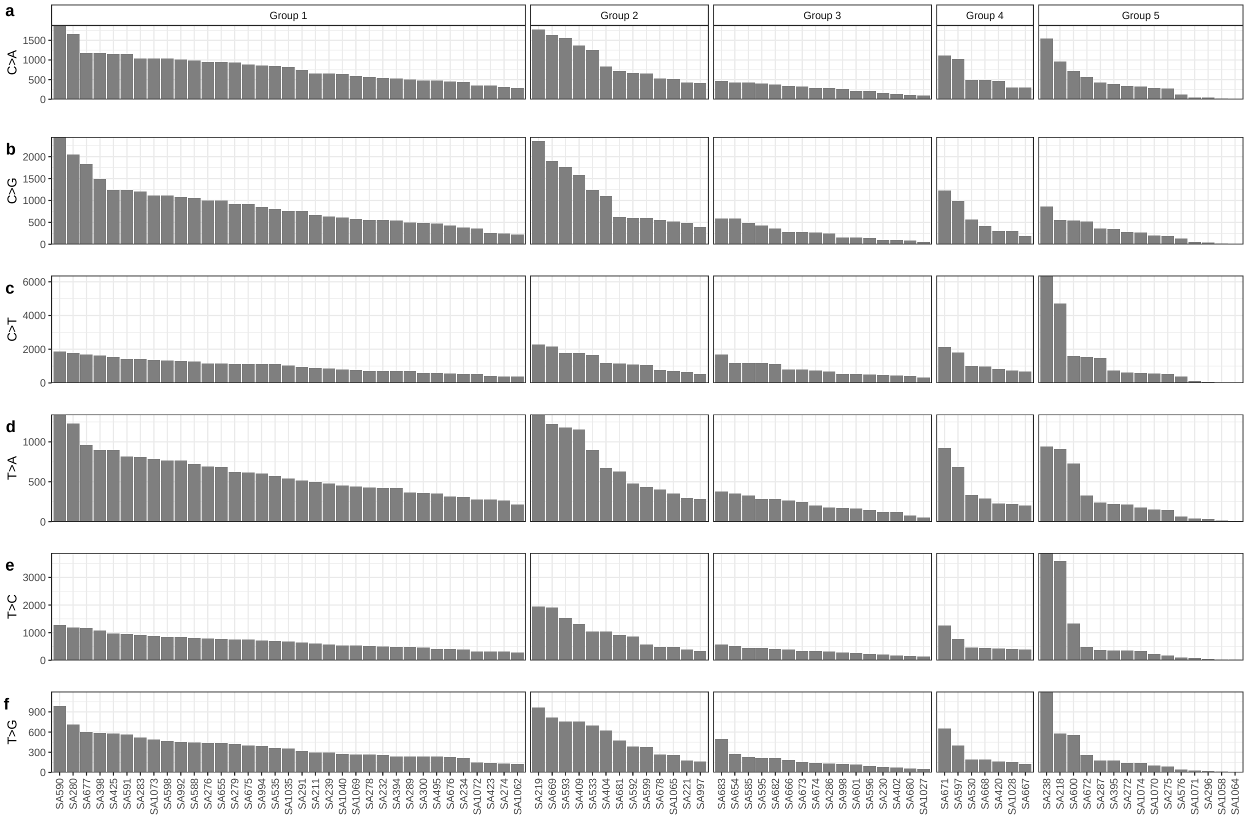


Figure 3.21: Subgroup trinucleotide distributions: **(a)** C→A substitutions: Means = 802.7714, 948.5385, 281.25, 598.1429, 404.6 for groups 1 - 5 respectively. **(b)** C→G substitutions: Means = 867.4571, 1055, 269.75, 570.1429, 290.4 for groups 1 - 5 respectively. **(c)** C→T substitutions: Means = 1001.286, 1279.308, 771.9375, 1154, 1279.4 for groups 1 - 5 respectively. **(d)** T→A substitutions: Means = 590.1714, 717, 208.75, 409.7143, 279.2 for groups 1 - 5 respectively. **(e)** T→C substitutions: Means = 680, 984.2308, 325.125, 590.1429, 758.2667 for groups 1 - 5 respectively. **(f)** T→G substitutions: Means = 371.7429, 514.4615, 162.3125, 265.1429, 233.7333 for groups 1 - 5 respectively.

3.3.7.4 Subgroup comparative analyses from a clinical perspective

To determine the association between the identified subgroups and clinical outcomes, comparative analyses of patient groups by age, tumour_size, node status, grade and overall survival were conducted. Comparison of subgroups by age identified no large variance in subgroup mean ages (means = 50.08, 52.01, 58.85, 64.5, 62.11 for groups 1 - 5 respectively), however, as seen in (Fig. 3.22: **a)**), group 4 (FBI + Cl-SV) was identified having the highest average age (range: 47 - 81) while group 1 (HRD + S-Dup) had patients with the youngest ages (range: 26 - 80), with more than half of the patients having age < 50. We would expect that younger patients would be associated with a lower mutation burden but this was on the contrary as younger patients were clustered in one of the groups with the highest mutation burden (HRD + S-Dup group) suggestive of the likelihood that younger patients have more proliferative disease and therefore the tumors accumulate mutations at a higher rate.

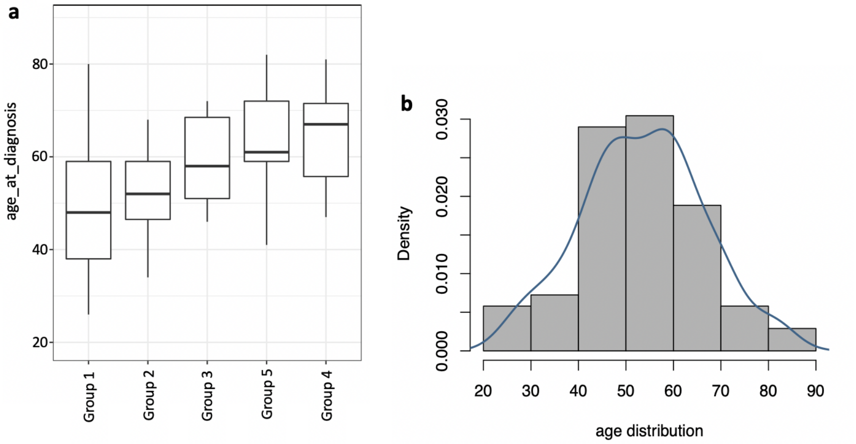
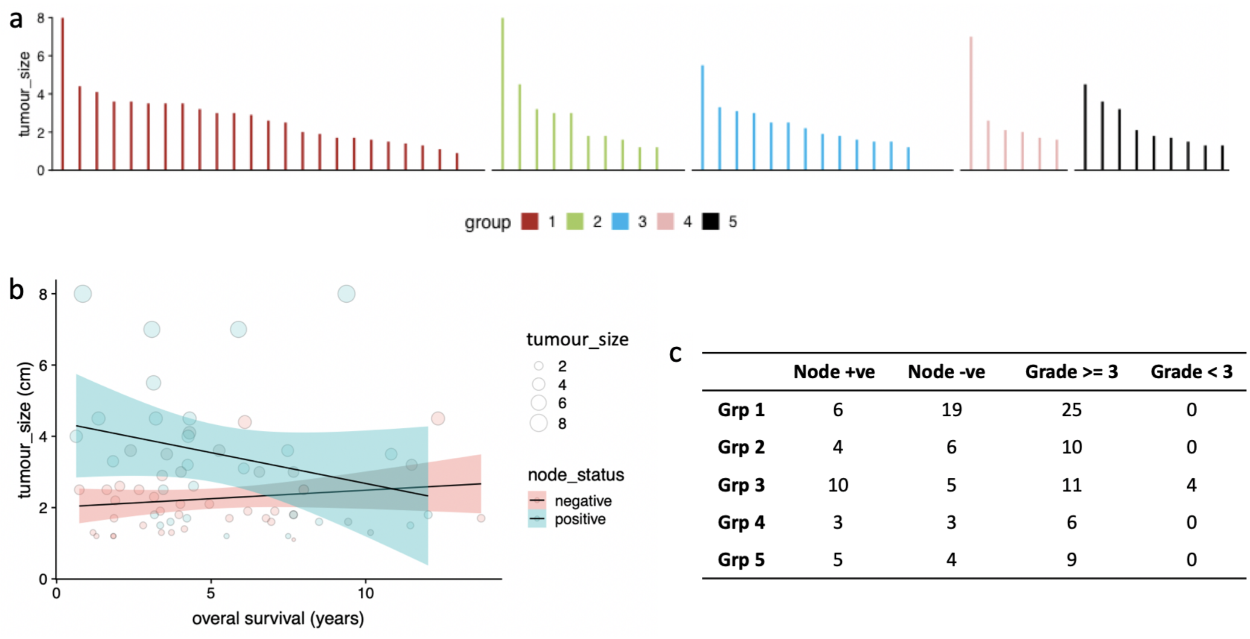


Figure 3.22: Subgroup comparative analyses based on clinical outcomes - age: **(a)** Age distribution per subgroup. **(b)** Age distribution of cases in this TNBC cohort: average age = 55.4years, range = 26years - 82years.

Based on tumour size and grade, group 2 (HRD) had the highest average tumour_size while group 5 (HET) had the lowest (Fig. 3.23: **a**)). As shown in (Fig. 3.23: **c**)), 24% of the patients in group 1 (HRD + S-Dup) were found node positive, while 40%, 67%, 50% and 54% of the patients in groups 2, 3, 4 and 5 were found node positive respectively. Noted from these analyses is that node positivity is not associated with mutation burden as groups with high mutation loads had fewer patients whose tumours were node positive. 67% of patients in the lowest mutation load group (group 3 (FBI)) were node positive while only 24% of the patients in a higher mutation load group were node positive (group 1 (HRD + S-Dup)). Also, tumour size was found dissociated from node positivy as seen with groups 1 (HRD + S-Dup) (24% node +ve, average tumour_ size = 2.7), 3 (FBI) (67% node +ve, average tumour_size = 2.4) and 2 (HRD) (40% node +ve, average tumour_size = 2.9). All cases in each subgroup presented with high grade tumours but for group 3 (FBI) that had some cases (27%) with low grade tumors.



	Node +ve	Node -ve	Grade >= 3	Grade < 3
Grp 1	6	19	25	0
Grp 2	4	6	10	0
Grp 3	10	5	11	4
Grp 4	3	3	6	0
Grp 5	5	4	9	0

Figure 3.23: Subgroup comparative analyses based on clinical outcomes - tumour size, node status and grade: **(a)** Tumour_size distribution: Means = 2.77cm, 2.93cm, 2.43cm, 2.83cm, 2.33cm for groups 1 - 5 respectively. **(b)** Relationship between tumour size, node status and overall survival. **(c)** Chi-square contingency table of node status and grade observations per subgroup.

A preliminary sniff into the overall survival of the identified patient subgroups identified group 5 (HET) to putatively have the best survival outcomes while group 2 (HRD) has the worst. Despite the HRD group (group 2) having the highest mutation burden and the worst overall survival (OS), we also see that the FBI group (group 3) with the lowest mutation burden is not associated with the best overall survival but group 5 (HET) is. This goes to show that the mutation burden may not necessarily have a large bearing on patient outcomes. Secondly, group 1 (HRD + S-Dup) with a higher average mutation burden has better OS than the FBI group with a lower average mutation burden.

From these statistical analyses we can putatively deduce that:

1. Mutation burden may not necessarily have a large bearing on patient outcomes as the FBI group (group 3) with the lowest mutation burden was not found associated with the best overall survival (OS). Secondly, group 1 (HRD + S-Dup) with a higher average mutation burden was found to have a better OS than the FBI group with a lower average mutation burden.

2. Node positivity is not associated with mutation burden as groups with high mutation loads had fewer patients whose tumours were node positive. 67% of patients in the lowest mutation load group (group 3 (FBI)) were node positive while only 24% of the patients in a higher mutation load group were node positive (group 1 (HRD + S-Dup)).

3. Younger patients were identified in one of the groups with the highest mutation burden (HRD + S-Dup, group 1) suggestive of more proliferative disease that could lead to a higher rate at which tumours accumulate mutations in younger patients.

4. Low grade tumours were found associated with a low mutation load subgroup (group 3 (FBI)).

5. Tumour size has no bearing on node positivy as seen with groups 1 (HRD + S-Dup) (24% node +ve, average tumour_ size = 2.7), 3 (FBI) (67% node +ve, average tumour_size = 2.4).

** All the above deductions pend validation with a larger cohort. This will also go a long way in identifying a solid association between SMGs and clinical outcomes.

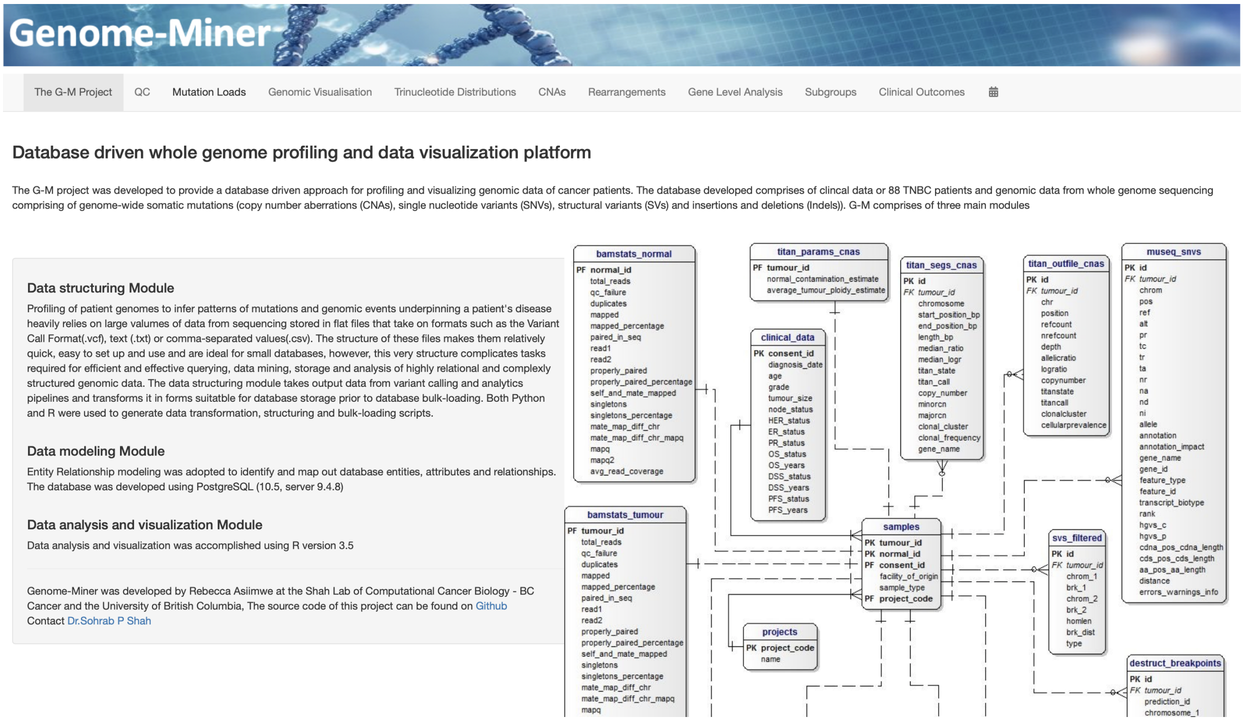
Chapter 4

Data Access and Visualization Interface

Current advances in sequencing technologies have led to the generation of vast amounts of sequencing data that have come with a salient need for data management, access, analysis and visualisation. Secondly, as genomic research has become increasingly collaborative, it has become crucial to access and share data in a way that is understandable to research teams, both technical and non technical. We extended this functionality to the implemented database by developing Genome-Miner, a flexible, convenient, and interactive web-based database interface to support global data access, interactive exploration, querying, analysis, visualization and sharing of clinical outcomes and whole genome profiling data in the developed database, results of which are rendered dynamically into the front-end web application and directly from the database for utilisation by researchers, biologists and clinicians using intuitive and interactive plots and data tables, as shown in this chapter. The developed interface also allows researchers to download all results and upload files for data analysis and visualization without the need of any programming knowledge. This establishment will go a long way in helping researchers generate novel insights and hypotheses by visualizing clinical outcomes and genomic variant data on CNAs, SNVs, SVs and indels.

4.1 Genome-Miner

The front page to the developed platform provides users with the objective and overview of Genome-Miner (Fig. 4.1). This page also provides navigational links to the interface implementations of the analyses conducted on the data in the database based on quality control, mutation burden, genomic visualization, trinucleotide distributions, CNAs, rearrangements, gene-level analysis, subgroup discovery and clinical outcomes analyses and visualizations.



Database driven whole genome profiling and data visualization platform

The G-M project was developed to provide a database driven approach for profiling and visualizing genomic data of cancer patients. The database developed comprises of clincal data or 88 TNBC patients and genomic data from whole genome sequencing comprising of genome-wide somatic mutations (copy number aberrations (CNAs), single nucleotide variants (SNVs), structural variants (SVs) and insertions and deletions (Indels)). G-M comprises of three main modules

Data structuring Module

Profiling of patient genomes to infer patterns of mutations and genomic events underpinning a patient's disease heavily relies on large valumes of data from sequencing stored in flat files that take on formats such as the Variant Call Format(.vcf), text (.txt) or comma-separated values(.csv). The structure of these files makes them relatively quick, easy to set up and use and are ideal for small databases, however, this very structure complicates tasks required for efficient and effective querying, data mining, storage and analysis of highly relational and complexly structured genomic data. The data structuring module takes output data from variant calling and analytics pipelines and transforms it in forms suitatble for database storage prior to database bulk-loading. Both Python and R were used to generate data transformation, structuring and bulk-loading scripts.

Data modeling Module

Entity Relationship modeling was adopted to identify and map out database entities, attributes and relationships. The database was developed using PostgreSQL (10.5, server 9.4.8)

Data analysis and visualization Module

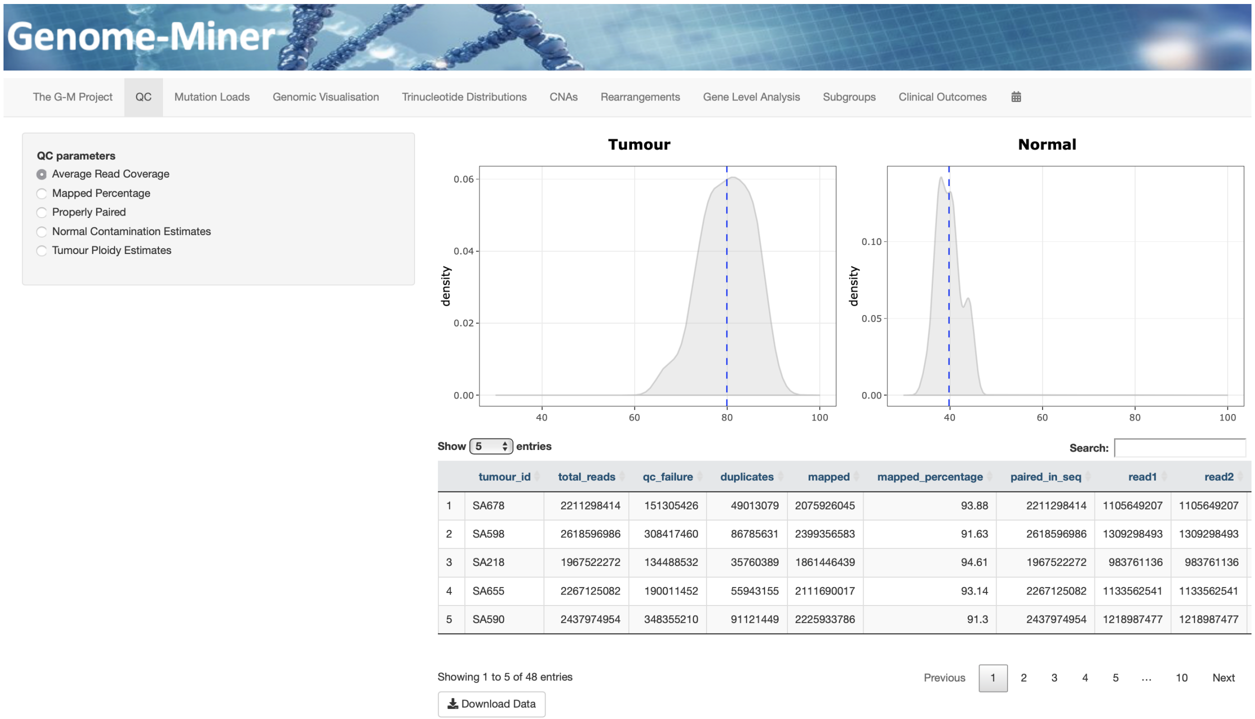
Data analysis and visualization was accomplished using R version 3.5

Genome-Miner was developed by Rebecca Asiimwe at the Shah Lab of Computational Cancer Biology - BC Cancer and the University of British Columbia, The source code of this project can be found on Github
Contact Dr.Sohrab P Shah

Figure 4.1: Genome-Miner: Front page showing navigational links to the main analysis and visualisation themes availed through this platform (QC, Mutation Loads, Genomic Visualization, Trinucleotide Distributions, CNAs, Rearrangements, Gene Level Analysis, Subgroups and Clinical Outcomes).

4.2 Quality control analyses and visualizations

a



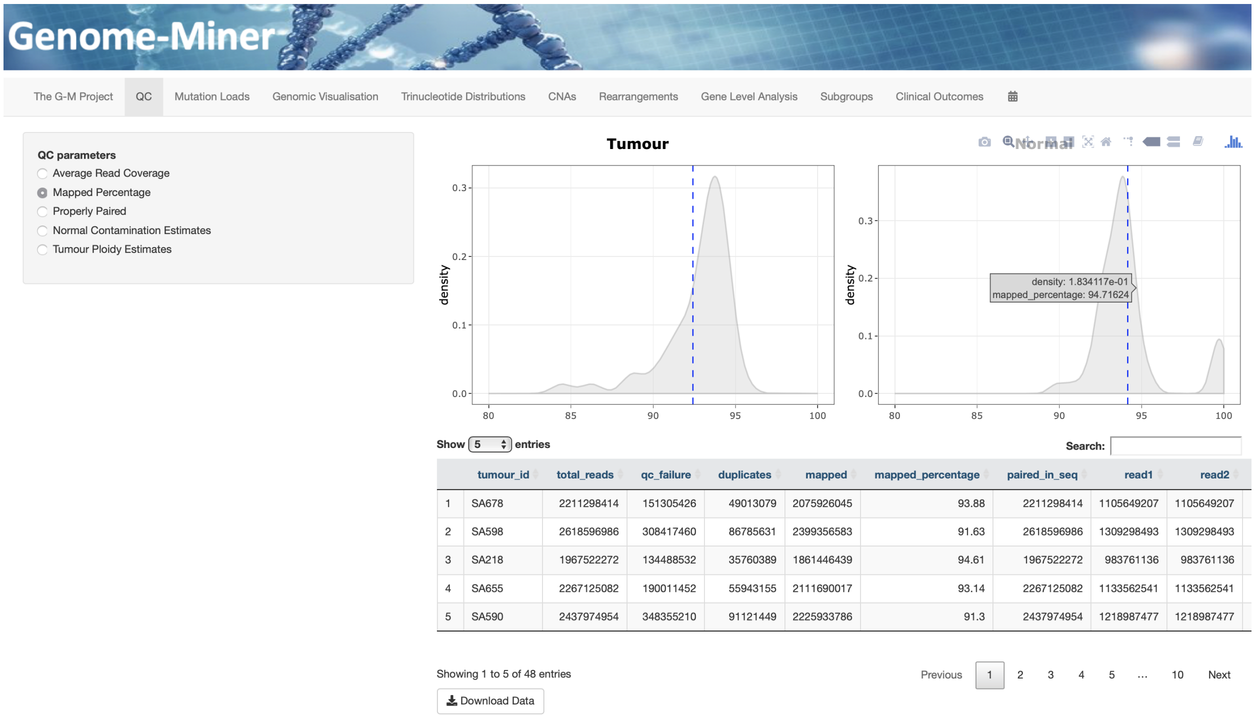
QC parameters

● Average Read Coverage
○ Mapped Percentage
○ Properly Paired
○ Normal Contamination Estimates
○ Tumour Ploidy Estimates

	tumour_id	total_reads	qc_failure	duplicates	mapped	mapped_percentage	paired_in_seq	read1	read2
1	SA678	2211298414	151305426	49013079	2075926045	93.88	2211298414	1105649207	1105649207
2	SA598	2618596986	308417460	86785631	2399356583	91.63	2618596986	1309298493	1309298493
3	SA218	1967522272	134488532	35760389	1861446439	94.61	1967522272	983761136	983761136
4	SA655	2267125082	190011452	55943155	2111690017	93.14	2267125082	1133562541	1133562541
5	SA590	2437974954	348355210	91121449	2225933786	91.3	2437974954	1218987477	1218987477

Showing 1 to 5 of 48 entries

b

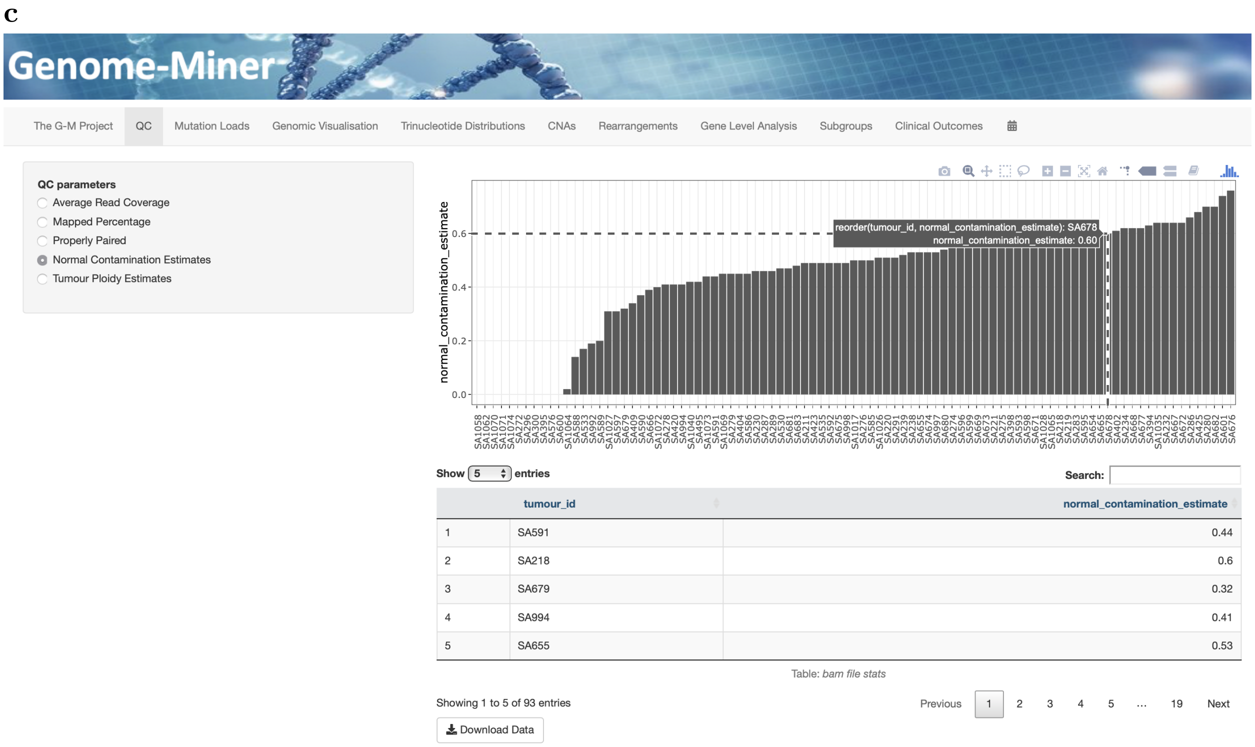


QC parameters

○ Average Read Coverage
● Mapped Percentage
○ Properly Paired
○ Normal Contamination Estimates
○ Tumour Ploidy Estimates

	tumour_id	total_reads	qc_failure	duplicates	mapped	mapped_percentage	paired_in_seq	read1	read2
1	SA678	2211298414	151305426	49013079	2075926045	93.88	2211298414	1105649207	1105649207
2	SA598	2618596986	308417460	86785631	2399356583	91.63	2618596986	1309298493	1309298493
3	SA218	1967522272	134488532	35760389	1861446439	94.61	1967522272	983761136	983761136
4	SA655	2267125082	190011452	55943155	2111690017	93.14	2267125082	1133562541	1133562541
5	SA590	2437974954	348355210	91121449	2225933786	91.3	2437974954	1218987477	1218987477

Showing 1 to 5 of 48 entries


c
Genome-Miner
The G-M Project
QC
Mutation Loads
Genomic Visualisation
Trinucleotide Distributions
CNAs
Rearrangements
Gene Level Analysis
Subgroups
Clinical Outcomes
QC parameters
○ Average Read Coverage
○ Mapped Percentage
○ Properly Paired
● Normal Contamination Estimates
○ Tumour Ploidy Estimates
normal_contamination_estimate
0.6
0.4
0.2
0.0
reorder(tumour_id, normal_contamination_estimate): SA678
normal_contamination_estimate: 0.60
SA1058 SA1062 SA1070 SA1071 SA1074 SA272 SA296 SA300 SA395 SA576 SA600 SA1064 SA588 SA533 SA992 SA589 SA1027 SA597 SA679 SA409 SA590 SA666 SA1072 SA278 SA420 SA994 SA1040 SA495 SA1073 SA591 SA1069 SA279 SA404 SA586 SA230 SA287 SA289 SA530 SA681 SA683 SA211 SA423 SA535 SA592 SA675 SA998 SA1017 SA276 SA585 SA1026 SA220 SA291 SA239 SA238 SA655 SA674 SA997 SA680 SA274 SA596 SA599 SA669 SA673 SA221 SA275 SA398 SA593 SA598 SA671 SA1028 SA1065 SA218 SA219 SA283 SA595 SA654 SA665 SA678 SA402 SA234 SA668 SA677 SA394 SA1035 SA232 SA667 SA672 SA286 SA425 SA280 SA682 SA601 SA676
Show 5 entries
Search:
Previous
1
2
3
4
5
...
19
Next
Download Data

	tumour_id	normal_contamination_estimate
1	SA591	0.44
2	SA218	0.6
3	SA679	0.32
4	SA994	0.41
5	SA655	0.53

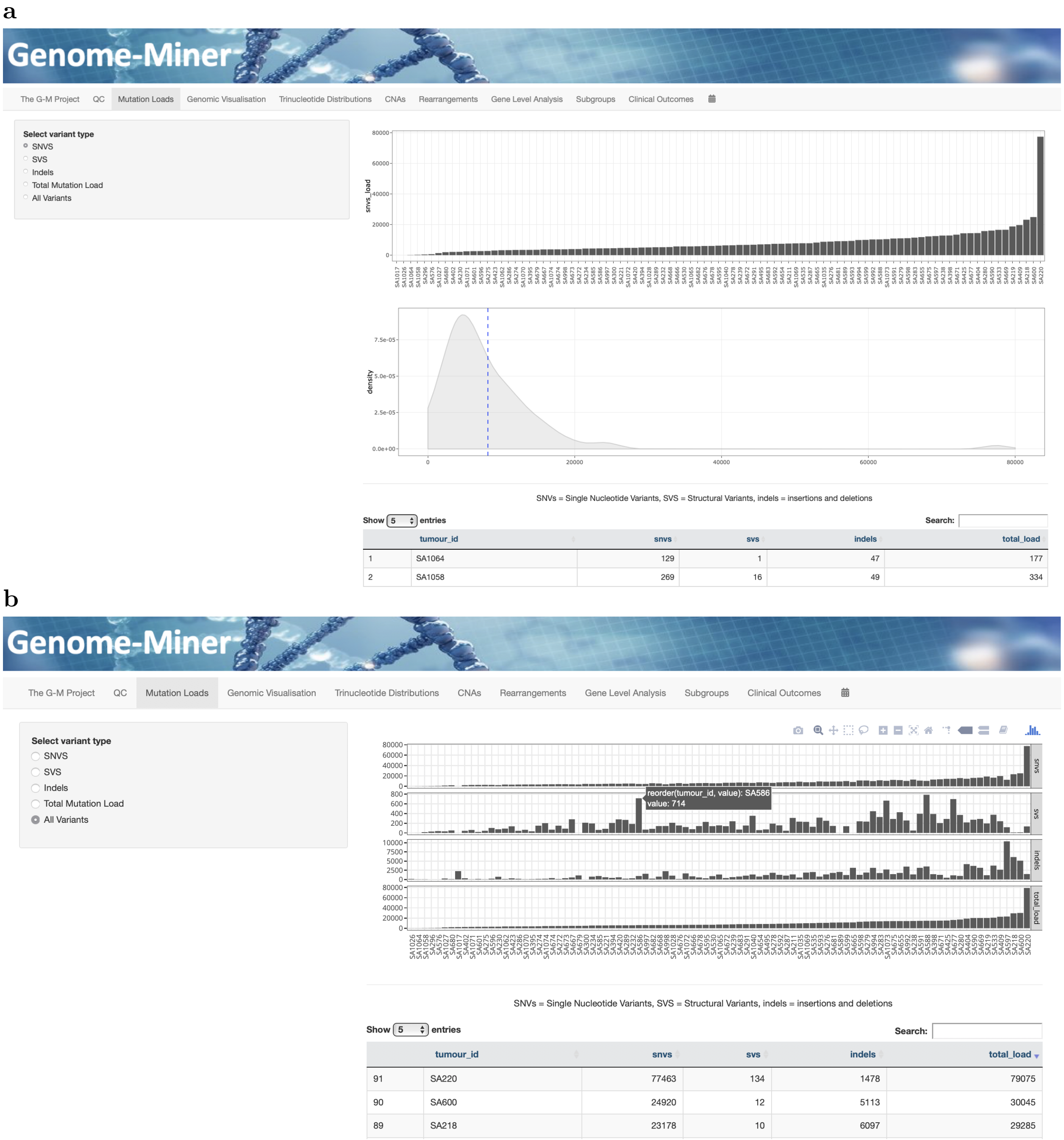
Table: *bam file stats*

Showing 1 to 5 of 93 entries

Figure 4.2: Database interface - user defined QC explorations and visualizations for **a)** average read coverage **b)** mapped percentage and **c)** normal contamination estimates.

QC explorations, analyses and visualizations are based on 5 main parameters: average read coverage, mapped percentage, properly paired reads, normal contamination estimates and tumour ploidy estimates. These analyses are based on sequencing statistics extracted from bam files (average read coverage, mapped percentage, properly paired reads) and Titan output files (normal contamination and ploidy estimates) providing an overview of the distribution of specified parameters across the cohort. More data on each of the samples in the cohort is availed through data tables that can be manipulated to dynamically trigger visualizations of interest.

4.3 Mutation load analyses and visualizations per sample and across the cohort



a

Select variant type

- ● SNVS
- ○ SVS
- ○ Indels
- ○ Total Mutation Load
- ○ All Variants

SNVs = Single Nucleotide Variants, SVS = Structural Variants, indels = insertions and deletions

Show 5 entries Search:

	tumour_id	snvs	svs	indels	total_load
1	SA1064	129	1	47	177
2	SA1058	269	16	49	334

b

Select variant type

- ○ SNVS
- ○ SVS
- ○ Indels
- ○ Total Mutation Load
- ● All Variants

SNVs = Single Nucleotide Variants, SVS = Structural Variants, indels = insertions and deletions

Show 5 entries Search:

	tumour_id	snvs	svs	indels	total_load
91	SA220	77463	134	1478	79075
90	SA600	24920	12	5113	30045
89	SA218	23178	10	6097	29285

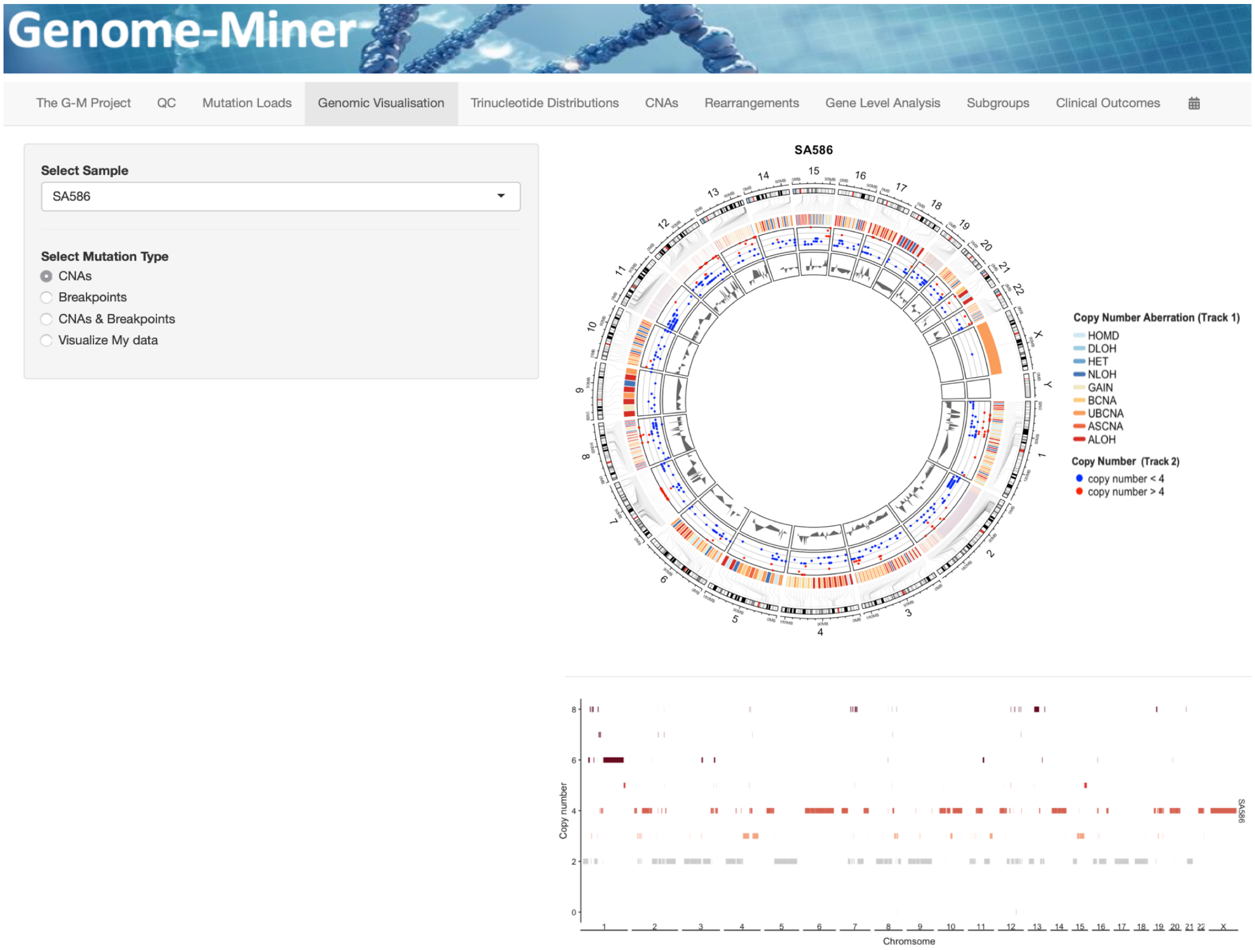
Figure 4.3: Database interface - mutation loads: User defined explorations, analyses and visualizations for **a)** SNV mutation loads and **b)** total mutation loads.

Interface enabled explorations, analyses and visualizations of mutation loads per sample and across the cohort based on variant types: SNVs, SVs, indels and the total mutation load are shown in Fig. 4.3. Data tables provide sample specific mutation loads that can be ordered (ascending or descending) by variant type and based on a user's visualization preference. Users also have the ability to refer to the average read coverage for cases of interest.

4.4 Genomic visualizations

The database interface also enables user-triggered visualizations for genomic events across a patient's chromosome by CNAs (Fig. 4.4 **a**)), breakpoints or by a combination of CNAs and breakpoints (Fig. 4.4 **b**)). A user specifies two parameters: the sample of interest and the mutation type, parameters that are subsequently used to generate and render circos plots on the fly as is the case with all other plots.

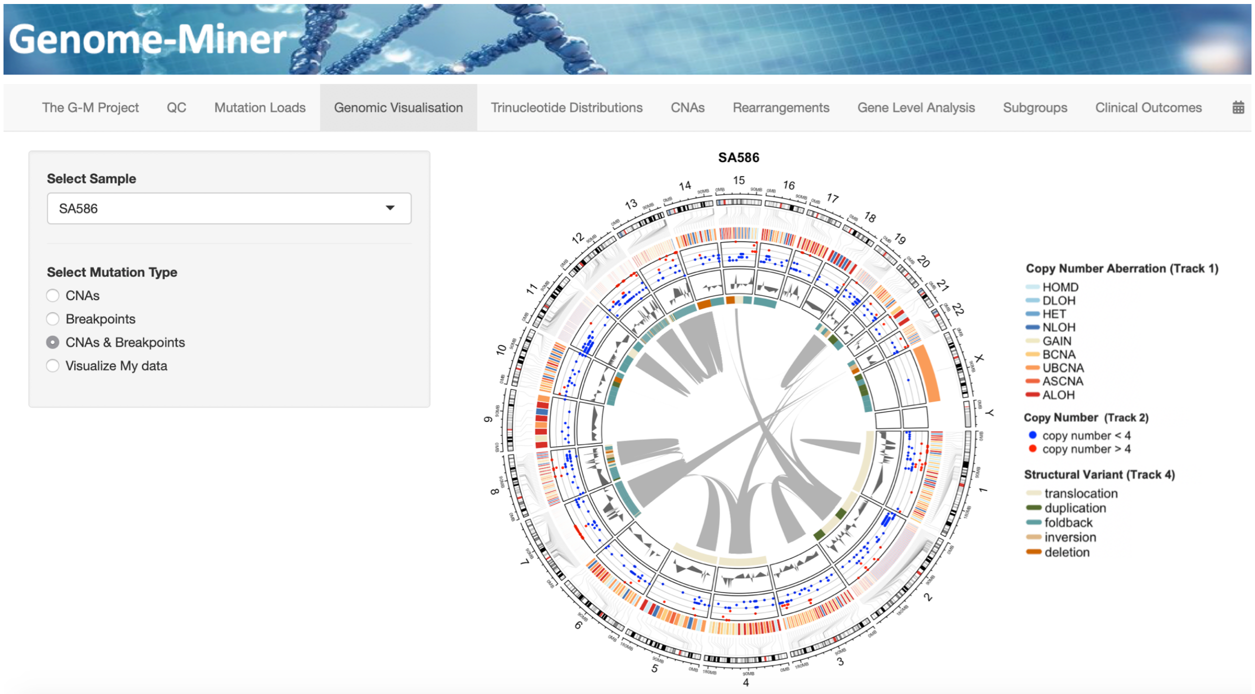
a



Select Sample
SA586

Select Mutation Type
● CNAs
○ Breakpoints
○ CNAs & Breakpoints
○ Visualize My data

b



Select Sample

SA586

Select Mutation Type

○ CNAs

○ Breakpoints

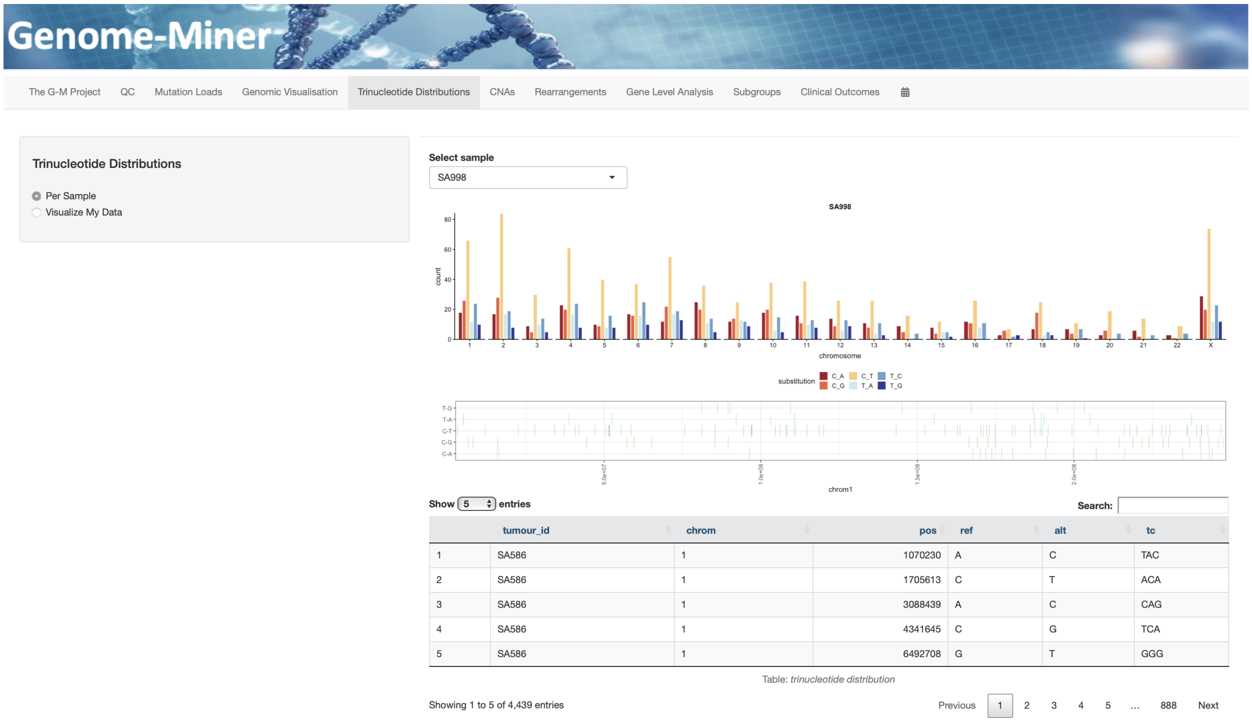
● CNAs & Breakpoints

○ Visualize My data

Figure 4.4: Database interface - genomic visualization: **a)** Genomic visualization by CNAs. **b)** Genomic visualization by CNAs and SNVs.

Fig. 4.4 **a)** (top plot) shows a circos plot rendered to the user where track 1 in the plot shows the type of copy number aberration per genomic position while the copy number variations are shown in track 2 and 3. The bottom plot shows the copy number profile of the sample of interest (case SA586) and shows copy number variations (y-axis) along the genome denoted by coordinates representing genomic positions (x-axis). Fig. 4.4 **b)** on the other hand shows a user-triggered circos plot for both copy number aberrations and structural variants. Structural variations across the chromosome (translocations, duplications, foldback inversions, deletions, and inversions) are shown in track 4 followed by corresponding links between genomic positions. The interface will also support visualizations of multiple circos plots to support comparative visualizations of the genomic structures of multiple cases, for example, visualizations by subgroup.

4.5 Intra-sample trinucleotide distribution



	tumour_id	chrom	pos	ref	alt	tc
1	SA586	1	1070230	A	C	TAC
2	SA586	1	1705613	C	T	ACA
3	SA586	1	3088439	A	C	CAG
4	SA586	1	4341645	C	G	TCA
5	SA586	1	6492708	G	T	GGG

Table: *trinucleotide distribution*

Showing 1 to 5 of 4,439 entries

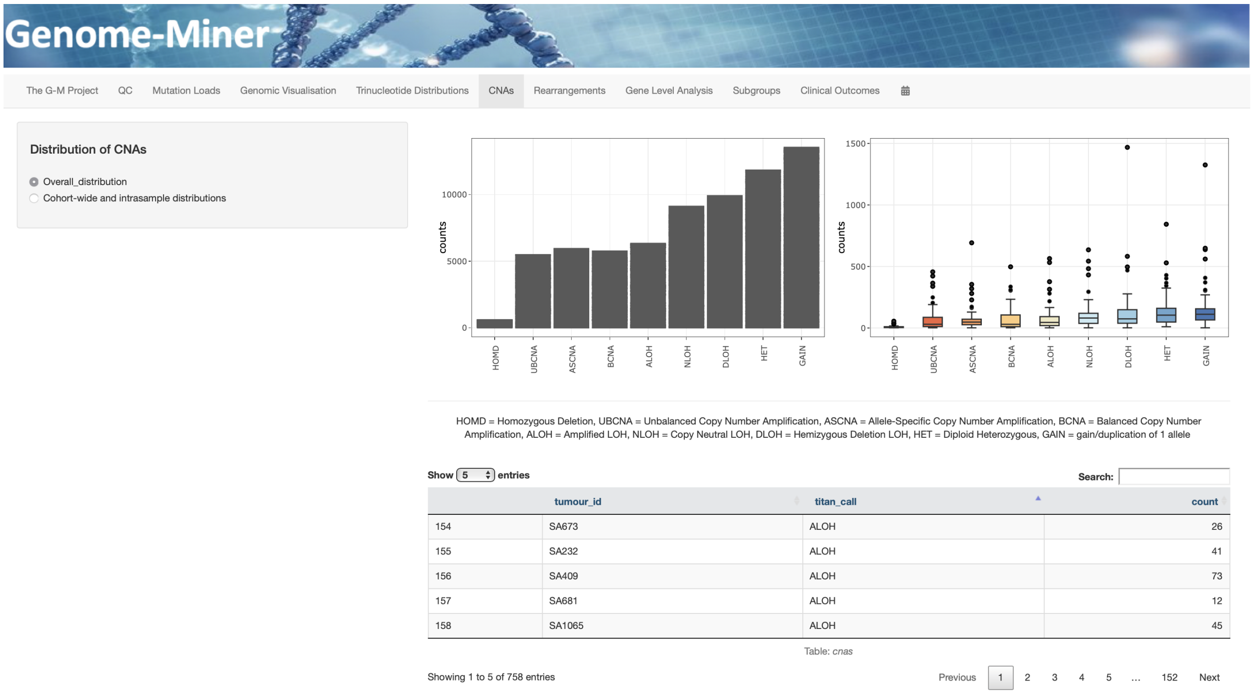
Figure 4.5: Database interface - trinucleotide distribution per sample.

The developed interface supports visualizations of the trinucleotide distribution across the chromosome of each case as shown in Fig. 4.5. Trinclueotide substitution (C>A, C>G, C>T, T>A, T>C, T>G) counts (y-axis) are shown for every chromosome (x-axis) with an extract of required source data and details provided in the data table.

4.6 CNAs analysis and visualizations per sample and across the cohort

Explorations and visualizations of the distribution and abundance of copy number aberrations across the cohort and within each sample as triggered by the interface user are shown in Fig. 4.6 **a)** and **b)** respectively.

a



Distribution of CNAs

● Overall_distribution
○ Cohort-wide and intrasample distributions

HOMD = Homozygous Deletion, UBCNA = Unbalanced Copy Number Amplification, ASCNA = Allele-Specific Copy Number Amplification, BCNA = Balanced Copy Number Amplification, ALOH = Amplified LOH, NLOH = Copy Neutral LOH, DLOH = Hemizygous Deletion LOH, HET = Diploid Heterozygous, GAIN = gain/duplication of 1 allele

Show 5 entries

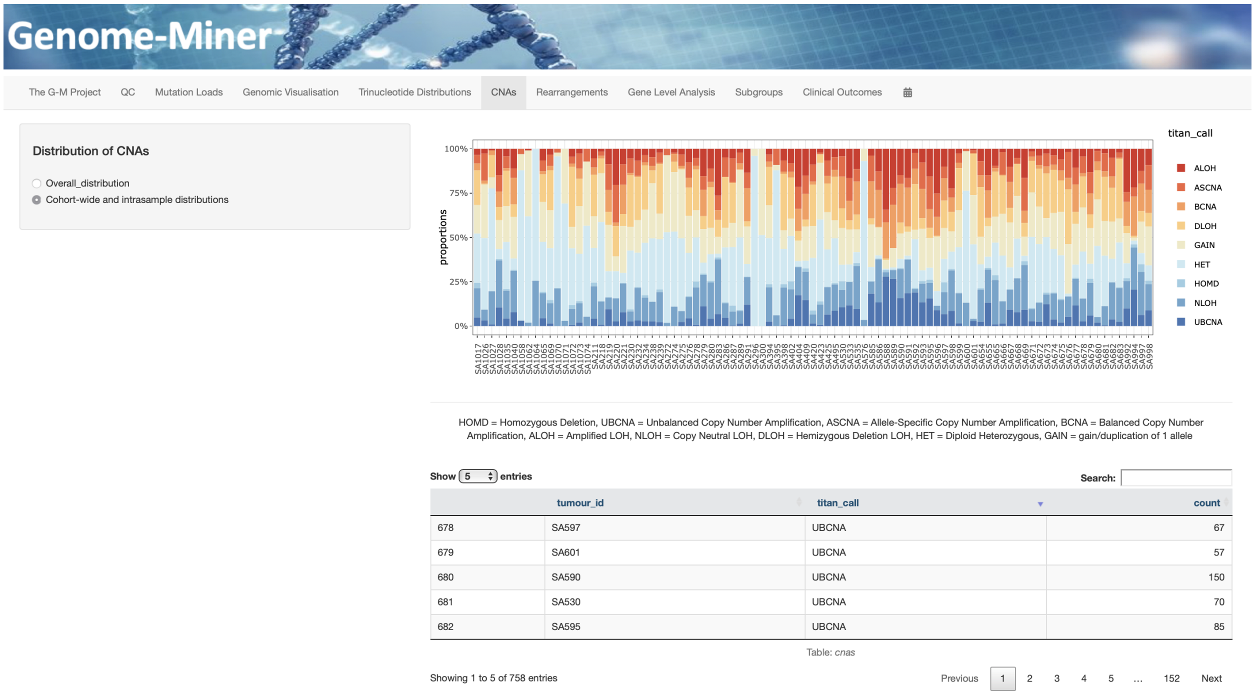
Search:

	tumour_id	titan_call	count
154	SA673	ALOH	26
155	SA232	ALOH	41
156	SA409	ALOH	73
157	SA681	ALOH	12
158	SA1065	ALOH	45

Table: *cnas*

Showing 1 to 5 of 758 entries

b



Distribution of CNAs

○ Overall_distribution
● Cohort-wide and intrasample distributions

HOMD = Homozygous Deletion, UBCNA = Unbalanced Copy Number Amplification, ASCNA = Allele-Specific Copy Number Amplification, BCNA = Balanced Copy Number Amplification, ALOH = Amplified LOH, NLOH = Copy Neutral LOH, DLOH = Hemizygous Deletion LOH, HET = Diploid Heterozygous, GAIN = gain/duplication of 1 allele

Show 5 entries

Search:

	tumour_id	titan_call	count
678	SA597	UBCNA	67
679	SA601	UBCNA	57
680	SA590	UBCNA	150
681	SA530	UBCNA	70
682	SA595	UBCNA	85

Table: *cnas*

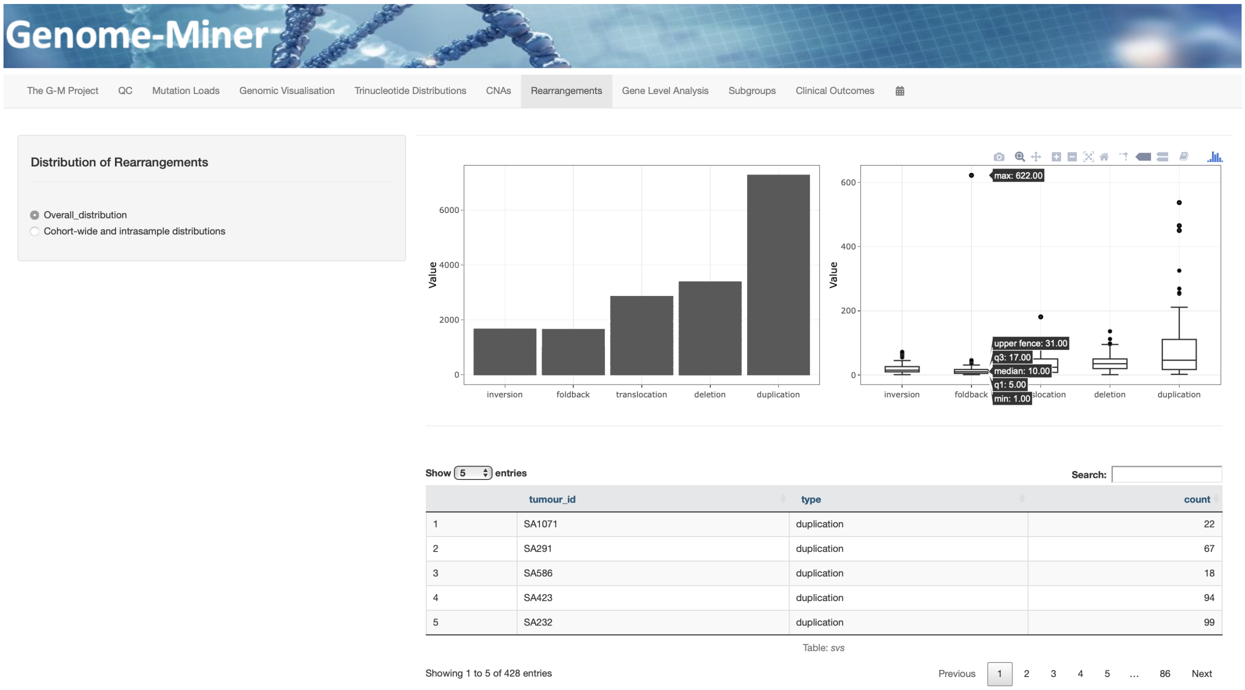
Showing 1 to 5 of 758 entries

Figure 4.6: Database interface - distribution of CNAs: **a)** Overall distribution **b)** Cohort-wide and intrasample distribution.

4.7 SVs analysis and visualizations per sample and across the cohort

The distribution and abundance of structural variants across the cohort and within each sample are shown in Fig. 4.7 **a)** and **b)** respectively. By specifying parameters of interest, the interface renders reactive plots and data tables to support user interaction with both the data and rendered plots as shown in Fig. 4.7.

a



Distribution of Rearrangements

● Overall_distribution
○ Cohort-wide and intrasample distributions

Show 5 entries

Search:

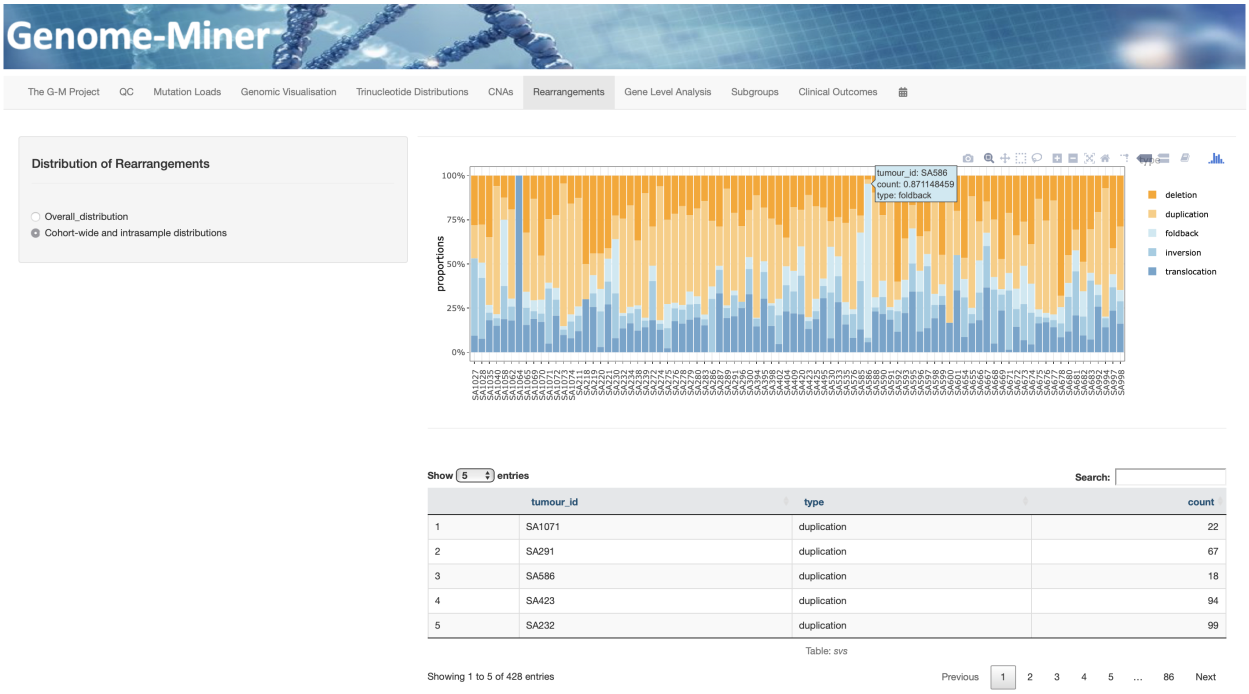
	tumour_id	type	count
1	SA1071	duplication	22
2	SA291	duplication	67
3	SA586	duplication	18
4	SA423	duplication	94
5	SA232	duplication	99

Table: *svs*

Showing 1 to 5 of 428 entries

Previous 1 2 3 4 5 … 86 Next

b



Distribution of Rearrangements

○ Overall_distribution

● Cohort-wide and intrasample distributions

Show 5 entries

Search:

	tumour_id	type	count
1	SA1071	duplication	22
2	SA291	duplication	67
3	SA586	duplication	18
4	SA423	duplication	94
5	SA232	duplication	99

Table: *svs*

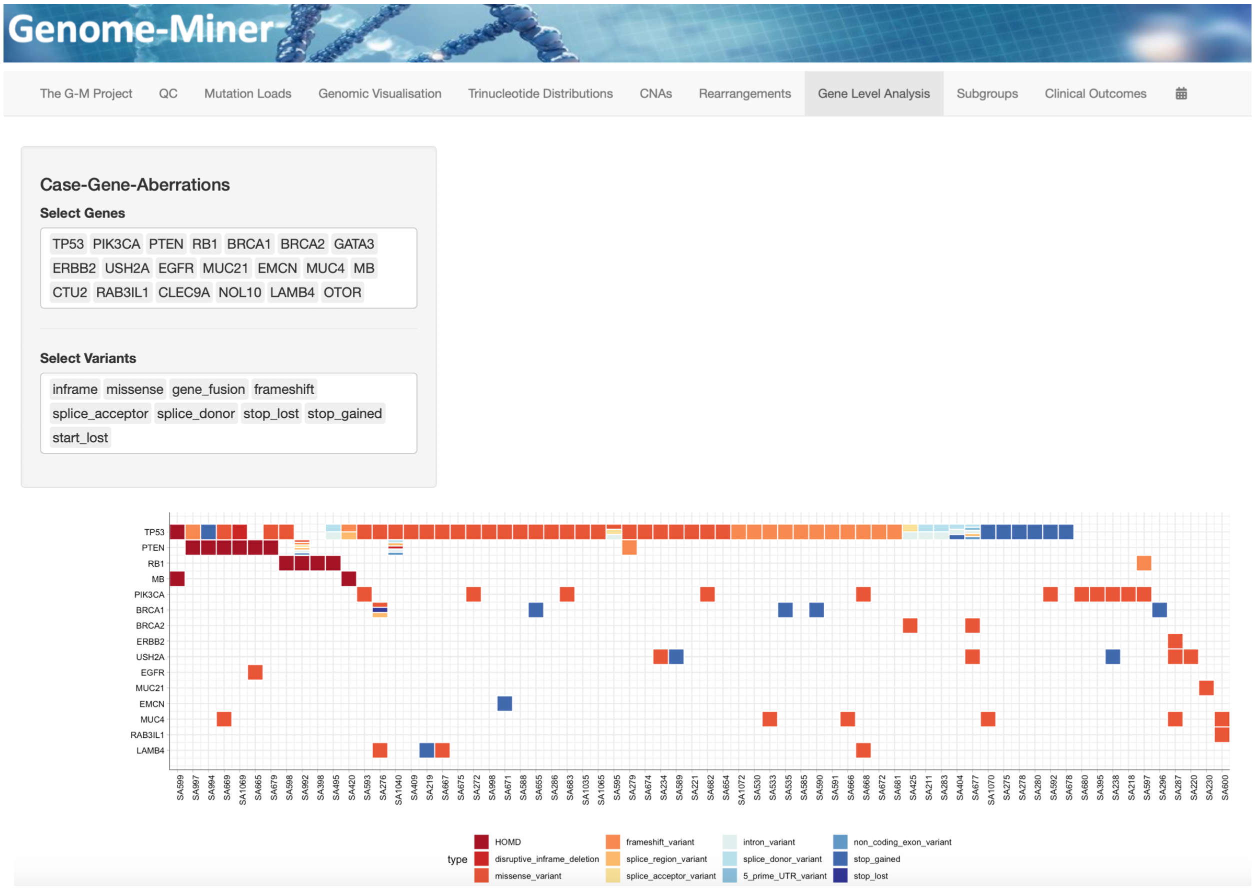
Showing 1 to 5 of 428 entries

Previous 1 2 3 4 5 ... 86 Next

Figure 4.7: Database interface - distribution of SVs: **a)** Overall distribution **b)** Cohort-wide and intrasample distribution.

4.8 Gene-level analysis

Genome-Miner also supports user defined visualizations of mutations in each gene per case (Fig. 4.8). Here, a user specifies genes and variants of interest and an oncoplot showing which cases have the specified mutations in selected genes is generated. Each row represents a gene and each column represents a case. Multiple hits in a gene are represented by multiple colors representative of specific mutation types in a single gene.


Genome-Miner
The G-M Project
QC
Mutation Loads
Genomic Visualisation
Trinucleotide Distributions
CNAs
Rearrangements
Gene Level Analysis
Subgroups
Clinical Outcomes
TP53
PTEN
RB1
MB
PIK3CA
BRCA1
BRCA2
ERBB2
USH2A
EGFR
MUC21
EMCN
MUC4
RAB3IL1
LAMB4
SA599 SA997 SA994 SA669 SA1069 SA665 SA679 SA598 SA992 SA398 SA495 SA420 SA593 SA276 SA1040 SA409 SA219 SA667 SA675 SA272 SA998 SA671 SA588 SA655 SA286 SA683 SA1035 SA1065 SA595 SA279 SA674 SA234 SA589 SA221 SA682 SA654 SA1072 SA530 SA533 SA535 SA585 SA590 SA591 SA666 SA668 SA672 SA681 SA425 SA211 SA283 SA404 SA677 SA1070 SA275 SA278 SA280 SA592 SA678 SA680 SA395 SA238 SA218 SA597 SA296 SA287 SA220 SA230 SA600
type
HOMD
disruptive_inframe_deletion
missense_variant
frameshift_variant
splice_region_variant
splice_acceptor_variant
intron_variant
splice_donor_variant
5_prime_UTR_variant
non_coding_exon_variant
stop_gained
stop_lost

Case-Gene-Aberrations

Select Genes

TP53 PIK3CA PTEN RB1 BRCA1 BRCA2 GATA3 ERBB2 USH2A EGFR MUC21 EMCN MUC4 MB CTU2 RAB3IL1 CLEC9A NOL10 LAMB4 OTOR

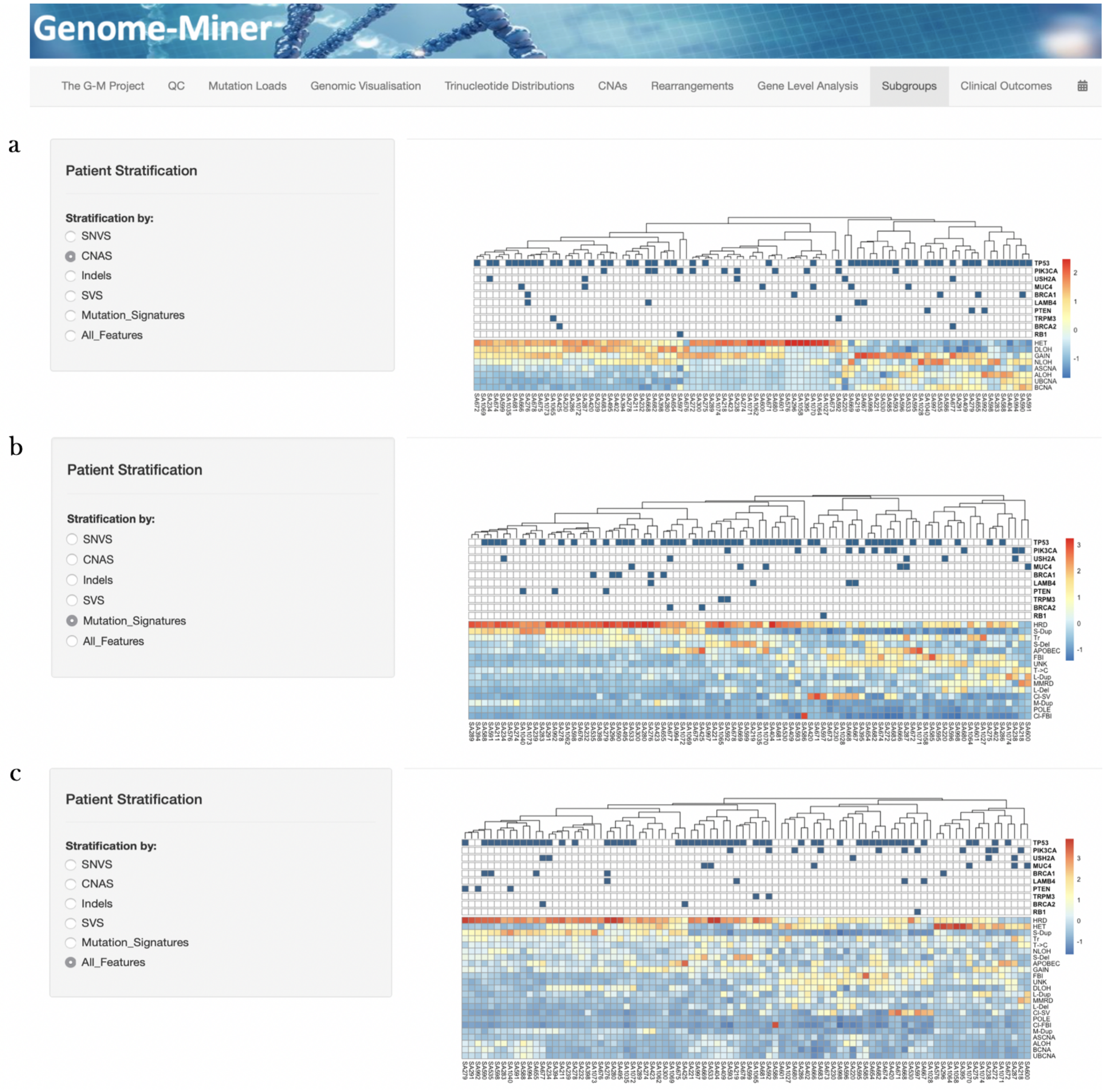
Select Variants

inframe missense gene_fusion frameshift splice_acceptor splice_donor stop_lost stop_gained start_lost

Figure 4.8: Database interface - gene-level analysis.

4.9 TNBC subgroup analysis and visualizations

Database-derived stratification of patients as enabled by the developed interface (Fig. 4.9) can be done based on the user's choice of individual variants: SNVs, CNAs, indels, SVs, mutation signatures or by the integration of all features (SNVs, CNAs, indels, SVs and mutation signatures). For inclusion will be user-based selections of genes of interest to apply for annotating the clustered heatmap and choices of variants from various variant domains for inclusion as stratification features.



a

Patient Stratification

Stratification by:

○ SNVS
● CNAS
○ Indels
○ SVS
○ Mutation_Signatures
○ All_Features

b

Patient Stratification

Stratification by:

○ SNVS
○ CNAS
○ Indels
○ SVS
● Mutation_Signatures
○ All_Features

c

Patient Stratification

Stratification by:

○ SNVS
○ CNAS
○ Indels
○ SVS
○ Mutation_Signatures
● All_Features

Figure 4.9: Database interface - subgroup discovery by **a)** CNAs **b)** mutation signatures and **c)** integrated genomic features.

Chapter 5

Conclusions and Future Work

Relational databases have long been used as an indispensable tool in modelling and organizing vast amounts of data, including biological data. Currently, profiling of patient genomes to infer patterns of mutations and genomic events underpinning a patient's disease heavily relies on data stored in flat files whose structure complicates tasks required for analyzing relational and complexly structured genomic data. To the best of our knowledge, this is the first research of its kind that solves this problem by implementing a database driven approach to integrate data from whole genome sequences with clinical outcomes for the exploration of the genomic landscapes and mutation characteristics underpinning cancers. The developed clinical outcomes and genomic variants database was further applied to support the mining and comprehensive analysis of clinical outcomes and whole genome profiles of 88 TNBC patients. Functionality of the database was extended to support global data access, interactive exploration, querying, analysis, visualization and sharing of collected data among various research groups through the birth of Genome-Miner, a flexible, convenient, and interactive web-based platform.

We demonstrate the applicability of the database to effectively support and enforce quality control checks and measures by filtering data to meet down stream analysis requirements. We also demonstrate the utility of the developed database for the exploration and analysis of somatic alterations (CNAs, SNVs, SVs and indels), results of which show a varying distribution of mutation loads in this TNBC cohort. Also identified was the variation in the mutation type loads within each sample and the heterogeneity of TNBCs at CNA, SV, SNV and indel level. In this study cohort, *TP53* (56.9%), *PIK3CA* (8.9%), *PTEN* (7.3%), *BRCA1* (5.7%), *USH2A* (4.9%), *MUC4* (4.9%) and *RB1* (4.1%) were identified as the most frequently mutated genes.

We further applied the database to mine and compute genomic features used for patient stratification and demonstrate for the first time and to the best of our knowledge, the discovery of 5 putative and distinct TNBC genomic subgroups revealed by 23 significant genomic features, 8 of which were mined and computed from the developed database (proportions of: HET, GAIN, DLOH, NLOH, ASCNA, ALOH, BCNA, UBCNA) and the other 15 (mutation signatures: POLE, APOBEC, HRD, UNK, MMRD, T→C, M-Dup, S-Del, Cl-SV, FBI, Cl-FBI, L-Del, S-Dup, Tr and L-Dup) derived using the multi-modal correlated topic model (MMCTM). Each of the identified subgroups exhibited distinct genomic and clinical characteristics.

Results from this research show and confirm our hypothesis that TNBC patients can be stratified into distinct subgroups based on their whole genome profiles and that the identified TNBC subgroups exhibit distinct clinical and genomic characteristics. Our results also show that mutation signatures enriched in identified subgroups associate with specific driver mutations or mutations in DNA damage repair genes. These results provide an improved understanding of TNBCs and will further provide valuable insights into subgroup specific clinically actionable events, options for novel therapeutic modalities and the identification of patients most likely to respond to specific modalities. These results also show for the utility of the genome as a potential discriminant biomarker in patient treatment.

5.1 Limitations and future work

The research herein focused on the analysis of whole genome profiles and clinical outcomes of TNBC patients, however, the elucidation of TNBC subgroups and their respective characteristics further requires a multi-omics approach that integrates and analyses data from all platforms (DNA methylation, messenger RNA arrays, exome sequencing, microRNA sequencing and reverse-phase protein arrays) for the discovery of more informative subgroups. Also, the analysis of a larger cohort is still needed to provide more insights into the genomic underpinnings of TBCs and to corroborate the identified subgroups in this study or identify additional subgroups in TNBC.

Secondly, the database was structured to suit the data output from respective variant callers (mutationSeq, Titan, deStruct, Lumpy and Strelka). A more generic approach of storing and mining variants discovered using other variant calling tools other than those used in this study is needed. Also, to note is that the variants in this study were annotated using snpEff. A more generic solution will require the mapping and inclusion of annotations from other variant annotation tools such as VEP.

Future work will involve implementing a more generalizable database and bulk-loading genomic data from other cancers such as ovary into the database to support integrated analyses and inferences from different cancer types. Prospects to include more data such as gene expression data are underway.

Bibliography

[1] C. G. A. Network *et al.*, “Comprehensive molecular portraits of human breast tumours,” *Nature*, vol. 490, no. 7418, p. 61, 2012.

[2] C. M. Perou, T. Sørlie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen *et al.*, “Molecular portraits of human breast tumours,” *nature*, vol. 406, no. 6797, p. 747, 2000.

[3] Z. Hu, C. Fan, D. S. Oh, J. S. Marron, X. He, B. F. Qaqish, C. Livasy, L. A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. G. Ewend, L. R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Ruiz Orrico, D. Dreher, J. P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. F. Quackenbush, M. J. Ellis, O. I. Olopade, P. S. Bernard, and C. M. Perou, “The molecular portraits of breast tumors are conserved across microarray platforms,” *BMC Genomics*, vol. 7, p. 96, Apr 2006.

[4] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, “Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies,” *The Journal of clinical investigation*, vol. 121, no. 7, pp. 2750–2767, 2011.

[5] C. M. Perou, “Molecular stratification of triple-negative breast cancers,” *Oncologist*, vol. 16 Suppl 1, pp. 61–70, 2011.

[6] F. Bertucci, P. Finetti, and D. Birnbaum, “Basal breast cancer: a complex and deadly molecular subtype,” *Curr. Mol. Med.*, vol. 12, no. 1, pp. 96–110, Jan 2012.

[7] G. J. Logan, D. J. Dabbs, P. C. Lucas, R. C. Jankowitz, D. D. Brown, B. Z. Clark, S. Oesterreich, and P. F. McAuliffe, “Molecular drivers of lobular carcinoma in situ,” *Breast Cancer Research*, vol. 17, no. 1, p. 76, 2015.

[8] D. Krug and R. Souchon, “Radiotherapy of ductal carcinoma in situ,” *Breast Care*, vol. 10, no. 4, pp. 259–264, 2015.

[9] G. K. Malhotra, X. Zhao, H. Band, and V. Band, “Histological, molecular and functional subtypes of breast cancers,” *Cancer biology & therapy*, vol. 10, no. 10, pp. 955–960, 2010.

[10] E. A. Rakha, J. S. Reis-Filho, F. Baehner, D. J. Dabbs, T. Decker, V. Eusebi, S. B. Fox, S. Ichihara, J. Jacquemier, S. R. Lakhani *et al.*, “Breast cancer prognostic classification in the molecular era: the role of histological grade,” *Breast Cancer Research*, vol. 12, no. 4, p. 207, 2010.

[11] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan *et al.*, “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” *Nature*, vol. 486, no. 7403, p. 346, 2012.

[12] E. Lerma, G. Peiro, T. Ramón, S. Fernandez, D. Martinez, C. Pons, F. Munoz, J. M. Sabate, C. Alonso, B. Ojeda *et al.*, “Immunohistochemical heterogeneity of breast carcinomas negative for estrogen receptors, progesterone receptors and her2/neu (basal-like breast carcinomas),” *Modern Pathology*, vol. 20, no. 11, p. 1200, 2007.

[13] N. A. Makretsov, D. G. Huntsman, T. O. Nielsen, E. Yorida, M. Peacock, M. C. Cheang, S. E. Dunn, M. Hayes, M. van de Rijn, C. Bajdik *et al.*, “Hierarchical clustering analysis of tissue microarray immunostaining data identifies prognostically significant groups of breast carcinoma,” *Clinical cancer research*, vol. 10, no. 18, pp. 6143–6151, 2004.

[14] J. Reis-Filho and A. Tutt, “Triple negative tumours: a critical review,” *Histopathology*, vol. 52, no. 1, pp. 108–118, 2008.

[15] B. G. Haffty, Q. Yang, M. Reiss, T. Kearney, S. A. Higgins, J. Weidhaas, L. Harris, W. Hait, and D. Toppmeyer, “Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer,” *Journal of clinical oncology*, vol. 24, no. 36, pp. 5652–5657, 2006.

[16] J. D. Marotti, F. B. de Abreu, W. A. Wells, and G. J. Tsongalis, “Triple-negative breast cancer: next-generation sequencing for target identification,” *The American journal of pathology*, vol. 187, no. 10, pp. 2133–2138, 2017.

[17] R. Dent, M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod, "Triple-negative breast cancer: clinical features and patterns of recurrence," *Clinical cancer research*, vol. 13, no. 15, pp. 4429–4434, 2007.

[18] V. G. Abramson, B. D. Lehmann, T. J. Ballinger, and J. A. Pietenpol, "Subtyping of triple-negative breast cancer: implications for therapy," *Cancer*, vol. 121, no. 1, pp. 8–16, 2015.

[19] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston *et al.*, "Race, breast cancer subtypes, and survival in the carolina breast cancer study," *Jama*, vol. 295, no. 21, pp. 2492–2502, 2006.

[20] K. R. Bauer, M. Brown, R. D. Cress, C. A. Parise, and V. Caggiano, "Descriptive analysis of estrogen receptor (er)-negative, progesterone receptor (pr)-negative, and her2-negative invasive breast cancer, the so-called triple-negative phenotype: a population-based study from the california cancer registry," *Cancer*, vol. 109, no. 9, pp. 1721–1728, 2007.

[21] H. G. Kaplan, J. A. Malmgren, and M. Atwood, "T1n0 triple negative breast cancer: risk of recurrence and adjuvant chemotherapy," *The breast journal*, vol. 15, no. 5, pp. 454–460, 2009.

[22] H. Kennecke, R. Yerushalmi, R. Woods, M. C. U. Cheang, D. Voduc, C. H. Speers, T. O. Nielsen, and K. Gelmon, "Metastatic behavior of breast cancer subtypes," *Journal of clinical oncology*, vol. 28, no. 20, pp. 3271–3277, 2010.

[23] R. Dent, W. M. Hanna, M. Trudeau, E. Rawlinson, P. Sun, and S. A. Narod, "Pattern of metastatic spread in triple-negative breast cancer," *Breast cancer research and treatment*, vol. 115, no. 2, pp. 423–428, 2009.

[24] N. U. Lin, J. R. Bellon, and E. P. Winer, "Cns metastases in breast cancer," *Journal of clinical oncology*, vol. 22, no. 17, pp. 3608–3617, 2004.

[25] F. Heitz, P. Harter, A. Traut, H. Lueck, B. Beutel, and A. du Bois, "Cerebral metastases (cm) in breast cancer (bc) with focus on triple-negative tumors," *Journal of Clinical Oncology*, vol. 26, no. 15_suppl, pp. 1010–1010, 2008.

[26] G. Von Minckwitz, M. Untch, J.-U. Blohmer, S. D. Costa, H. Eidtmann, P. A. Fasching, B. Gerber, W. Eiermann, J. Hilfrich, J. Huober *et al.*, "Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes," *J Clin oncol*, vol. 30, no. 15, pp. 1796–1804, 2012.

[27] A. Lee and M. B. Djamgoz, “Triple negative breast cancer: emerging therapeutic modalities and novel combination therapies,” *Cancer treatment reviews*, vol. 62, pp. 110–122, 2018.

[28] M. D. Burstein, A. Tsimelzon, G. M. Poage, K. R. Covington, A. Contreras, S. A. Fuqua, M. I. Savage, C. K. Osborne, S. G. Hilsenbeck, J. C. Chang *et al.*, “Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer,” *Clinical Cancer Research*, vol. 21, no. 7, pp. 1688–1698, 2015.

[29] J. Stagg and B. Allard, “Immunotherapeutic approaches in triple-negative breast cancer: latest research and clinical prospects,” *Therapeutic advances in medical oncology*, vol. 5, no. 3, pp. 169–181, 2013.

[30] B. D. Lehmann, B. Jovanović, X. Chen, M. V. Estrada, K. N. Johnson, Y. Shyr, H. L. Moses, M. E. Sanders, and J. A. Pietenpol, “Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection,” *PloS one*, vol. 11, no. 6, p. e0157368, 2016.

[31] R. Gao, A. Davis, T. O. McDonald, E. Sei, X. Shi, Y. Wang, P.-C. Tsai, A. Casasent, J. Waters, H. Zhang *et al.*, “Punctuated copy number evolution and clonal stasis in triple-negative breast cancer,” *Nature genetics*, vol. 48, no. 10, p. 1119, 2016.

[32] S. P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari *et al.*, “The clonal and mutational evolution spectrum of primary triple-negative breast cancers,” *Nature*, vol. 486, no. 7403, p. 395, 2012.

[33] X. Wang and C. Guda, “Integrative exploration of genomic profiles for triple negative breast cancer identifies potential drug targets,” *Medicine*, vol. 95, no. 30, 2016.

[34] C. J. Lord and A. Ashworth, “Brcaness revisited,” *Nature Reviews Cancer*, vol. 16, no. 2, p. 110, 2016.

[35] O. A. Stefansson, J. G. Jonasson, O. T. Johannsson, K. Olafsdottir, M. Steinarsdottir, S. Valgeirsdottir, and J. E. Eyfjord, “Genomic profiling of breast tumours in relation to brca abnormalities and phenotypes,” *Breast Cancer Research*, vol. 11, no. 4, p. R47, 2009.

[36] K. A. Kwei, Y. Kung, K. Salari, I. N. Holcomb, and J. R. Pollack, “Genomic instability in breast cancer: pathogenesis and clinical implications,” *Molecular oncology*, vol. 4, no. 3, pp. 255–266, 2010.

[37] M. L. Telli, K. M. Timms, J. Reid, B. Hennessy, G. B. Mills, K. C. Jensen, Z. Szallasi, W. T. Barry, E. P. Winer, N. M. Tung *et al.*, “Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer,” *Clinical cancer research*, 2016.

[38] N. Turner, A. Tutt, and A. Ashworth, “Hallmarks of’brcaness’ in sporadic cancers,” *Nature reviews cancer*, vol. 4, no. 10, p. 814, 2004.

[39] M. K. Graeser, A. McCarthy, C. J. Lord, K. Savage, M. Hills, J. Salter, N. Orr, M. Parton, I. E. Smith, J. Reis-Filho *et al.*, “A marker of homologous recombination predicts pathological complete response to neoadjuvant chemotherapy in primary breast cancer,” *Clinical Cancer Research*, pp. clincanres–1027, 2010.

[40] T. Jiang, W. Shi, V. B. Wali, L. S. Pongor, C. Li, R. Lau, B. Győrffy, R. P. Lifton, W. F. Symmans, L. Pusztai *et al.*, “Predictors of chemosensitivity in triple negative breast cancer: an integrated genomic analysis,” *PLoS medicine*, vol. 13, no. 12, p. e1002193, 2016.

[41] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, p. 415, 2013.

[42] T. Funnell, A. Zhang, Y.-J. Shiah, D. Grewal, R. Lesurf, S. McKinney, A. Bashashati, Y. K. Wang, P. Boutros, and S. Shah, “Integrated single-nucleotide and structural variation signatures of dna-repair deficient human cancers,” *bioRxiv*, p. 267500, 2018.

[43] T. Helleday, S. Eshtad, and S. Nik-Zainal, “Mechanisms underlying mutational signatures in human cancers,” *Nature Reviews Genetics*, vol. 15, no. 9, p. 585, 2014.

[44] S. A. Roberts, M. S. Lawrence, L. J. Klimczak, S. A. Grimm, D. Fargo, P. Stojanov, A. Kiezun, G. V. Kryukov, S. L. Carter, G. Saksena *et al.*, “An apobec cytidine deaminase mutagenesis pattern is widespread in human cancers,” *Nature genetics*, vol. 45, no. 9, p. 970, 2013.

[45] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, “Deciphering signatures of mutational processes operative in human cancer,” *Cell reports*, vol. 3, no. 1, pp. 246–259, 2013.

[46] D. Ramazzotti, A. Lal, K. Liu, R. Tibshirani, and A. Sidow, “De novo mutational signature discovery in tumor genomes using sparsesignatures,” *bioRxiv*, 2018.

[47] S. Nik-Zainal, H. Davies, J. Staaf, M. Ramakrishna, D. Glodzik, X. Zou, I. Martincorena, L. B. Alexandrov, S. Martin, D. C. Wedge *et al.*, "Landscape of somatic mutations in 560 breast cancer whole-genome sequences," *Nature*, vol. 534, no. 7605, p. 47, 2016.

[48] H. A. Wahba and H. A. El-Hadaad, "Current approaches in treatment of triple-negative breast cancer," *Cancer biology & medicine*, vol. 12, no. 2, p. 106, 2015.

[49] C. Crutcher, L. Cornwell, and A. Chagpar, "Effect of triple-negative status on surgical decision making." ASCO, 2010.

[50] E. B. C. T. C. Group *et al.*, "Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials," *The Lancet*, vol. 366, no. 9503, pp. 2087–2106, 2005.

[51] G. M. Freedman, P. R. Anderson, T. Li, and N. Nicolaou, "Locoregional recurrence of triple-negative breast cancer after breast-conserving surgery and radiation," *Cancer*, vol. 115, no. 5, pp. 946–951, 2009.

[52] J. Panoff, J. Hurley, C. Takita, I. Reis, W. Zhao, V. Sujoy, C. Gomez, M. Jorda, L. Koniaris, and J. Wright, "Risk of locoregional recurrence by receptor status in breast cancer patients receiving modern systemic therapy and post-mastectomy radiation," *Breast cancer research and treatment*, vol. 128, no. 3, pp. 899–906, 2011.

[53] B. S. Abdulkarim, J. Cuartero, J. Hanson, J. Deschênes, D. Lesniak, and S. Sabri, "Increased risk of locoregional recurrence for women with T1-2N0 triple-negative breast cancer treated with modified radical mastectomy without adjuvant radiation therapy compared with breast-conserving therapy," *Journal of Clinical Oncology*, vol. 29, no. 21, p. 2852, 2011.

[54] C. K. Anders and L. A. Carey, "Biology, metastatic patterns, and treatment of patients with triple-negative breast cancer," *Clinical breast cancer*, vol. 9, pp. S73–S81, 2009.

[55] T. Ballinger, J. Kremer, and K. Miller, "Triple negative breast cancer-review of current and emerging therapeutic strategies," 2016.

[56] R. Thirumaran, G. C. Prendergast, and P. B. Gilman, "Cytotoxic chemotherapy in clinical treatment of cancer," in *Cancer Immunotherapy.* Elsevier, 2007, pp. 101–116.

[57] P. C. Fong, T. A. Yap, D. S. Boss, C. P. Carden, M. Mergui-Roelvink, C. Gourley, J. De Greve, J. Lubinski, S. Shanley, C. Messiou *et al.*, "Poly (adp)-ribose polymerase inhibition: frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval," *Journal of clinical oncology*, vol. 28, no. 15, pp. 2512–2519, 2010.

[58] C. K. Anders, E. P. Winer, J. M. Ford, R. Dent, D. P. Silver, G. W. Sledge, and L. A. Carey, "Poly (adp-ribose) polymerase inhibition:"targeted" therapy for triple-negative breast cancer," *Clinical Cancer Research*, vol. 16, no. 19, pp. 4702–4710, 2010.

[59] J. S. Lim and D. S. Tan, "Understanding resistance mechanisms and expanding the therapeutic utility of parp inhibitors," *Cancers*, vol. 9, no. 8, p. 109, 2017.

[60] L. A. Carey, "Directed therapy of subtypes of triple-negative breast cancer," *The oncologist*, vol. 16, no. Supplement 1, pp. 71–78, 2011.

[61] N. T. Ueno and D. Zhang, "Targeting egfr in triple negative breast cancer," *Journal of Cancer*, vol. 2, p. 324, 2011.

[62] F. Sobande, L. Dusek, A. Matejková, T. Rozkos, J. Laco, and A. Ryska, "Egfr in triple negative breast carcinoma: significance of protein expression and high gene copy number," *Cesk Patol*, vol. 51, no. 2, pp. 80–86, 2015.

[63] A. Bahnassy, M. Mohanad, M. F. Ismail, S. Shaarawy, A. El-Bastawisy, and A.-R. N. Zekri, "Molecular biomarkers for prediction of response to treatment and survival in triple negative breast cancer patients from egypt," *Experimental and molecular pathology*, vol. 99, no. 2, pp. 303–311, 2015.

[64] A. Aleshin and R. S. Finn, "Src: a century of science brought to the clinic," *Neoplasia*, vol. 12, no. 8, pp. 599–607, 2010.

[65] M. Anbalagan, K. Moroz, A. Ali, L. Carrier, S. Glodowski, and B. G. Rowan, "Subcellular localization of total and activated src kinase in african american and caucasian breast cancer," *PloS one*, vol. 7, no. 3, p. e33017, 2012.

[66] E. M. Kim, K. Mueller, E. Gartner, and J. Boerner, "Dasatinib is synergistic with cetuximab and cisplatin in triple-negative breast cancer cells," *journal of surgical research*, vol. 185, no. 1, pp. 231–239, 2013.

[67] F. E. Vera-Badillo, A. J. Templeton, P. de Gouveia, I. Diaz-Padilla, P. L. Bedard, M. Al-Mubarak, B. Seruga, I. F. Tannock, A. Ocana, and E. Amir, “Androgen receptor expression and outcomes in early breast cancer: a systematic review and meta-analysis,” *Journal of the National Cancer Institute*, vol. 106, no. 1, p. djt319, 2013.

[68] E. Hilborn, J. Gacic, T. Fornander, B. Nordenskjöld, O. Stål, and A. Jansson, “Androgen receptor expression predicts beneficial tamoxifen response in oestrogen receptor-α-negative breast cancer,” *British journal of cancer*, vol. 114, no. 3, p. 248, 2016.

[69] N. Tung, J. E. Garber, M. R. Hacker, V. Torous, G. J. Freeman, E. Poles, S. Rodig, B. Alexander, L. Lee, L. C. Collins *et al.*, “Prevalence and predictors of androgen receptor and programmed death-ligand 1 in brca1-associated and sporadic triple-negative breast cancer,” *NPJ Breast Cancer*, vol. 2, p. 16002, 2016.

[70] R. M. Layman, A. S. Ruppert, M. Lynn, E. Mrozek, B. Ramaswamy, M. B. Lustberg, R. Wesolowski, S. Ottman, S. Carothers, A. Bingman *et al.*, “Severe and prolonged lymphopenia observed in patients treated with bendamustine and erlotinib for metastatic triple negative breast cancer,” *Cancer chemotherapy and pharmacology*, vol. 71, no. 5, pp. 1183–1190, 2013.

[71] M. Schuler, M. Uttenreuther-Fischer, M. Piccart-Gebhart, N. Harbeck, study group, and trial team, “Bibw 2992, a novel irreversible egfr/her1 and her2 tyrosine kinase inhibitor, for the treatment of patients with her2-negative metastatic breast cancer after failure of no more than two prior chemotherapies.” *Journal of Clinical Oncology*, vol. 28, no. 15_suppl, pp. 1065–1065, 2010.

[72] K. Gelmon, R. Dent, J. Mackey, K. Laing, D. McLeod, and S. Verma, “Targeting triple-negative breast cancer: optimising therapeutic outcomes,” *Annals of oncology*, vol. 23, no. 9, pp. 2223–2234, 2012.

[73] Y. K. Wang, A. Bashashati, M. S. Anglesio, D. R. Cochrane, D. S. Grewal, G. Ha, A. McPherson, H. M. Horlings, J. Senz, L. M. Prentice *et al.*, “Genomic consequences of aberrant dna repair mechanisms stratify ovarian cancer histotypes,” *Nature genetics*, vol. 49, no. 6, p. 856, 2017.

[74] A. Schuh, H. Dreau, S. J. Knight, K. Ridout, T. Mizani, D. Vavoulis, R. Colling, P. Antoniou,

E. M. Kvikstad, M. M. Pentony *et al.*, “Clinically actionable mutation profiles in patients with cancer identified by whole-genome sequencing,” *Molecular Case Studies*, vol. 4, no. 2, p. a002279, 2018.

[75] P. H. Sudmant, T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. H.-Y. Fritz *et al.*, “An integrated map of structural variation in 2,504 human genomes,” *Nature*, vol. 526, no. 7571, p. 75, 2015.

[76] R. Shepherd, S. A. Forbes, D. Beare, S. Bamford, C. G. Cole, S. Ward, N. Bindal, P. Gunasekaran, M. Jia, C. Y. Kok *et al.*, “Data mining using the catalogue of somatic mutations in cancer biomart,” *Database*, vol. 2011, 2011.

[77] E. R. Mardis, “The 1, 000*genome, the*100,000 analysis?” *Genome medicine*, vol. 2, no. 11, p. 84, 2010.

[78] E. F. Codd, “A relational model of data for large shared data banks,” *Communications of the ACM*, vol. 13, no. 6, pp. 377–387, 1970.

[79] T. J. Lee, Y. Pouliot, V. Wagner, P. Gupta, D. W. Stringer-Calvert, J. D. Tenenbaum, and P. D. Karp, “Biowarehouse: a bioinformatics database warehouse toolkit,” *BMC bioinformatics*, vol. 7, no. 1, p. 170, 2006.

[80] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson *et al.*, “Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal,” *Sci. Signal.*, vol. 6, no. 269, pp. pl1–pl1, 2013.

[81] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, and S. W. Scherer, “The database of genomic variants: a curated collection of structural variation in the human genome,” *Nucleic acids research*, vol. 42, no. D1, pp. D986–D992, 2013.

[82] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks *et al.*, “Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data,” *Genome research*, pp. gr–180 281, 2014.

[83] A. McPherson, S. P. Shah, and S. C. Sahinalp, “destruct: Accurate rearrangement detection using breakpoint specific realignment,” *bioRxiv*, p. 117523, 2017.

[84] R. M. Layer, C. Chiang, A. R. Quinlan, and I. M. Hall, “Lumpy: a probabilistic framework for structural variant discovery,” *Genome biology*, vol. 15, no. 6, p. R84, 2014.

[85] J. Ding, A. Bashashati, A. Roth, A. Oloumi, K. Tse, T. Zeng, G. Haffari, M. Hirst, M. A. Marra, A. Condon *et al.*, “Feature-based classifiers for somatic mutation detection in tumour–normal paired sequencing data,” *Bioinformatics*, vol. 28, no. 2, pp. 167–175, 2011.

[86] C. T. Saunders, W. S. Wong, S. Swamy, J. Becq, L. J. Murray, and R. K. Cheetham, “Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs,” *Bioinformatics*, vol. 28, no. 14, pp. 1811–1817, 2012.

[87] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts *et al.*, “Mutational heterogeneity in cancer and the search for new cancer-associated genes,” *Nature*, vol. 499, no. 7457, p. 214, 2013.

[88] A. Taylor-Weiner, C. Stewart, T. Giordano, M. Miller, M. Rosenberg, A. Macbeth, N. Lennon, E. Rheinbay, D.-A. Landau, C. J. Wu *et al.*, “Detin: overcoming tumor-in-normal contamination,” *Nature Methods*, vol. 15, no. 7, p. 531, 2018.

[89] M. J. Taghiyar, J. Rosner, D. Grewal, B. M. Grande, R. Aniba, J. Grewal, P. C. Boutros, R. D. Morin, A. Bashashati, and S. P. Shah, “Kronos: a workflow assembler for genome analytics and informatics,” *GigaScience*, vol. 6, no. 7, p. gix042, 2017.

Appendix A

Examples of Data Structuring, Bulk-loading and Data Extraction Scripts

A.1 Examples of data structuring and loading scripts

A.1.1 Script to structure and load bam file statics derived by flagstats

```
#!/home/rasiimwe/miniconda3/bin/python
import psycopg2
import sys
import csv
import os
import string
import subprocess

con = None
try:
        con = psycopg2.connect("host=‘localhost’ dbname=‘genomic_variants’
        user=‘ ’ password=‘ ’")

        cur = con.cursor()

        #creating tumour_bamstats database object
        create table tumour_bamstats (tumour_id varchar primary key references
        samples (tumour_id), normal_id varchar, total_reads bigint, qc_failure
        bigint, duplicates bigint, mapped  bigint, mapped_percentage float,
        paired_in_seq bigint, read1 bigint, read2 bigint, properly_paired
        bigint, properly_paired_percentage float,self_and_mate_mapped bigint,
        singletons bigint, singletons_percentage float, mate_map_diff_chr
        bigint, mate_map_diff_chr_mapq bigint, mapq  varchar, mapq2 varchar,
        avg_read_coverage float)

        #Calling extracted stats, structuring and loading into database
#------------------------------------------------------------------------------
        path = ""
        os.chdir(path)
        for file in os.listdir(path):
                tumour_id = ‘_illumina’.join(file.split(‘_illumina’)[:-1])
                f = open(file, "r")
                for i, line in enumerate(f):
                        if i == 0: #in total e.g. 843990938
                                total = int(filter(str.isdigit, line))

                        elif i == 1: #QC failure e.g. 0
                                qc.f = int(filter(str.isdigit, line))

                        elif i == 2: #duplicates e.g. 623568175
                                duplicates = int(filter(str.isdigit, line))

                        elif i == 3: #mapped e.g. 801225533, percentage
                        #mapped eg  (94.93%)  # pass mapped and percentage
                        #mapped as 2 diff variales
                                mapped = ‘ ’.join(line.split(‘ ’)[:1])
                                da = ‘ ’.join(line.split(‘ ’)[-1:])
                                db = str(da)[1:-1]
                                dc = ‘)’.join(db.split(‘)’)[:-1])
                                mapped_percentage= ‘%’.join(dc.split(‘%’)[:-1])

                        elif i==4: #paired in sequencing e.g. 843990938
                                paired_in_seq = int(filter(str.isdigit, line))

                        elif i == 5: #read1 e.g. 421995469
                                read1 = ‘ ’.join(line.split(‘ ’)[:1])

                        elif i == 6: #read2 e.g. 421995469
                                read12 = ‘ ’.join(line.split(‘ ’)[:1])

                        elif i == 7: #properly paired e.g. 791170058,
                        #percentage of properly paired (93.74%)
                                properly_paired = ‘ ’.join(line.split(‘ ’)[:1])
                                ha = ‘ ’.join(line.split(‘ ’)[-1:])
                                hb = str(ha)[1:-1]
                                hc = ‘)’.join(hb.split(‘)’)[:-1])
                                properly_paired_percentage = ‘%’.join(hc.split
                                (‘%’)[:-1])

                        elif i == 8: #with itself and mate mapped
                        #e.g. 795636451
                                self_and_mate_mapped = int(filter(str.isdigit,
                                line))

                        elif i == 9: #singletons e.g. 42764211 percentage
                        #of singletons (5.07%)
                        singletons = ‘ ’.join(line.split(‘ ’)[:1])
                                ja = ‘ ’.join(line.split(‘ ’)[-1:])
                                jb = str(ja)[1:-1]
```

```
                                jc = ‘)’.join(jb.split(‘)’)[:-1])
                                singletons_percentage =
                                ’%’.join(jc.split(‘%’)[:-1])

                        elif i == 10: #with mate mapped to a different chr
                        #e.g.  2882215
                        mate.mapped.chr= int(filter(str.isdigit, line))

                        elif i == 11: #with mate mapped to a different chr
                        #e.g. 2178551, mapQ (mapQ>=5)
                                mate.mapped.chr.diff =
                                ‘ ’.join(line.split(‘ ’)[:1])
                                la = ‘ ’.join(line.split(‘ ’)[-1:])
                                lb = str(la)[1:-1]
                                mapq = ‘)’.join(lb.split(‘)’)[:-1])
                                mapq2 = ‘Q’.join(l2.split(‘Q’)[-1:])

                cur.execute ("""
                        update tumour_bamstats set
                        total_reads=(%s), qc_failure=(%s),duplicates=(%s),
                        mapped=(%s), mapped_percentage=(%s),
                        paired_in_seq=(%s), read1=(%s), read2=(%s),
                        properly_paired=(%s), properly_paired_percentage=(%s),
                        self_and_mate_mapped=(%s),
                        singletons=(%s), singletons_percentage=(%s),
                        mate_map_diff_chr=(%s), mate_map_diff_chr_mapq=(%s),
                        mapq=(%s), mapq2=(%s)
                        where tumour_id = (select tumour_id from samples where
                        tumour_archive_id =(%s) and
                        samples.tumour_id = tumour_bamstats.tumour_id)
                        """, [total, qc.f, duplicates, mapped,
                        mapped_percentage, paired_in_seq, read1,
                        read2, properly_paired,properly_paired_percentage,
                        self_and_mate_mapped, singletons,
                        singletons_percentage, mate.mapped.chr,
                        mate.mapped.chr.diff, mapq, mapq2, tumour_id])

#-----------------------------------------------------------------------------
        con.commit()
except psycopg2.DatabaseError, e:
        if con:
                con.rollback()
        print ‘Eror %s’ % e
        sys.exit(1)
```

A.1.2 Script to structure and load mutationSeq data

```
library(VariantAnnotation)
library(dplyr)
library(tidyr)
library(splitstackshape)
library(RPostgreSQL)

pw <- { " "}
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "genomic_variants", host = "localhost",
                user = "rasiimwe", password = pw)
rm(pw)

dbExistsTable(con, "pipeline_result_paths")
museqsnvs <- dbGetQuery(con, "select tumour_id, mutationseq from
                        pipeline_result_paths")
museqsnvs <- as.data.frame(museqsnvs)
tumour_id1 = museqsnvs[1]
museqsnvs.path = museqsnvs[2]

for(i in museqsnvs.path){
        files <- Sys.glob(file.path(i, "*.vcf"))
        for (f in files){
                x <- matrix(unlist(strsplit(as.character(f), ‘/’)), ncol=1,
                byrow=TRUE)
                tumour_id <- as.character(x[5])

                vcf <- readVcf(f, "hg19")
                if (dim(vcf)[1]! = 0){
                        initial <- data.frame(info(vcf))
                        initial <- tibble::rownames_to_column(initial,
                        "chrom_pos_ref_alt")


                        split1 <- matrix(unlist(strsplit(as.character
                        (initial$chrom_pos_ref_alt), ‘:’)), ncol=2,byrow=TRUE)
                        df <- cbind(initial$chrom_pos_ref_alt,
                        as.data.frame(split1))
                        names(df) <- c("chrom_pos_ref_alt", "chrom", "pos")
                        split2 <- matrix(unlist(strsplit(as.character
                        (df$pos), ‘_’)), ncol = 2, byrow=TRUE)
                        df2 <- cbind(df, split2)
                        names(df2) <- c("chrom_pos_ref_alt", "chrom",
                        "pos1", "pos", "ref_alt")
                        split3 <- matrix(unlist(strsplit(as.character
                        (df2$ref_alt), ‘/’)), ncol = 2, byrow = TRUE)
                        df3 <- cbind(df2, split3)
                        names(df3) <- c("chrom_pos_ref_alt", "chrom",
                        "pos1", "pos", "ref_alt", "ref", "alt")


                        initial <- cbind (df3$chrom, df3$pos,
                        df3$ref,df3$alt,initial)
                        names(initial)[names(initial)==‘df3$chrom’]<-‘chrom’
                        names(initial)[names(initial) == ‘df3$pos’]<- ‘pos’
                        names(initial)[names(initial) == ‘df3$ref’]<- ‘ref’
                        names(initial)[names(initial) == ‘df3$alt’]<- ‘alt’
                        initial$chrom_pos_ref_alt <- NULL


                        newann <- cSplit(initial, 13, sep = ",",
                        direction = "long", fixed = FALSE, drop = TRUE,
                        stripWhite = TRUE, makeEqual =FALSE,
                        type.convert = TRUE)
                        newann <- as.data.frame(newann)
```

```
                        newann[] <- lapply(newann, gsub, pattern=‘"’,
                        replacement=‘’)
                        newann <- cSplit(newann, "ANN", "|")
                        names(newann)[names(newann) %in% ...
                        newann[] <- lapply(newann, gsub,
                        pattern = ‘\\(’, replacement=‘’)
                        newann[] <- lapply(newann, gsub, pattern=‘)’,
                        replacement = ‘’)
                        newann$allele[] <- lapply(newann$allele, gsub,
                        pattern = ‘c’, replacement=‘’)

                        newann$tumour_id <- " "
                        newann$tumour_id <- tumour_id
                        newann <- setNames(newann, tolower(colnames(newann)))
                        museq_unfiltered <- newann[,c(35,1:12,19:34,13:18)]
                        #museq_unfiltered <- newann[,c(1:12,19:34,13:18)]
                        museq_unfiltered<-cbind("id"=1:nrow(museq_unfiltered),
                        museq_unfiltered)

                        dbWriteTable(con,"museq_unfiltered", museq_unfiltered,
                        append=TRUE, row.names=0)
                }
        }

}

dbDisconnect(con)
```

A.2 Examples of data extraction scripts

A.2.0.1 Extracting mutation loads per case:

```
dbExistsTable(con, "snvs_intersect", "svs_filtered", "strelka_indels")
#snvs
mutation.load <- dbGetQuery(con,
                       "SELECT load1.tumour_id, snvs_load, svs_load,
                        indel_load
                        FROM (SELECT DISTINCT tumour_id, COUNT(DISTINCT(
                        tumour_id, chrom, pos)) AS snvs_load
                        FROM snvs_intersect
                        WHERE pr >= 0.9  GROUP BY 1 ORDER BY 2)load1
                        LEFT OUTER JOIN LATERAL
                        (SELECT COUNT(tumour_id) AS svs_load
                        FROM svs_filtered
                        WHERE load1.tumour_id = svs_filtered.tumour_id
                        GROUP BY tumour_id )load2
                        ON TRUE LEFT OUTER JOIN LATERAL
                        (SELECT COUNT(DISTINCT(tumour_id, chrom, pos))
                        AS indel_load
                        FROM strelka_indels
                        WHERE load1.tumour_id = strelka_indels.tumour_id)load3
                        ON TRUE")

#sample output
------------------------------------------------------------------------------
 tumour_id | snvs_load | svs_load | indel_load
-----------+-----------+----------+-------------
 SA1064    |       129 |        1 |          18
 SA1058    |       269 |       16 |          23
 SA296     |       450 |       28 |           2
```

```
SA576     |       645 |       37 |           5
SA1027    |      1376 |       32 |          78
SA680     |      1954 |       51 |          32
SA402     |      2127 |       43 |         121
SA230     |      2193 |       75 |         239
SA1071    |      2536 |       61 |          29
SA601     |      2640 |       20 |          45
SA596     |      2672 |      102 |          71
SA275     |      2746 |       44 |          32
SA423     |      3040 |      136 |         161
SA1062    |      3250 |       89 |          42
SA286     |      3293 |       43 |          77
```

A.2.0.2 Extracting samples with specified mutations in genes of interest:

```
gene.effect <- function(gene, annotation){
  if (!is.character(gene) | !is.character(annotation)) {
    stop(paste("Expecting gene or annotation to be of type character.
             You supplied", typeof(gene), "for gene and ",
               typeof(annotation),"for annotation"))
  }
  else
    if(is_empty(gene) | is_empty(annotation)){
      stop(paste("Expecting gene or annotation elements.
                 Your vector is empty"))
    }
  else
  {
    genes <- as.vector(gene)
    for(i in genes){
      annotation <- paste(annotation, "%", sep="")
      annotation <- paste( "%", annotation, sep="")
      annotations <- as.vector(annotation)

      for(j in annotations){
        #magic_for(print, silent = TRUE)
        gene <- i
        effect <- j

        query <- fn$identity("select distinct tumour_id, gene_name, annotation
        from strelka_indels where gene_name like ‘$gene’ and annotation like
        ‘$effect’ union select distinct tumour_id, gene_name, annotation from
        snvs_intersect where gene_name like ‘$gene’ and annotation like
        ‘$effect’ and pr >= 0.90 order by gene_name asc")

        x <- dbGetQuery(con, query)
        x <- as.data.frame(x)

        if (nrow(x)! = 0)
        {
          x <- as.data.frame(x)
          x <- na.omit(x)
          print(x)
        }
        else{
          next
        }
      }
    }
  }
}

output.gene.effect <- as.data.frame(capture.output(gene.effect(gene, annotation)))
```

A.2.0.3 Extracting copy number profile of a case of interest:

```
pw <- {
  " "
}
drv <- dbDriver("PostgreSQL")
con <- dbConnect(drv, dbname = "genomic_variants",
                 host = "localhost", port = 5433,
                 user = "rasiimwe", password = pw)
#on.exit(dbDisconnect(con))

sample <- ‘SA994’
query <- fn$identity("select chromosome, start_position_bp as start_pos,
                    end_position_bp as end_pos, titan_call, copy_number
                    from titan_segs_cnas where
                    tumour_id = ‘$sample’ order by 1, 5 asc")
cna.profile <- dbGetQuery(con, query)

#sample output
----------------------------------------------------------------------------
 chromosome | start_pos |  end_pos  | titan_call | copy_number
------------+-----------+-----------+------------+-------------
 1          |    774883 |    811735 | HOMD       |           0
 1          |   1809509 |   2831790 | NLOH       |           2
 1          |    814077 |   1514183 | NLOH       |           2
 1          |  17019673 |  17266878 | HET        |           2
 1          |   5812704 |   8640892 | NLOH       |           2
 1          |   3833193 |   5798176 | NLOH       |           2
 1          |   8695522 |  17005876 | NLOH       |           2
 1          |  17275895 |  19582338 | NLOH       |           2
 1          |  19587201 |  19806850 | NLOH       |           2
 1          |  37365498 |  39169657 | NLOH       |           2
```

```
1          | 232230516 | 249208389 | NLOH       |           2
1          |   2835487 |   3089976 | NLOH       |           2
1          |  20376717 |  36454013 | NLOH       |           2
1          | 120919943 | 146528957 | BCNA       |           4
1          | 231220183 | 232224588 | BCNA       |           4
1          | 226334605 | 231187205 | BCNA       |           4
1          | 225887453 | 226292868 | BCNA       |           4
1          | 197254186 | 225611118 | BCNA       |           4
1          | 192443500 | 197201223 | BCNA       |           4
1          |  36496696 |  37362015 | ALOH       |           4
```

Appendix B

Significantly Mutated Genes

B.1 50 top significantly mutated genes (SMGs) in this TNBC study cohort identified using MutsigCV

```
gene    expr reptime hic N_nonsilent N_silent N_noncoding n_nonsilent n_silent n_noncoding
nnei x X p q
EMCN 231787 807 -5 127075 34775 421590 6 0 2 50 10 19776510 0 0
TP53 2069567 213 34 200785 55445 348465 37 0 1 50 6 18124470 0 0
MUC21 2483101 261 36 248820 84435 111735 7 1 0 47 4 16643185 2.332409e-06 1.466463e-02
PIK3CA 401889 613 11 513110 127855 660270 9 0 0 26 5 12899575 1.166765e-05 5.501878e-02
MUC4 920866 365 49 551915 154570 713895 23 4 0 4 4 1561300 6.284072e-05 1.812296e-01
MB 936853 205 -13 74165 19240 108225 2 0 0 50 11 17708145 7.135792e-05 1.812296e-01
CTU2 802673 189 41 242840 70720 398190 3 0 0 50 12 20439705 7.879780e-05 1.812296e-01
RAB3IL1 1998419 212 69 177970 53105 165750 3 0 0 50 5 24505325 8.475045e-05 1.812296e-01
PTEN 259678 300 34 196820 45760 336765 3 0 0 50 10 20711470 8.647367e-05 1.812296e-01
CLEC9A 351158 445 9 119535 26715 384150 2 0 0 50 11 26085865 1.262553e-04 2.160980e-01
NOL10 804107 344 32 337220 81445 295815 3 0 0 50 11 20631910 1.449161e-04 2.160980e-01
LAMB4 487451 326 4 838240 218270 1281150 5 1 0 37 4 18093855 1.474913e-04 2.160980e-01
PLP1 472103 NaN 26 129610 37700 355875 2 0 0 24 2 13731445 1.489383e-04 2.160980e-01
SPATA4 134663 504 -4 144040 38870 182910 3 0 0 50 14 23871315 1.707353e-04 2.300293e-01
PRB3 311837 673 -1 137020 47840 184275 3 0 0 50 7 15785380 2.310394e-04 2.856918e-01
PDCD6IP 293448 409 9 411580 112970 586755 5 0 1 38 10 17206410 2.423427e-04 2.856918e-01
PIK3R1 48999 619 32 368940 91455 665730 6 1 0 50 12 19350110 2.707102e-04 2.901350e-01
MIDN 2214581 234 29 208325 70720 121875 2 0 0 50 2 15072785 2.768757e-04 2.901350e-01
SERPINB3 174293 370 -5 187460 46735 297960 2 0 0 50 9 23473580 3.351789e-04 3.311552e-01
SYT8 916434 481 33 181610 60190 156390 2 0 0 50 8 20130760 3.559895e-04 3.311552e-01
TFCP2L1 240173 568 9 231140 60970 481065 2 0 0 50 8 24944725 3.686915e-04 3.311552e-01
RUNX1 164915 429 45 226135 70265 319800 3 0 0 50 9 22489350 4.203889e-04 3.481186e-01
ANO2 349024 712 13 478335 125970 716430 5 0 0 11 1 5744960 4.279815e-04 3.481186e-01
TMEM41A 953031 366 31 121290 37245 147030 2 0 0 50 5 21343595 4.568487e-04 3.481186e-01
EZR 564251 163 55 283010 70525 564525 2 0 0 50 5 18855655 4.660136e-04 3.481186e-01
```

```
ZCCHC5 263052 NaN 20 216775 61685 11895 3 0 0 50 8 18993455 4.798581e-04 3.481186e-01

CREB3L1 801520 158 36 242255 70525 130845 2 0 0 50 5 19788925 5.353796e-04 3.740122e-01

DNAJC17 1011507 172 46 148395 38610 636285 2 0 0 15 2 7109895 5.709995e-04 3.846497e-01

C6orf89 836025 244 43 169130 44005 263250 2 0 0 50 4 20571005 7.313133e-04 4.680890e-01

OR4X1 400891 509 -27 137735 41080 20280 4 0 0 50 11 18862740 7.533784e-04 4.680890e-01
```

Appendix C

Database Data Dictionary

The data dictionary presented herein shows the database objects (entities) for which data was collected and their respective descriptions. Described are the fields (variables), their description, data type and constraints for each corresponding object.

Entity 1: Projects - Towards database expansion and analysis of data from different studies, the projects table contains data about the different projects to which each collected sample data belongs. All current samples in the developed database belong to the TNBC project.

Entity: Projects			
Field Name	**Description**	**Data Type**	**Constraints**
project_code	Project code uniquely identifies studies in the database for which variant data from WGS has been collected	CHARACTER_VARYING	PRIMARY_KEY
name	Project name	CHARACTER_VARYING	NOT NULL

Table C.1: Database entity - Projects

Entity 2: Clinical_data - This table contains available clinical outcomes data collected on the samples in this TNBC cohort. Clinical data was collected from various institutions (BC outcomes unit, Montreal, Alberta and McGill) from which samples were collected.

Entity: Clinical_data			
Field Name	**Description**	**Data Type**	**Constraints**
consent_id	Clinical id that uniquely identifies patients across systems	CHARACTER_VARYING	PRIMARY_KEY

diagnosis_date	Date of cancer diagnosis	DATE	NULL
age	Patient's age at diagnosis	INTEGER	NULL
grade	This field captures the patient's tumour grade and takes on values such as 3, 2 or 1 for grades 3, 2 and 1 respectively. The tumour grade can also be unknown, therefore the field accommodates null values	INTEGER	NULL
tumour_size	Patient's tumour size in cm. Null values are accommodated as a patient's tumour size may be unknown	FLOAT	NULL
node_status	Node status (can either be positive, negative or unknown (NULL))	CHARACTER_VARYING	NULL
HER_status	A patient's HER2 status as identified by IHC (can either be positive, negative or unknown (NULL))	CHARACTER_VARYING	NULL
ER_status	A patient's ER2 status as identified by IHC (can either be positive, negative or unknown (NULL))	CHARACTER_VARYING	NULL
PR_status	A patient's PR status as identified by IHC (can either be positive, negative or unknown (NULL))	CHARACTER_VARYING	NULL
OS_status	A patient's overall survival status (can either be alive or dead or unknown (NULL))	CHARACTER_VARYING	NULL
OS_years	A patient's overall survival in years	FLOAT	NULL
DSS_status	A patient's disease specific survival status (can either be alive or dead or unknown (NULL))	CHARACTER_VARYING	NULL
DSS_years	A patient's disease specific survival in years	FLOAT	NULL

PFS_status	A patient's progression free survival status (can either be alive or dead or unknown (NULL))	CHARACTER_VARYING	NULL
PFS_years	A patient's progression free survival in years	FLOAT	NULL

Table C.2: Database entity - Clinical_data

Entity 3: Samples - This table contains metadata on each of the patient samples. Data collected includes the sample identifier, clinical identifier, facility of origin, sample type and the associated project to which a sample belongs.

Entity: Samples			
Field Name	**Description**	**Data Type**	**Constraints**
tumour_id	This field captures the unique tumour identifier	CHARACTER_VARYING	COMPOSITE PRIMARY_KEY (unique id pair (tumour_id & normal_id))
normal_id	Normal sample identifier (part of the composite primary key). This value is not unique as a patient with 2 tumour samples will have 1 normal sample associated with each matched tumour sample	CHARACTER_VARYING	NOT NULL
consent_id	Patient clinical identifier	CHARACTER_VARYING	FOREIGN_KEY references clinical_data (consent_id)
facility_of_origin	This field captures the facility/institution of origin of each sample	CHARACTER_VARYING	NOT NULL

sample_type	The sample type can either be xenograft or primary	CHARACTER_ VARYING	NOT NULL
project_code	The code (identifier) of the project to which a sample belongs	CHARACTER_ VARYING	FOREIGN_ KEY references projects (project_code)

Table C.3: Database entity - Samples

Entity 4: Titan_params_cnas - This table contains relevant parameters (normal_contamination_estimate and average_tumour_ploidy_estimate) on each sample inferred by TITAN.

Entity: Titan_params_cnas			
Field Name	**Description**	**Data Type**	**Constraints**
tumour_id	Unique identifier for a tumour sample	CHARACTER_ VARYING	PRIMARY_ KEY
normal_ contamination_ estimate	This field captures the normal contamination estimate derived from TITAN and denotes the proportion of normal content in a sample	FLOAT	NOT NULL
average_ tumour_ploidy_ estimate	The average number of estimated copies in the genome (2 represents diploid)	FLOAT	NOT NULL

Table C.4: Database entity - Titan_params_cnas

Entity 5: Titan_segs_cnas - This table contains copy number aberrations per genomic segment inferred by TITAN.

Entity: Titan_segs_cnas			
Field Name	**Description**	**Data Type**	**Constraints**

id	Serial id	INTEGER	PRIMARY_KEY(SEQ)
tumour_id	Tumour/sample id associated with the variant called	CHARACTER_VARYING	
chromosome	Chromosome with copy number aberration	CHARACTER_VARYING	NOT NULL
start_position_bp	Segment start position	BIGINT	NOT NULL
end_position_bp	Segment end position	BIGINT	NOT NULL
length_bp	Number of SNPs in the segment	INTEGER	NOT NULL
median_ratio	Median allelic ratio across SNPs in a segment	FLOAT	NOT NULL
median_logr	Median log ratio across SNPs in the segment	FLOAT	NOT NULL
titan_state	State number used by TITAN	INTEGER	NOT NULL
titan_call	Interpretable TITAN state (Can be HOMD, DLOH, HET, NLOH, ALOH, ASCNA, BCNA, UBCNA)	CHARACTER_VARYING	NOT NULL
copy_number	Predicted TITAN copy number	INTEGER	NOT NULL
minorcn	Copy number of minor allele	INTEGER	NOT NULL
majorcn	Copy number of major allele	INTEGER	NOT NULL
clonal_cluster	Predicted TITAN clonal cluster	INTEGER	NOT NULL
clonal_frequency	Clonal frequency	INTEGER	NOT NULL
gene_name	Mutated gene(s) in segment	CHARACTER_VARYING	NOT NULL

Table C.5: Database entity - Titan_segs_cnas

Entity 6: Titan_outfile_cnas - This table contains CNAs derived from Titan.

Entity: Titan_outfile_cnas			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_ VARYING	NOT NULL
tumour_id	Tumour/sample id associated with the variant called	CHARACTER_ VARYING	
chr	Chromosome	CHARACTER_ VARYING	NOT NULL
position	Position	BIGINT	
refcount	Number of reads matching the reference base	INTEGER	NOT NULL
nrefcount	Number of reads matching the non-reference base	INTEGER	NOT NULL
depth	Total read depth at a position	INTEGER	NOT NULL
allelicratio	Refcount/depth	FLOAT	NOT NULL
logratio	Log2 ratio between normalized tumour and normal read depths	FLOAT	NOT NULL
copynumber	Predicted TITAN copy number	INTEGER	NOT NULL
titanstate	Internal state number used by TITAN	INTEGER	NOT NULL
titancall	Interpretable TITAN state (Can be HOMD, DLOH, HET, NLOH, ALOH, ASCNA, BCNA, UBCNA)	CHARACTER_ VARYING	NOT NULL
clonalcluster	Predicted TITAN clonal cluster	INTEGER	NOT NULL
cellular prevalence	Proportion of tumour cells containing genomic event	FLOAT	NOT NULL

Table C.6: Database entity - Titan_outfile_cnas

Entity 7: Museq_snvs - This table contains somatic single nucleotide variants (SNV) detected at each genomic position using mutationSeq for each sample in the cohort.

** The table snvs_intersect contains a mapping of snvs detected by mutationSeq with those detected by Strelka. The table therefore inherits fields and descriptions as those of the museq_snvs table.

Entity: Museq_snvs			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_ VARYING	NOT NULL
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_ VARYING	NOT NULL
chrom	Chromosome identifier from the reference genome	CHARACTER_ VARYING	NOT NULL
pos	Reference position	BIGINT	NOT NULL
ref	Reference nucleotide at position (pos) of the chromosome	CHARACTER_ VARYING	NOT NULL
alt	Alternate non-reference allele	CHARACTER_ VARYING	NOT NULL
pr	Probability of somatic mutation	FLOAT	NOT NULL
tc	Tri-nucleotide context	CHARACTER_ VARYING	NOT NULL
tr	Count of tumour with reference to REF	INTEGER	NOT NULL
ta	Count of tumour with reference to ALT	INTEGER	NOT NULL
nr	Count of normal with reference to REF	INTEGER	NOT NULL
na	Count of normal with reference to ALT	INTEGER	NOT NULL
nd	Number of deletions	INTEGER	NOT NULL
ni	Number of insertions	INTEGER	NOT NULL
allele	Identifies the alt being referred to in instances of multiple alt fields	CHARACTER_ VARYING	NOT NULL
annotation	Effect or consequence annotated using sequence ontology terms e.g splice_donor	CHARACTER_ VARYING	NOT NULL
annotation_ impact	Estimation of putative impact or deleteriousness (Can either be HIGH, MODERATE, LOW or MODIFIER)	CHARACTER_ VARYING	NOT NULL

gene_name	Common gene name (HGNC)	CHARACTER_VARYING	NOT NULL
gene_id	Gene ID	CHARACTER_VARYING	NOT NULL
feature_type	Transcript, motif, miRNA	CHARACTER_VARYING	NOT NULL
feature_id	Depending on the annotation, this may be: Transcript ID Motif ID, miRNA, ChipSeq peak, Histone mark, etc	CHARACTER_VARYING	NOT NULL
transcript_biotype	Description on whether the transcript is coding or noncoding	CHARACTER_VARYING	NOT NULL
rank	Exon or intron rank or total number of exons or introns	CHARACTER_VARYING	NOT NULL
hgvs_c	Variant using HGVS notation (DNA level)	CHARACTER_VARYING	NOT NULL
hgvs_p	If variant is coding, this field describes the variant using HGVS notation (protein level)	CHARACTER_VARYING	NOT NULL
cdna_pos_cdna_length	Position in cDNA and trancript's cDNA length	INTEGER	NOT NULL
cds_pos_cds_length	Position and number of coding bases	INTEGER	NOT NULL
aa_pos_aa_length	Position and number of AA	INTEGER	NOT NULL

distance	Context and implementation dependent. E.g. when the variant is “intronic” the annotation may show the distance to the closest exon; when the variant is “intergenic” it may show the distance to the closest gene; and when the variant is “upstream or downstream” may show the distance to the closest 5’UTR or 3’UTR base	INTEGER	NOT NULL
errors_warning_info	Warnings or information messages	CHARACTER_VARYING	NOT NULL

Table C.7: Database entity - Museq_snvs

Entity 8: Strelka_snvs - This table contains somatic single nucleotide variants (SNV) detected at each genomic position using Strelka for each sample in the cohort.

Entity: Strelka_snvs			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_VARYING	NOT NULL
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_VARYING	NOT NULL
chrom	Chromosome identifier from the reference genome	CHARACTER_VARYING	NOT NULL
pos	Reference position	BIGINT	NOT NULL
ref	Reference nucleotide at position (pos) of the chromosome	CHARACTER_VARYING	NOT NULL
alt	Alternate non-reference allele	CHARACTER_VARYING	NOT NULL
qss	SNV quality score	INTEGER	NOT NULL

tqss	Data tier used to compute QSS	INTEGER	NOT NULL
nt	Genotype of the normal in all data tiers, as used to classify somatic variants	CHARACTER_VARYING	NOT NULL
qss_nt	Quality score reflecting the joint probability of a somatic variant and NT	INTEGER	NOT NULL
tqss_nt	Data tier used to compute QSS_NT	INTEGER	NOT NULL
sgt	Most likely somatic genotype excluding normal noise states	CHARACTER_VARYING	NOT NULL
somatic	Denoting somatic mutation	CHARACTER_VARYING	NOT NULL
allele	Identifies the alt being referred to in instances of multiple alt fields	CHARACTER_VARYING	NOT NULL
annotation	Effect or consequence annotated using sequence ontology terms e.g splice_donor	CHARACTER_VARYING	NOT NULL
annotation_impact	Estimation of putative impact or deleteriousness (Can either be HIGH, MODERATE, LOW or MODIFIER)	CHARACTER_VARYING	NOT NULL
gene_name	Common gene name (HGNC)	CHARACTER_VARYING	NOT NULL
gene_id	Gene ID	CHARACTER_VARYING	NOT NULL
feature_type	Transcript, motif, miRNA	CHARACTER_VARYING	NOT NULL
feature_id	Depending on the annotation, this may be: Transcript ID Motif ID, miRNA, ChipSeq peak, Histone mark, etc	CHARACTER_VARYING	NOT NULL
transcript_biotype	Description on whether the transcript is coding or noncoding	CHARACTER_VARYING	NOT NULL
rank	Exon or intron rank or total number of exons or introns	CHARACTER_VARYING	NOT NULL

hgvs_c	Variant using HGVS notation (DNA level)	CHARACTER_VARYING	NOT NULL
hgvs_p	If variant is coding, this field describes the variant using HGVS notation (Protein level)	CHARACTER_VARYING	NOT NULL
cdna_pos_cdna_length	Position in cDNA and trancript's cDNA length	INTEGER	NOT NULL
cds_pos_cds_length	Position and number of coding bases	INTEGER	NOT NULL
aa_pos_aa_length	Position and number of AA	INTEGER	NOT NULL
distance	Context and implementation dependent. E.g. when the variant is "intronic" the annotation may show the distance to the closest exon; when the variant is "intergenic" it may show the distance to the closest gene; and when the variant is "upstream or downstream" may show the distance to the closest 5'UTR or 3'UTR base	INTEGER	NOT NULL
errors_warning_info	Warnings or information messages	CHARACTER_VARYING	NOT NULL

Table C.8: Database entity - Strelka_snvs

Entity 9: Strelka_indels - This table contains insertions and deletions detected at each genomic position using Strelka for each sample in the cohort.

Entity: Strelka_indles			
Field Name	**Description**	**Data Type**	**Constraints**

id	Serial id	CHARACTER_ VARYING	NOT NULL
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_ VARYING	NOT NULL
chrom	Chromosome identifier from the reference genome	CHARACTER_ VARYING	NOT NULL
pos	Reference position	BIGINT	NOT NULL
ref	Reference nucleotide at position (pos) of the chromosome	CHARACTER_ VARYING	NOT NULL
alt	Alternate non-reference allele	CHARACTER_ VARYING	NOT NULL
qsi	Quality score for variant	INTEGER	NOT NULL
tqsi	Data tier used to compute QSI	INTEGER	NOT NULL
nt	Genotype of the normal in all data tiers, as used to classify somatic variants	CHARACTER_ VARYING	NOT NULL
qsi_nt	Quality score reflecting the joint probability of a somatic variant and NT	INTEGER	NOT NULL
tqsi_nt	Data tier used to compute QSI_NT	INTEGER	NOT NULL
sgt	Most likely somatic genotype excluding normal noise states	CHARACTER_ VARYING	NOT NULL
ru	Smallest repeating sequence unit in inserted or deleted sequence	CHARACTER_ VARYING	NOT NULL
rc	Number of times RU repeats in the reference allele	INTEGER	NOT NULL
ic	Number of times RU repeats in the indel allele	INTEGER	NOT NULL
ihp	Largest reference interrupted homopolymer length intersecting with the indel	INTEGER	NOT NULL
svtype	Type of structural variant	CHARACTER_ VARYING	NOT NULL

somatic	Flag denoting somatic mutation	CHARACTER_VARYING	NOT NULL
overlap	Flag denoting somatic indel possibly overlaps a second indel	CHARACTER_VARYING	NOT NULL
allele	Identifies the alt being referred to in instances of multiple alt fields	CHARACTER_VARYING	NOT NULL
annotation	Effect or consequence annotated using sequence ontology terms e.g splice_donor	CHARACTER_VARYING	NOT NULL
annotation_impact	Estimation of putative impact or deleteriousness (Can either be HIGH, MODERATE, LOW or MODIFIER)	CHARACTER_VARYING	NOT NULL
gene_name	Common gene name (HGNC)	CHARACTER_VARYING	NOT NULL
gene_id	Gene ID	CHARACTER_VARYING	NOT NULL
feature_type	Transcript, motif, miRNA	CHARACTER_VARYING	NOT NULL
feature_id	Depending on the annotation, this may be: Transcript ID Motif ID, miRNA, ChipSeq peak, Histone mark, etc	CHARACTER_VARYING	NOT NULL
transcript_biotype	Description on whether the transcript is coding or noncoding	CHARACTER_VARYING	NOT NULL
rank	Exon or intron rank or total number of exons or introns	CHARACTER_VARYING	NOT NULL
hgvs_c	Variant using HGVS notation (DNA level)	CHARACTER_VARYING	NOT NULL
hgvs_p	If variant is coding, this field describes the variant using HGVS notation (protein level)	CHARACTER_VARYING	NOT NULL

cdna_pos_cdna_length	Position in cDNA and trancript's cDNA length	INTEGER	NOT NULL
cds_pos_cds_length	Position and number of coding bases	INTEGER	NOT NULL
aa_pos_aa_length	Position and number of AA	INTEGER	NOT NULL
distance	Context and implementation dependent. E.g. when the variant is “intronic” the annotation may show the distance to the closest exon; when the variant is “intergenic” it may show the distance to the closest gene; and when the variant is “upstream or downstream” may show the distance to the closest 5’UTR or 3’UTR base	INTEGER	NOT NULL
errors_warning_info	Warnings or information messages	CHARACTER_VARYING	NOT NULL

Table C.9: Database entity - Strelka_indels

Entity 10: Destruct_breakpoints - This table contains structural variants derived from deStruct.

Entity: Destruct_breakpoints			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_VARYING	NOT NULL
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_VARYING	NOT NULL
prediction_id	Unique identifier of the breakpoint prediction	INTEGER	NOT NULL

chromosome_1	Chromosome for breakend 1	CHARACTER_VARYING	NOT NULL
strand_1	Strand for breakend 1	CHARACTER_VARYING	NOT NULL
position_1	Position of breakend 1	BIGINT	NOT NULL
chromosome_2	Chromosome for breakend 2	CHARACTER_VARYING	NOT NULL
strand_2	Strand for breakend 2	CHARACTER_VARYING	NOT NULL
position_2	Position of breakend 2	BIGINT	NOT NULL
homology	Sequence homology at the breakpoint	INTEGER	NOT NULL
num_split	Total number of discordant reads split by the breakpoint	INTEGER	NOT NULL
inserted	Nucleotides inserted at the breakpoint	CHARACTER_VARYING	NOT NULL
mate_score	Average score of mate reads aligning as if concordant	FLOAT	NOT NULL
template_length_1	Length of region to which discordant reads align at breakend 1	INTEGER	NOT NULL
log_cdf	Mean cdf of discordant read alignment likelihood	FLOAT	NOT NULL
template_length_2	Length of region to which discordant reads align at breakend 2	INTEGER	NOT NULL
log_likelihood	Mean likelihood of discordant read alignments	FLOAT	NOT NULL
template_length_min	Minimum of template_length_1 and template_length_2	INTEGER	NOT NULL
num_reads	Total number of discordant reads	INTEGER	NOT NULL
num_unique_reads	Total number of discordant reads, potential PCR duplicates removed	INTEGER	NOT NULL

type	Breakpoint orientation type deletion: +- , inversion: ++ or –, duplication -+, translocation: 2 different chromosomes	CHARACTER_VARYING	NOT NULL
num_inserted	Number of untemplated nucleotides inserted at the breakpoint	INTEGER	NOT NULL
sequence	Sequence as predicted by discordant reads and possibly split reads	CHARACTER_VARYING	NOT NULL
gene_id_1	Ensembl gene id for gene at or near breakend 1	CHARACTER_VARYING	NOT NULL
gene_name_1	Name of the gene at or near breakend 1	CHARACTER_VARYING	NOT NULL
gene_location_1	Location of the gene with respect to the breakpoint for breakend 1	CHARACTER_VARYING	NOT NULL
gene_id_2	Ensembl gene id for gene at or near breakend 2	CHARACTER_VARYING	NOT NULL
gene_name_2	Name of the gene at or near breakend 2	CHARACTER_VARYING	NOT NULL
gene_location_2	Location of the gene with respect to the breakpoint for breakend 2	CHARACTER_VARYING	NOT NULL
dgv_ids	Database of genomic variants annotation for germline variants	CHARACTER_VARYING	NOT NULL

Table C.10: Database entity - Destruct_breakpoints

Entity 11: Lumpy_svs - This table contains sample specific structural variants discovered by Lumpy.

Entity: Lumpy_svs			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_VARYING	NOT NULL

tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_VARYING	NOT NULL
chrom	Chromosome	CHARACTER_VARYING	
chstart	Chromosome start	INTEGER	NOT NULL
chend	Chromosome end	INTEGER	NOT NULL
width	Width of range	CHARACTER_VARYING	NOT NULL
strand	Segment strand	CHARACTER_VARYING	NOT NULL
paramrangeid	Distinguishes which records came from which range	CHARACTER_VARYING	NULL
ref	Reference nucleotide at position (pos) of the chromosome	CHARACTER_VARYING	NOT NULL
alt	Alternate non-reference allele	CHARACTER_VARYING	NOT NULL
qual	Phred_quality score	CHARACTER_VARYING	NOT NULL
filter	PASS if the probability of being somatic is greater than threshold		NOT NULL
svtype	Type of structural variant	CHARACTER_VARYING	NOT NULL
svlen	Difference in length between REF and ALT alleles	INTEGER	NOT NULL
end	End position of the variant described in this record	BIGINT	NOT NULL
strands	Strand orientation of the adjacency in BEDPE format (DEL:+-, DUP:-+, INV:++/–)	CHARACTER_VARYING	NOT NULL
imprecise	Flag denoting imprecise structural variation	CHARACTER_VARYING	NOT NULL

cipos	Confidence interval around POS for imprecise variants	INTEGER	NOT NULL
ciend	Confidence interval around END for imprecise variants	INTEGER	NOT NULL
cipos95	Confidence interval (95%) around POS for imprecise variants	INTEGER	NOT NULL
ciend95	Confidence interval (95%) around END for imprecise variants	INTEGER	NOT NULL
mateid	ID of mate breakends	CHARACTER_ VARYING	NOT NULL
event	ID of event associated to breakend	CHARACTER_ VARYING	NOT NULL
secondary	Flag denoting secondary breakend in a multi-line variants	CHARACTER_ VARYING	NOT NULL
su	Number of pieces of evidence supporting the variant across all samples	INTEGER	NOT NULL
pe	Number of paired-end reads supporting the variant across all samples	INTEGER	NOT NULL
sr	Number of split reads supporting the variant across all samples	INTEGER	NOT NULL
bd	Amount of BED evidence supporting the variant across all samples	INTEGER	NOT NULL
ev	Type of LUMPY evidence contributing to the variant call	CHARACTER_ VARYING	NOT NULL
prpos	Probability curve of the POS breakend	CHARACTER_ VARYING	NOT NULL
prend	Probability curve of the END breakend	CHARACTER_ VARYING	NOT NULL

Table C.11: Database entity - Lumpy_svs

Entity 12: Svs_filtered - Table containing breakpoints filtered for low mapability regions from both deStruct and lumpy

Entity: Svs_filtered			
Field Name	**Description**	**Data Type**	**Constraints**
id	Serial id	CHARACTER_ VARYING	NOT NULL
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_ VARYING	NOT NULL
chrom_1	Chromosome for breakend 1	CHARACTER_ VARYING	NOT NULL
brk_1	Break 1	BIGINT	NOT NULL
chrom_2	Chromosome for breakend 2	CHARACTER_ VARYING	NOT NULL
brk_2	Break 2	BIGINT	NOT NULL
homlen	Length of base pair identical micro-homology at event breakpoints	BIGINT	NOT NULL
brk_dist	Break distance	BIGINT	NOT NULL
type	Type of structural variant	CHARACTER_ VARYING	NOT NULL

Table C.12: Database entity - Svs_filtered

Entity 13: Bamstats_tumour - This table contains sequencing statistics derived from the bam file of each tumour sample.

Entity: Bamstats_tumour			
Field Name	**Description**	**Data Type**	**Constraints**
tumour_id	Sample identifier associated with the variant called at each position	CHARACTER_ VARYING	NOT NULL
total_reads	Number of reads that are in a sample's bam file	INTEGER	NOT NULL
qc_failure	Number of reads marked QC failure	INTEGER	NOT NULL

duplicates	Number of duplicate reads	INTEGER	NOT NULL
mapped	Number of reads marked mapped in the flag	INTEGER	
mapped_percentage	Number of mapped reads / total reads(%)	FLOAT	NOT NULL
paired_in_seq	Reads paired in sequencing	INTEGER	NOT NULL
read1	Count read1	INTEGER	NOT NULL
read2	Count read2	INTEGER	
properly_paired	Properly paired reads	INTEGER	NOT NULL
properly_paired_percentage	Percentage of properly paired reads	FLOAT	NOT NULL
self_and_mate_mapped	Number of reads for which both reads mapped	INTEGER	NOT NULL
singletons	Reads that mapped but the mate didn't	INTEGER	
singletons_percentage	Percentage of reads that mapped but the mate didn't	FLOAT	NOT NULL
mate_map_diff_chr	Number of reads with a mate mapped on a different chromosome	INTEGER	NOT NULL
mate_map_diff_chr_mapq	Number of reads with a mate mapped on a different chromosome - mapping quality	INTEGER	NOT NULL
mapq	The phred scaled probability of the alignment/base being wrong	FLOAT	NOT NULL
avg_read_coverage	Average read coverage	FLOAT	NOT NULL

Table C.13: Database entity - Bamstats_tumour

Entity 14: Bamstats_normal - This table contains sequencing statistics derived from the bam file of each normal sample.

Entity: Bamstats_normal

Field Name	**Description**	**Data Type**	**Constraints**
normal_id	Matched normal-sample identifier	CHARACTER_ VARYING	NOT NULL
total_reads	Number of reads that are in a sample's bam file	INTEGER	NOT NULL
qc_failure	Number of reads marked QC failure	INTEGER	NOT NULL
duplicates	Number of duplicate reads	INTEGER	NOT NULL
mapped	Number of reads marked mapped in the flag	INTEGER	
mapped_ percentage	Number of mapped reads / total reads(%)	FLOAT	NOT NULL
paired_in_seq	Reads paired in sequencing	INTEGER	NOT NULL
read1	Count read1	INTEGER	NOT NULL
read2	Count read2	INTEGER	
properly_paired	Properly paired reads	INTEGER	NOT NULL
properly_paired _percentage	Percentage of properly paired reads	FLOAT	NOT NULL
self_and_mate _mapped	Number of reads for which both reads mapped	INTEGER	NOT NULL
singletons	Reads that mapped but the mate didn't	INTEGER	
singletons _percentage	Percentage of reads that mapped but the mate didn't	FLOAT	NOT NULL
mate_map_diff _chr	Number of reads with a mate mapped on a different chromosome	INTEGER	NOT NULL
mate_map_diff _chr_mapq	Number of reads with a mate mapped on a different chromosome - mapping quality	INTEGER	NOT NULL
mapq	The phred scaled probability of the alignment/base being wrong	FLOAT	NOT NULL

avg_read_coverage	Average read coverage	FLOAT	NOT NULL

Table C.14: Database entity - Bamstats_normal