UNIVERSITÀ DI TRENTO
Dipartimento di
Biologia Cellulare, Computazionale e Integrata
CIBIO

BACHELOR'S DEGREE IN Biomolecular Sciences and Technology

Development of a web-based resource for breast cancer genomic data exploration

Supervisor:

Alessandro Romanel

Graduant:

Andrea Tonina

Co-supervisor:

Nicola Casiraghi

Contents

Abstract

The constant increase of genomic data generated from rapid and cost-effective high-throughput sequencing technologies represent a fundamental resource to get detailed insights of tumor samples composition by a comprehensive view of the mutational landscape from single base mutations to chromosomal or whole genome-scale events. Consequently, the need for computational suites for analyzing, summarizing, and extract meaningful information are in high demand.

Breast cancer (BRCA) has emerged as the second major cancer type comprising almost 25% of all cancers among women. In the last decades, scientific research focused on the characterization of its molecular subtypes to identify driver somatic events and consequently improve the efficacy of current clinical treatments. In this work, I present BroadBand, a web-based resource developed with the Shiny R package, designed for the exploration of genomic data selected among a comprehensive selection of breast cancer (BRCA) studies with available information on somatic single nucleotide variants (SNVs), somatic copy number alterations (SCNAs), and stratified by BRCA molecular subtypes. BroadBand allows an intuitive and custom exploration of SNVs and SCNAs data which can be filtered and summarized with a resolution level from gene to chromosome level and stratified across BRCA molecular subtypes and tumor classifications.

All retrieved information is reported and summarized as figures and tables, resulting in a valuable resource for additional downstream analysis or novel experimental designs. The BroadBand GitHub repository is available at https://github.com/cibiobcg/BroadBand and it is also possible to see how the application works at https://bcglab.cibio.unitn.it/broadband.

1 Introduction

This section will provide an overview of cancer evolution models, tumor heterogeneity, and the characterization of primary and metastatic tumors. Furthermore, I will introduce molecular and clinical aspects specific to breast cancer and ultimately the relevance of computational approaches to exploit genomic data in the oncology field.

1.1 Evolutionary tumor models

Peter Nowell published in 1976 a landmark perspective on cancer as an evolutionary process that is driven by stepwise, somatic-cell mutations with sequential, subclonal selection [1]. Based on this paradigm, the evolution of tumors, that are influenced by endogenous and exogenous mutation, follows the evolutionary system of Darwin based on the selection of the fittest hereditary genetic variants. This model has been deeply investigated thanks to modern cancer biology and genomic data obtained through analysis.

Cancer evolves by an iterative process of clonal expansion, genetic diversification, and clonal selection within the adaptive landscapes of tissue ecosystem [2]. The acquisition of heritable alterations and genetic drift are random processes but on the other hand Darwinian selection is a deterministic process.

For example, if multiple subclones possess different driver mutations, displaying different phenotypes can lead to clonal competitions and the fitness of an individual subclone is then defined in relation to the fitness of other competing clones.

Mutation accumulation in all cancer types can occur gradually, constituting the perpetual adaptation of the tumor but can also occur through punctuated episodes that are highly disruptive. This subdivision has been included in the argumentation of micro-evolution, that stays for gradual accumulation of point mutations, in opposition of macro-evolution, which emphasizes the importance of large-scale chromosomal alterations and bursts of mutations [3], meaning, that the genome evolution of cancer is not always a gradual process.

About the clonal heterogeneity, tumoral clones evolve through interactions of selectively advantageous “driver” lesions, selectively neutral “passenger” lesions and deleterious lesions. Moreover, the “mutator” lesions increase the rate of the other genetic changes [4,5]. In this scenario, selective pressures allow some subclone mutations to expand and other to extinct or remain dominant; also, the mutation profile of a tumor represents the historical record of its evolutionary history which can be used to understand the temporal order of mutation events which can determine which of those are clonal and which are instead subclonal [6].

Evidence suggests that certain diver alterations can be more likely to be subclonal than other and

specific [3,7–12] driver mutations tend to be clonal in some cancer types but not in other [7,12,13]. It also seems that somatic mutations that lend a clonal advantage are positively selected during evolution.

1.2 Tumor heterogeneity

As a result of evolutionary forces of variation and selection, extensive genetic and phenotypic variations exist not only across cancers (inter-tumor heterogeneity) but also within individual tumors (intra-tumor heterogeneity). Tumors that originate from different tissues and cell types vary in terms of their genomic

landscapes, prognosis, and their response to treatments. Mutational frequencies of oncogenes and tumor suppressors vary between tumors of different tissues, probably reflecting the importance of distinct tissue dependent signaling pathways.

Within tumors, genetically distinct subclonal populations of cells arise through inter-cellular genetic variation, followed by selective outgrowth of clones having a phenotypic advantage within a given tumor environmental context [1,14]. If a new clone takes over the entire population by replacing ancestral ones, this will result in a homogenous cell population. Otherwise, if during linear evolution a new clone fails to outcompete its predecessors, a degree of heterogeneity will be observed [15] and if distinct subclones evolve in parallel (branched tumor evolution) this will result in extensive subclonal diversity [2]. Analysis of large cancer databases support evidence of the genetic heterogeneity between cancers and even within an individual cancer type. In addition to the heterogeneity of cancer genes, there is considerable diversity in the nature, number, and distribution of mutations within and across different cancer histologies [16]. These studies have revealed that the degree of intra-tumor heterogeneity can be highly variable, with between zero and thousands coding mutations found to be heterogeneous within primary tumors or between primary and metastatic or recurrence sites [17]. Genomic copy number heterogeneity can also be extensive within tumors. Large scale chromosomal alterations may have profound impact on the genome architecture, disrupting hundreds of genes, and can be considered macro-evolutionary events, which may contribute to tumor progression [13,18,19].

1.3 Personalized cancer medicine

Recognition of tumor heterogeneity led to the concept of personalized cancer medicine: deciphering individual cancer genomic profiles should provide precise insights into disease biology and allow the targeting of genetically encoded susceptibilities for therapeutic benefit. Indeed, intra-tumor genetic heterogeneity results in phenotypic diversity affecting clinically relevant parameters such as gene expression signatures that reflect prognosis and response to therapeutic agents. It is also important to note that phenotypic heterogeneity is not only mediated through genetic diversity; genetically

homogenous subclones can behave in functionally distinct ways after exposure to chemotherapy [20]. Therapeutic intervention may destroy subclones and alter their favorable microenvironment, but it can also provide a potent selective pressure for the expansion of resistant variants.

1.4 Primary and metastatic tumor

Tumor metastasis is frequently cited to be responsible for about 90% of all cancer-related deaths. The process has been linked to a speciation event with macro-evolutionary leaps required to endow a tumor cell with metastatic potential [21]. Next-generation DNA sequencing has made it apparent that most primary tumors do not consist of a single population of genetically identical cells. They are a collection of subpopulations of genetically identical cells that can be distinguished from other subclones by the mutations they harbor. The evolutionary paths from primary tumors to metastasis that are taken by tumor cells are many and represent a challenging research issue. It is often assumed that one disseminated a tumor cell initiates metastatic outgrowth (monoclonal seeding) and there is debate about whether metastases derive from multiple branched spreading events involving disseminated cells from the primary tumor as well as metastases (polyclonal seeding).

The monoclonal origin model indicates that all metastatic lesions are derived from a common cancer cell ancestor traceable back to one distinct focus that is present in the primary prostate tumor [22,23]. In the polyclonal origin model, multiple genomically distinct foci in the primary tumor, without sharing a common cancer cell ancestor, can independently progress and metastases can harbor multiple distinct clonal aberrations originating from the primary tumor [22,24]. Additionally, in both models the acquisition of subsequent mutations can also occur during disease progression and/or metastasis-to-metastasis cross-seeding (in which subclones within a metastasis originated from another metastatic site, rather than from the primary tumor) [22,24], leading to substantial genomic diversity.

1.5 Breast cancer epidemiology

Globally, the incidence of breast cancer has been rising with an annual increase of 3.1% with a start of 641.000 cases in 1980, and an increase to more than 1.6 million in 2010 [25] and the trend is likely to continue. This proliferation is increasing in countries regardless of the income level due to the growth and aging of the population; indeed, the female population represents 49.5% of the global population and a larger portion of the population >60 years of age. Data also shows a higher incidence in high-income countries than lower income regions [26,27] thanks to better awareness of risk factors and the broader availability of mammography tests. Consequently, breast cancer is often diagnosed early and the prognosis is good in high-income countries. On the other hand, in low and middle-income countries, the tumor is often diagnosed later, resulting in a higher mortality rate [28].

In the last decade, clinical treatments evolved mainly considering that breast cancer is a molecular heterogeneous disease. The classification proposed by Perou and Sorlie in 2000 stratifies the breast cancer disease into four molecular subtypes: luminal A and luminal B which express estrogen receptor (ER), basal like and human epidermal growth factor receptor 2 (HER2, encoded by *ERBB2*) characterized by a lack of ER expression [29]. Breast cancer expressing estrogen receptor and/or progesterone receptor (PR) are considered hormone receptor-positive breast cancer, where tumors that don't express ER, PR or HER2 are classified as triple negative breast cancer (TNBC) [30] (**Table 1, and 2**).

Table 1: Overview and full annotations of breast cancer molecular subtypes.

BRCA molecular subtypes	Hormone receptors	HER2	Ki-67	Annotation
Luminal A	+	-	Low	Low-grade, tend to grow slowly and have the best prognosis
Normal-like	+	-	Low	Slightly worse prognosis than luminal A
Luminal B	+	+	Low	Grow slightly faster than Luminal A cancers and prognosis is slightly worse
	+	-	High	
Triple-negative	-	-	No consensus	More common in women with *BRCA1* mutations
HER2-enriched (non-luminal)	-	+	High	Grow faster than luminal cancers and have worse prognosis

Table 2: Summary and major classification of breast cancer molecular subtypes.

BRCA molecular subtypes	Hormone receptors	HER2	Label used in this work
HER2 positive	-	+	HER2+
ER positive	+	-	ER+
Triple-negative	-	-	TNBC

Different subtypes lead to different death rates, with HER2 subtypes showing the higher rate of death, followed by the TNBC, luminal A and luminal B subtypes [31]. Moreover, the incidence of tumor subtypes varies by ethnicity, for example African-American women have the highest rate of TNBC compared with any other ethnic groups and they also show higher rates of metastatic disease which is associated with a lower survival rate [32].

Even though it varies by ethnicity, across countries, and molecular subtypes, approximately 10% of all breast cancers are inherited and associated with family history [33]. Indeed, evidence suggests that individuals with first degree relatives who had breast cancer have an elevated relative risk of early onset breast cancer [34].

1.6 Breast cancer pathophysiology

The mechanism by which breast cancer begins is still unknown, nevertheless much effort has been made to molecularly characterize breast cancer and delineate its genesis and progression. Hormones are a stimulus for breast development during puberty, menstrual cycles and pregnancy, but they are also a major risk factors for sporadic breast cancer where estrogen is a promoter of breast cancer for binding ER located in the nucleus, which is a ligand-activated transcription factor.

At the morphological level, there is a continuum of lesions and genetic modifications from normal glands to cancer. At the molecular level, there is evidence that shows two distinct molecular pathways of progression for breast cancer, which are related to ER expression, tumor grade and its proliferation [30].

The first path (low-grade-like pathway) characterized by the gain of 1q, loss 16q, uncommon amplification of 17q12 and gene expression signature with major part of gene associated with ER phenotype, diploid karyotypes and low tumor grade, luminal A and to a limited degree luminal B fall in this pathway.

The second path (high-grade-like pathway) is distinguished by the loss of 13q, gain of chromosomal region 11q13, amplification of 17q12 that contains *ERBB2* (which encodes for HER2) and genes involved in cell cycle and proliferation [35]; HER2 positive tumors and TNBC fall in this category [36]. HER2 it's a member of the human epidermal growth factor family, this family protein includes an extracellular ligand-binding domain, a transmembrane domain, and a tyrosine kinase intracellular catalytic domain. *ERBB2* is amplified in 13–15% of breast cancers, causing the activation of the HER2 pathway. The activation of HER2 takes place through dimerization after binding of the ligand, but as of now no specific ligand has been identified [30]. HER2 pathway proliferation, cell survival, metastasis, and adhesion through different pathways, targeting HER2 shows to be effective in HER2

positive breast cancer that are defined by protein overexpression or gene amplification.

In early-stage breast cancer, the genes that are reported as the most mutated by a somatic point mutation or copy number amplification are *TP53* (41% of tumors), *PIK3CA* (30%), *MYC* (20%), *PTEN* (16%), *CCND1* (16%), *ERBB2* (13%), *FGFR1* (11%) and *GATA3* (10%) [37]. These genes encode cell-cycle modulators which can be repressed or activated that are important for proliferation and/or inhibiting apoptosis, inhibiting oncogene pathways, or inhibiting elements that are no longer repressed. Most breast cancers are caused by multiple, low-penetrant mutations that act cumulatively. Luminal A tumors have a high prevalence of *PIK3CA* mutations (49%), whereas a high prevalence of *TP53* mutations is a hallmark of basal-like tumors (84%). TNBC, different molecular drivers under-line its subtypes, for metastatic stage, specific predictive alterations like *PIK3CA* mutations, can be also detected non-invasively from plasma samples in circulating free tumor DNA [38].

Genes in breast cancer can be globally hypomethylated leading to gene activation, upregulation of oncogenes and chromosomal instability or more focally hypermethylated (locus-specific) leading to gene repression and genetic instability due the silencing of DNA repair genes; those are less frequently. Other epigenetic mechanisms involve histone tail modifications by DNA methylation, inducing changes in chromatin structure which lead to silent gene expression and remodeling of nucleosomes. These changes are reversible, enzyme-mediated and potentially targetable [39]. An example is for luminal-like breast cancer cell lines, inhibition of histone deacetylase with specific inhibitors such as Vorinostat [40] or Chidamide [41] can reverse resistance to endocrine therapy via inhibition of the resistance pathway driven by epidermal growth factor receptor signaling.

Most of the driver alterations of primary breast cancer are also found in the metastatic sites, however different metastatic sites may hold “private” aberrations including new drivers, resulting in intra-patient tumor heterogeneity. Specifically, these alterations occur later, and some are triggered by treatment pressure. As an example, *ESR1* mutations can emerge after treatment with aromatase inhibitors. These acquired mutations influence the ligand-binding domain and are identified in metastatic tissue or plasma in women with breast cancer after being treated successfully with aromatase inhibitor in 23-40% of cases. During the metastatic development, different deposits (tumor deposits are irregular discrete tumor masses in adipose tissue, discontinuous from the primary tumor, that are described in various cancers [42]) show linear, parallel or poly-clonal evolutionary pathways from the primary tumor, which lead to various genetic and epigenetic evolutions [43]. Discrepancies between primary and metastatic breast cancer regarding the expression of ER, PR and HER2 can be explained by the subclonal diversification. These molecular targets are more commonly lost than

newly acquired; indeed, for HER2 positive primary tumors the 13% generates HER2 negative metastases but only 5% of HER2 negative primary tumors generate positive metastases [44].

1.7 Computational tools for cancer genomics

High throughput sequencing technologies have been proved extremely useful to unravel the mutational landscape of human tumors. Indeed, targeted sequencing, whole-exome sequencing (WES), and whole-genome sequencing (WGS) assays have been extensively exploited to study tumor evolution and heterogeneity across multiple cancer types. These assays and the analysis of generated genomic data allowed the discovery and characterization of novel genomic alterations, and the expansion of the catalog of known driver mutations (which are somatic alterations in DNA sequence that lead to selective advantages). Moreover, an accurate characterization of somatic alterations based on genomic sequencing data is critical to get molecular insights of the tumor. However, both the application and the interpretation of the results from a computational analysis is not always straightforward. As an example, different bioinformatics methods could provide divergent results in the identification of somatic variants when harbored by a small fraction of tumor cells or in the accurate detection of complex structural genomic alterations.

Detection of Single Nucleotide Variants

One of the most common type of somatic alterations detected in tumors are the single nucleotide variants (SNV, also referred as somatic point mutations), which are changes in the DNA sequence involving one base pair.

In a typical tumor-normal pair analysis, somatic SNVs are identified as those genomic loci where tumor reads supporting an alternative allele are not present in the matched normal sample. The computation of the *variant allele fraction* (VAF), which is the proportion of sequencing tumor reads supporting a candidate mutation, is a determinant measure for an accurate detection. Furthermore, since the VAF is proportional to the number of tumor cells carrying the mutations, this provides valuable information about its level of clonality.

Most of available computational methods designed for SNVs detection, utilize matched normal samples (i.e. peripheral blood samples) to identify germline (non-somatic) single nucleotide polymorphisms, and then apply custom thresholds (i.e. minimum number of reads, minimum VAF) to define a final set of putative somatic point mutations. More advanced algorithms, like MuTect2, CaVEMan, STrelka2, VarDict and MuSE [45–49] evaluate the genotype of normal and tumor samples together with prior probabilities for genotypes and frequencies by taking into account different sample-specific features such as the type of the mutation (transitions versus transversions), the tumor

purity and ploidy, and local features such as copy number alterations. Recently, machine learning based tools, like ISOWN [50], can discriminate between somatic and germline variants by using data of both somatic (COSMIC [51]) and germline (dbSNP [52]) polymorphisms.

Identifying cancer driver mutations among all the detected SNVs is another challenging task. Thanks to the availability of a large amount of cancer genomic data and data-driven methods (MutSig2CV [53], dNdScv [54]), a higher mutation frequency than expected across samples is usually evidence of positive selection. These alterations are considered driver events. Another important step is the annotation and prioritization of detected somatic mutations based on their potential functional consequences. Variant Effect Predictor(VEP) [55], SnpEff [56], ANNOVAR [57] are the most used annotators, and they report the effect of the genomic mutation on transcripts and establish if an observed mutation was already reported in a previously published cancer study and if it is pathogenic or potentially actionable. Databases such as COSMIC [51], ClinVar [58], and OncoKb [59] are fundamental for this step.

Detection of Somatic Copy Number Alterations

Somatic copy number alterations (SCNAs) include deletion and amplification of DNA regions with sizes varying from a few kilobases up to entire chromosomes. SCNAs are particularly common in cancer and contribute to its development.

Even though data from WGS assays show to be the best method for detecting SCNAs, multiple computational method have been developed and optimized to infer SCNA profiles also from WES [60–66] or targeted sequencing data [67] though with a lower resolution and accuracy.

Computational methods for identifying SCNAs from sequencing data adapt techniques from the analysis of CHG (comparative genomic hybridization) and SNP-array assays data. Indeed, by computing the “read depth” along the genome (the average number of sequencing reads spanning a defined genomic region of interest) as the main component, the methods segment the genome into regions characterized by specific copy numbers. Most segmentation algorithms are based on Hidden Markov Models (HMM), circular binary segmentation (CBS) [68], and piecewise constant fitting regression [69]. Since the "read depth" varies along the genome because of the GC content, DNA mappability, and other genomic features, it is necessary to apply matched control and perform ad-hoc pre-processing normalization steps [70].

Most advanced methods use the frequencies of minor alleles, defined as *BAF* (B-allele frequency,

which is the fraction of sequencing reads supporting one allele at a heterozygous single nucleotide polymorphism locus) for the segmentation and detection of the allelic specific copy number alterations and highlight eventual copy-neutral LOH events (the complete loss of only one allele in a bi-allelic region). Methods like EXOMECNV [63] and CONTRA [65] perform normalization of the read coverage and run a segmentation algorithm on tumor-normal ratio profile, where others like CopywriteR [71] and CNVkit [72] try to recapitulate the genome-wide profile from WES and targeted assays by including in the computation also off-target reads. More recent algorithms like Sequenza [73] and FACETS [68] detect allele-specific copy number alterations and infer tumor purity and ploidy by utilizing both sequencing read depth and BAF data.

When SCNAs are found in a set of samples with an alteration frequency higher than expected, the usage of algorithms such as GISTIC [74] can identify those regions that are likely to be driven events.

2 Aim

This project aims to fulfill the need for a resource that allows data exploration to list and rank genomic regions (from gene to chromosome-arm level) according to their frequency of alteration in breast cancers (BRCA). To achieve this result, I collected genomic data from different publicly available BRCA cohorts to get a detailed and comprehensive overview of critical somatic alterations, specifically somatic single nucleotide variants and somatic copy number aberrations. Thus, I first stratified the collected data in different classes based on available clinical and molecular BRCA subtypes. Then I computed the frequency of aberrations of genes and wider genomic regions in each category. As a result, my thesis work is a practical and interactive web-based resource, named BroadBand, able to summarize and visualize genomic data. Moreover, BroadBand highlights the importance of stratifying data based on clinical and molecular information to achieve higher resolution and precision in assessing genomic features across known BRCA subtypes.

3 Materials and Methods

3.1 Breast Cancer Datasets

Breast cancers (BRCA) datasets of interest have been selected and inspected through the cBioPortal for Cancer Genomics resource (https://www.cbioportal.org). The selected studies consist of cohorts where tumor samples from BRCA oncological patients have been molecularly profiled via Whole Exome Sequencing (WES) approach. For each dataset (**Table 3**), I downloaded, when available, pre-

processed and ready-to-use data of both somatic single nucleotide variants and somatic copy number alterations. Specifically, SNVs data were filtered to retain only those mutations causing a non-synonymous base substitution in a protein coding gene and SCNAs data were elaborated to convert copy number information from gene-level to a wider genomic range, such as genomic cytoband. Moreover, I also collected molecular details relative to the BRCA molecular subtypes (HER2+, HR+, TNBC) and clinical classification of the tumor (Metastatic or Primary) to annotate each patient (and associated data) accordingly.

Table 3. Datasets used by BroadBand. The first column contains the names of each dataset that are included and used by default in BroadBand, the second column indicates the number of patients available in each dataset and the third and fourth columns report the datasets references.

Names	Numbers of patients	Paper references	PMID
brca_metabric	1980	Breast Cancer (METABRIC, Nature 2012 & Nat Commun 2016)	27161491, 30867590, 22522925
brca_igr_2015	208	Metastatic breast cancer (INSERM-PLoS Med 2016)	28027327
brca_mbcproject_wagle_2017	106	The Metastatic Breast Cancer Project (Provisional - February 2020)	https://mbcproject.org/data-release
breast_msk_2018	1739	MSK-IMPACT Clinical sequencing Cohort (MSKK, Nat Med 2017)	28481359
brca-tcga_pan_can_atlas_2018	945	Breast Invasive Carcinoma (TCGA, PanCancer Atlas)	29625048, 29596782, 29622463, 29617662, 29625055, 29625050, 29617662, 30643250, 32214244, 29625049, 29850653

3.2 Computation of SNVs aberration frequencies

SNVs data of BRCA datasets of interest (**Table 1**) have been downloaded, when available, from the cBioPortal online resource (https://www.cbioportal.org) as VCF (the Variant Call Format) files. Then, VCF files have been filtered to retain only those somatic mutations occurring in a protein coding gene and causing a non-synonymous base substitution. These filtered VCFs have been used to compute for each gene its frequency of aberration across BRCA molecular subtypes and tumor classifications. Since different BRCA datasets comprise a different number of patients/samples, the frequency of aberration of each gene is computed as weighted mean so that frequencies are weighted accordingly with the number of samples to obtain a more reliable measure.

3.3 Computation of SCNAs aberration frequencies

SCNAs data of BRCA datasets of interest (**Table 1**) have been downloaded, when available, from the cBioPortal online resource (https://www.cbioportal.org). Specifically, the copy number status of each gene is encoded as -2 and -1 to indicate a homozygous and a hemizygous deletion, respectively. Gain and amplification events are encoded as 1 and 2, respectively. Genes with normal diploid copy number status are labeled as 0. Moreover, each gene has been annotated with its genomic location (chromosome, arm, cytoband) and flagged if it is a known cancer driver gene. Thus, I computed for each gene its frequency of aberration (one for each of the 4 SCNA events encoded) across BRCA molecular subtypes and tumor classifications. Frequencies of aberrations have been computed at two genomic levels. The first one is the gene level; here, frequencies are computed as the ratio between the number of samples with the considered gene affected by a specific SCNA event (i.e., hemizygous deletion) and the total number of samples with available SCNA data. Frequencies of aberration computed at the gene level have been combined to compute frequencies at the cytoband level. Indeed, the frequency of aberration of a given cytoband is calculated as the median of the frequencies of aberration of all genes that sit in it. For each cytoband, I also saved which gene is the most aberrant.

3.4 Code implementation and availability

I performed data analysis by designing and developing custom scripts in the R programming language. I mainly used functionalities of the ggplot and dplyr packages, both included in **Tidyverse** (https://www.tidyverse.org), a collection of numerous R packages designed specifically for data science. For data manipulation we used the library **dplyr**, which contains functions such as *summarise*, *group_by*, *filter* and *select*. This is extremely helpful to build up and remodeling a given dataset, for example by keeping only certain variables of interest (like the class of the tumor, the type, the genes), filtering the entries based on what the users want to inspect (i.e., the 5 most expressed

genes) and summarizing multiple values into a single one. I then used the **ggplot2** library for data visualization by plotting results of computed analysis as histograms, heat maps, and bar charts.

BroadBand is a web-based application powered by the **R shiny** library (https://shiny.rstudio.com). R shiny allows to structure and graphically personalize (via *shinytheme*) the user interface of a web application and customize it with several widgets. These widgets (i.e., filtering criteria and data selection) facilitate the interaction between the user and the visualized data in a reactive way by exploiting, for example, the combination management of some checkboxes (*shinyjs*). Every Shiny app is built on an HTML document that creates the apps' user interface.

For this project, I used Git as a versioning control system and the hosting service **Github** to make easier collaborations, keep track of changes in the code, and make it easily available to the scientific community. Moreover**,** I used **Singularity**, a platform that allows to create and run containers that package up pieces of software in a way that is both fully portable and reproducible.

The BroadBand github repository and the singularity image are available at: https://github.com/cibiobcg/BroadBand.

4 Results

4.1 Data exploration of Breast Cancer datasets

I here present a web-based resource to extract and summarize information derived from genomic data of BRCA oncological patients. The developed resource takes as input selected pre-processed somatic single nucleotides alterations and copy number aberrations data from publicly available datasets (**Table 1**) to assess the most frequently aberrant genomic regions, from gene- to chromosome-level, across BRCA tumor subtypes.

The web-resource, organized in 3 different sections, allows the user to explore, visualize and download information of interest. Each section reports results of a specific analysis that can be customized by tuning multiple settings.

The first section provides a data overview of considered datasets, reporting a) annotations about genomic data availability and tumor classification, b) overall summary statistics such as the number of samples stratified by tumor molecular subtypes. The second section focuses on gene aberration frequencies based on somatic single nucleotide variants while the third section analyses somatic copy number alterations and compute aberration frequencies at different resolutions: from gene to chromosome level.

4.2 Cohort overview and samples stratification

The “Overview” panel, the home page of the BroadBand resource, provides a detailed report (as bar chart and table) about the number of samples, carefully stratified by molecular features and data type availability, included in the selected cohort(s).

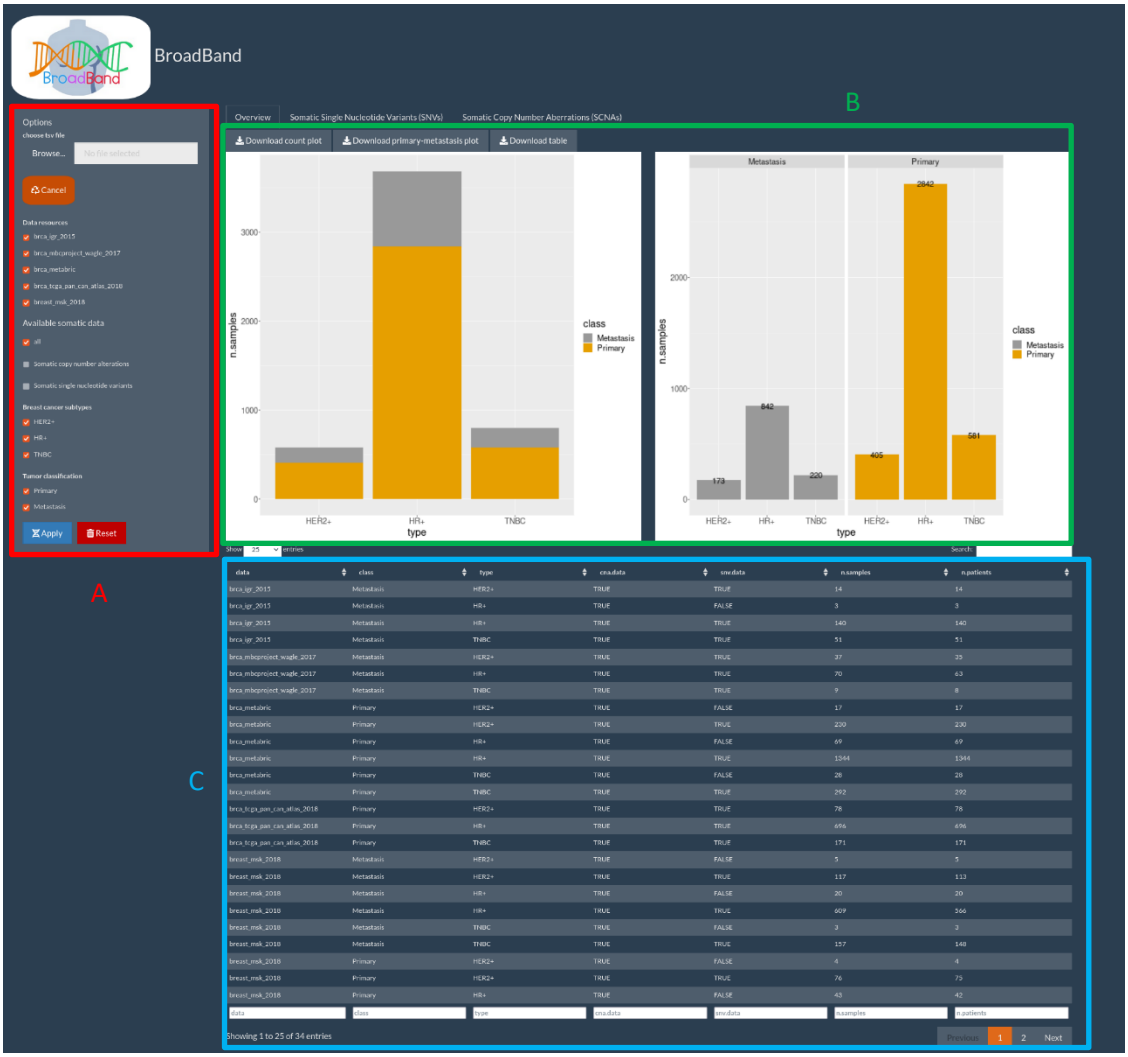

Figure 1. The “Overview” panel. The figure shows the layout of the home panel, named “Overview”, which reports the composition of selected BRCA cohorts. The panel is organized in 3 main sections: the widgets (A) allow the user to make a custom selection of data of interest, the area of the plots (B) where selected data are summarized and visually represented as bar charts or as a table (C).

The widget panel (**Figure 1A**) is populated with multiple selection options which will update the BRCA data displayed by default. Specifically, the widget panel allows the user to:

1. Upload a new BRCA dataset to be inspected in addition to those ones available by default. The *cancel* button will reverse the operation by removing the uploaded dataset among available data.
2. Select samples based on somatic aberrations data of interest. For example, the user can select only those samples with SCNAs and/or somatic (non-synonymous) SNVs data available.

3. Select BRCA cohorts of interest. If the user uploaded a custom dataset, it will be listed here.
4. Select data based on BRCA subtypes.
5. Select data based on tumor classification.
6. Reset all settings to default.
7. Apply selected options to data and update plots and the table accordingly.

In addition to the widgets bar, the page layout also includes two plots, a table and three downloads buttons which allow the user to save all outputs displayed. The plots and the table are the results of options selected by the user. Specifically, the first bar chart reports the number of samples stratified by cancer subtypes and tumor classification, the second plot is composed by two bar charts, one for each tumor class (primary and metastatic), and reports sample counts by cancer subtypes. The table (**Figure 1C**) reports data represented in all plots.

4.3 Genes aberrations frequencies based on somatic SNVs data

The second panel is focused on somatic single nucleotide variants data and provides a comprehensive report about the frequencies of aberration of genes.

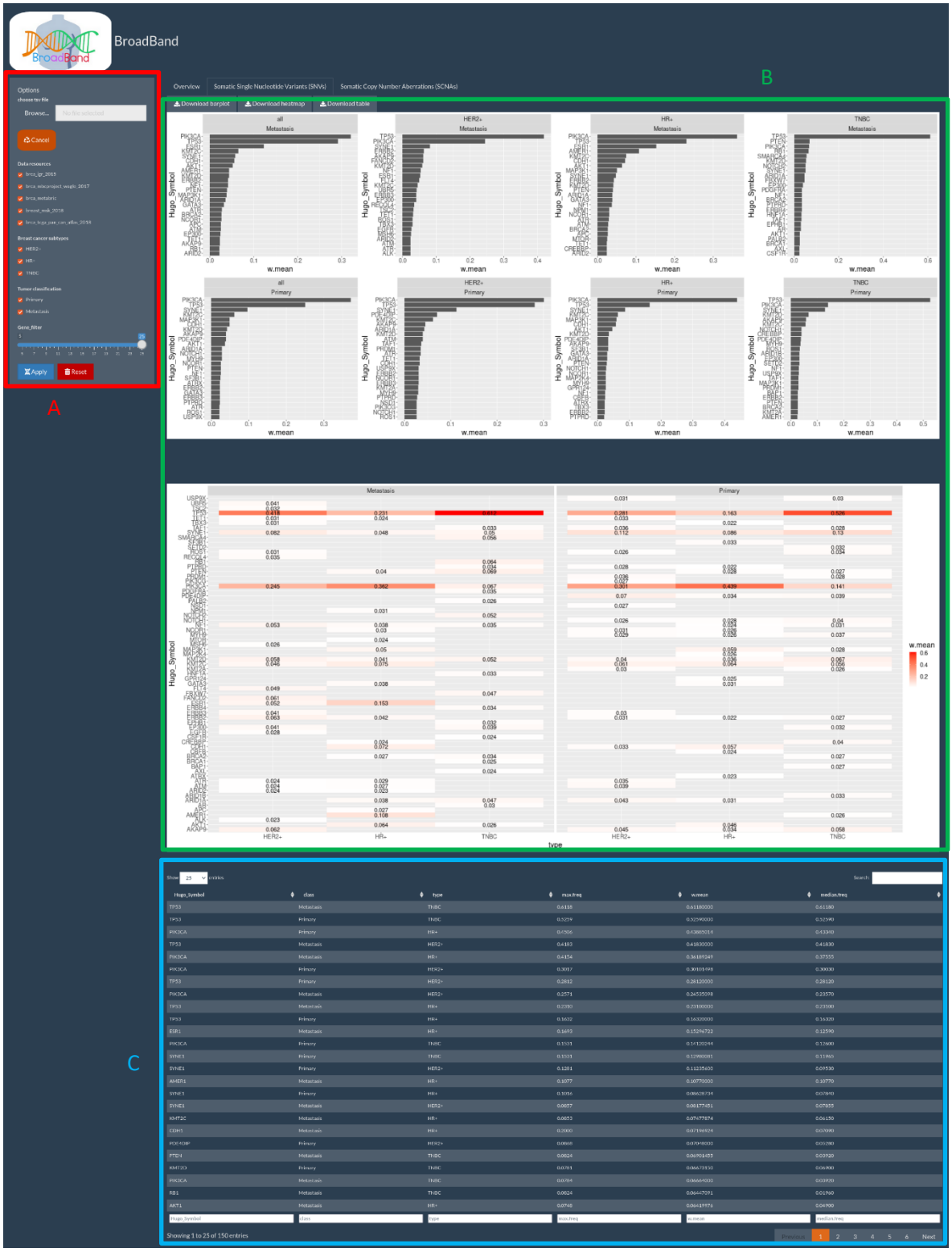

Figure 2. The "Somatic single nucleotide variants" panel. The figure shows the layout of the panel, named "Somatic single nucleotide variants", which ranks genes based on their frequency to be altered by a non-synonymous single nucleotide variant. The panel is divided into 3 main sections: the widgets (A) allow the user to make a custom selection of data of interest, the area of the plots (B) where selected data are summarized and visually represented as bar charts or as a table (C).

The widget panel (**Figure 2A**) is populated with multiple selection options which will update the BRCA data displayed by default. Specifically, the widget panel allows the user to:

1. Upload a new BRCA dataset to be inspected in addition to those ones available by default. The *cancel* button will reverse the operation by removing the uploaded dataset among available data.
2. Select BRCA cohorts of interest. If the user uploaded a custom dataset, it will be listed here.
3. Select data based on BRCA subtypes.
4. Select data based on tumor classification.
5. Select the N most frequently aberrant genes by using the provided slider.
6. Reset all settings to default.
7. Apply selected options to data and update plots and the table accordingly.

In addition to the widgets bar, the page layout also includes two plots, a table and three downloads buttons which allow the user to save all outputs displayed. The plots and the table are the results of options selected by the user. Specifically, the first plot (**Figure 2B**, left plot) comprises multiple bar charts, data represented are stratified by tumor classification (one for each row) and for breast cancer subtypes (one for each column; the first column "all" comprises tumor subtypes data altogether) as reported on top of each inset panel. Within each bar chart, genes (listed on the y-axis) are ranked based on their aberration frequency.

Alongside (**Figure 2B**, right plot), the heatmap compares aberration frequencies of genes between primary and metastatic and by cancer subtypes. Data represented in this way could be helpful to spot by visual inspection which genes are mutually exclusive within a single tumor class. The table (**Figure 2C**) reports data represented in both plots.

4.4 Chromosomal aberrations frequencies based on SCNAs data

The third panel is focused on somatic copy number aberrations data, this panel provides a full report about the frequencies of somatic copy number aberrations occurring in genomic regions of different ranges. Specifically, aberrations occurring in gene, chromosomal cytobands and chromosomal arms, are computed based on molecular features, tumor classification of available data included in the selected cohorts.

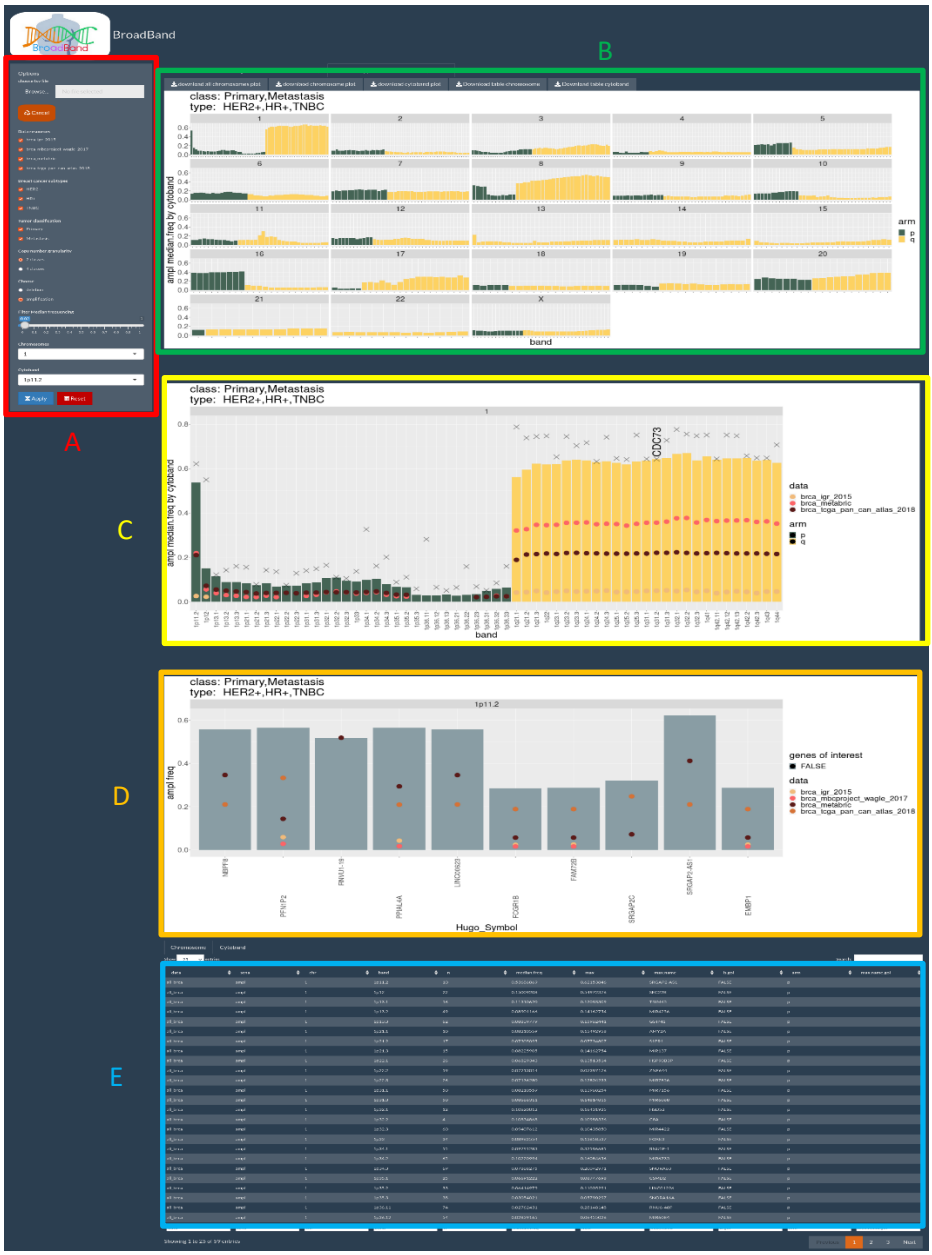

Figure 3. The "Somatic copy number aberrations" panel. The figure shows the layout of the panel, named "Somatic copy number aberrations", which reports the frequency of aberration computed in genomic regions of different ranges. The panel is divided into 3 main sections: the widgets (A) allow the user to make a custom selection of data of interest, the area of the plots (B, C, D) where selected data are summarized and visually represented as bar charts or as a table (E).

The widget panel (**Figure 3A**) is populated with multiple selection options which will update the

BRCA data displayed by default. Specifically, the widget panel allows the user to:

1. Upload a new BRCA dataset to be inspected in addition to those ones available by default. The *cancel* button will reverse the operation by removing the uploaded dataset among available data.
2. Select BRCA cohorts of interest. If the user uploaded a custom dataset, it will be listed here.
3. Select data based on BRCA subtypes.
4. Select data based on tumor classification.
5. Set granularity of SCNAs event. The user can select to compute and report the frequencies of aberrations based on 2 classes of aberrations: amplifications and deletion events or, alternatively, on 4 classes: homozygous deletion, hemizygous deletion, gain, and amplification events.
6. Select the SCNAs event of interest to be displayed.
7. Apply a threshold on aberration frequencies by using the provided slider.
8. Select the chromosome of interest to be displayed.
9. Select the cytoband of interest to be displayed.
10. Reset all settings to default.
11. Apply selected options to data and update plots and tables accordingly.

In addition to the widgets bar, the page layout also includes two plots, a table and five downloads buttons which allow the user to save all outputs displayed. The plots and the tables are the results of all options selected by the user. The first plot is a set of bar charts, one for each chromosome, to provide a landscape of SCNAs frequencies of aberrations. In the second plot the user can select a chromosome of interest to get a closer look at the frequencies of aberration across the chromosome cytobands. The third plot reports, within a selected cytoband of interest, the frequency of aberrations at the gene level. A gene is highlighted (**Figure 3D**) if it is a known cancer driver gene. Genomic location information (**Figure 3B, 3C**) and BRCA dataset used to compute frequencies of aberrations displayed (**Figure 3C, 3D**) are reported in the plot legend.

5 Conclusion and Discussion

The result of this project is the BroadBand application, a functional and interactive web-based resource to summarize and visualize genomic BRCA datasets, stratify available data based on clinical and molecular information, and achieve higher precision in the assessment of BRCA subtypes genomic features.

It provides an extensive collection of information from BRCA studies reporting SCNAs and SNVs data; at first it provides a complete overview about BRCA cohorts composition, such as number of samples and type of mutations available, stratified by molecular cancer subtypes and tumor classification (**Figure 1**). Moreover, detailed insights can be obtained by inspecting the two somatic aberrations panels (**Figure 2, 3**), where the user can customize its search by exploiting multiple settings and filters.

BroadBand is a comprehensive and useful tool, representing a valuable resource to explore data and filter those of interest. Indeed, raw data available from cBioPortal and other user-defined resources, are processed to get a more accurate and comprehensive genomic profile by focusing not just at the gene level but also at wider genomic regions. This feature, coupled with the possibility to stratify data based on different BRCA-specific molecular features, can be helpful to describe this disease better. It is possible to see how the application works at https://bcglab.cibio.unitn.it/broadband.

5.1 Future Perspectives

Currently, BroadBand is complete and functional; however, there is plenty of room for further improvements. For example, clinicians' feedback might be crucial to develop new features or settings to facilitate its usage and results interpretation in the clinical environment.

Moreover, I am planning to a) increase the number of data resources used by default (currently N = 5) to provide more detailed and robust results, b) add clinical information, such as patients' overall survival and exposure to therapy or medical treatments, to link the genomic background to resulting phenotypes, and c) improve data stratification when available for multiple molecular features (i.e., the molecular subtype classified as HR+ can be divided into the two distinct subtypes: Progesterone Receptor and Estrogen Receptor).

BroadBand has been developed by focusing exclusively on BRCA datasets. As future perspectives, the current version could be adapted and extended towards a more pan-cancer web-based resource allowing the exploration of multiple cancer types.

Bibliography

1. Nowell PC. The Clonal Evolution of Tumor Cell Populations. *Science*. 1976;194(4260):23-28. doi:10.1126/science.959840

2. Greaves M, Maley CC. Clonal evolution in cancer. *Nature*. 2012;481(7381):306-313. doi:10.1038/nature10762

3. Gerlinger M, McGranahan N, Dewhurst SM, Burrell RA, Tomlinson I, Swanton C. Cancer: evolution within a lifetime. *Annu Rev Genet*. 2014;48:215-236. doi:10.1146/annurev-genet-120213-092314

4. Bardelli A, Cahill DP, Lederer G, et al. Carcinogen-specific induction of genetic instability. *Proc Natl Acad Sci U S A*. 2001;98(10):5770-5775. doi:10.1073/pnas.081082898

5. Cahill DP, Kinzler KW, Vogelstein B, Lengauer C. Genetic instability and darwinian selection in tumours. *Trends Cell Biol*. 1999;9(12):M57-60.

6. Prandi D, Baca SC, Romanel A, et al. Unraveling the clonal hierarchy of somatic genomic aberrations. *Genome Biol*. 2014;15(8):439. doi:10.1186/s13059-014-0439-6

7. de Bruin EC, McGranahan N, Mitter R, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*. 2014;346(6206):251-256. doi:10.1126/science.1253462

8. Yates LR, Gerstung M, Knappskog S, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*. 2015;21(7):751-759. doi:10.1038/nm.3886

9. Uchi R, Takahashi Y, Niida A, et al. Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *PLOS Genet*. 2016;12(2):e1005778. doi:10.1371/journal.pgen.1005778

10. Harbst K, Lauss M, Cirenajwis H, et al. Multiregion Whole-Exome Sequencing Uncovers the Genetic Evolution and Mutational Heterogeneity of Early-Stage Metastatic Melanoma. *Cancer Res*. 2016;76(16):4765-4774. doi:10.1158/0008-5472.CAN-15-3476

11. Jj H, Dc L, Hq D, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet*. 2016;48(12). doi:10.1038/ng.3683

12. Bashashati A, Ha G, Tone A, et al. Distinct evolutionary trajectories of primary high-grade serous ovarian cancers revealed through spatial mutational profiling. *J Pathol*. 2013;231(1):21-34. doi:10.1002/path.4230

13. Murugaesu N, Wilson GA, Birkbak NJ, et al. Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discov*. 2015;5(8):821-831. doi:10.1158/2159-8290.CD-15-0412

14. Cairns J. Mutation selection and the natural history of cancer. *Nature*. 1975;255(5505):197-200. doi:10.1038/255197a0

15. Welch JS, Ley TJ, Link DC, et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*. 2012;150(2):264-278. doi:10.1016/j.cell.2012.06.023

16. Nathanson DA, Gini B, Mottahedeh J, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science*. 2014;343(6166):72-76. doi:10.1126/science.1241328

17. Johnson BE, Mazor T, Hong C, et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science*. 2014;343(6167):189-193. doi:10.1126/science.1239947

18. Notta F, Chan-Seng-Yue M, Lemire M, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature*. 2016;538(7625):378-382. doi:10.1038/nature19823

19. McPherson A, Roth A, Laks E, et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat Genet*. 2016;48(7):758-767. doi:10.1038/ng.3573

20. Kreso A, O'Brien CA, van Galen P, et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science*. 2013;339(6119):543-548. doi:10.1126/science.1227670

21. Turajlic S, Swanton C. Metastasis as an evolutionary process. *Science*. 2016;352(6282):169-175. doi:10.1126/science.aaf2784

22. ICGC Prostate UK Group, Gundem G, Van Loo P, et al. The evolutionary history of lethal metastatic prostate cancer. *Nature*. 2015;520(7547):353-357. doi:10.1038/nature14347

23. Liu W, Laitinen S, Khan S, et al. Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nat Med*. 2009;15(5):559-565. doi:10.1038/nm.1944

24. Hong MKH, Macintyre G, Wedge DC, et al. Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun*. 2015;6(1):6605. doi:10.1038/ncomms7605

25. Bray F, Ferlay J, Laversanne M, et al. Cancer Incidence in Five Continents: Inclusion criteria, highlights from Volume X and the global status of cancer registration. *Int J Cancer*. 2015;137(9):2060-2071. doi:10.1002/ijc.29670

26. Torre LA, Siegel RL, Ward EM, Jemal A. Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2016;25(1):16-27. doi:10.1158/1055-9965.EPI-15-0578

27. Ginsburg O, Bray F, Coleman MP, et al. The global burden of women's cancers: a grand challenge in global health. *Lancet Lond Engl*. 2017;389(10071):847-860. doi:10.1016/S0140-6736(16)31392-7

28. Allemani C, Weir HK, Carreira H, et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet Lond Engl*. 2015;385(9972):977-1010. doi:10.1016/S0140-6736(14)62038-9

29. Perou CM, Sørlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature*. 2000;406(6797):747-752. doi:10.1038/35021093

30. Harbeck N, Penault-Llorca F, Cortes J, et al. Breast cancer. *Nat Rev Dis Primer*. 2019;5(1):66. doi:10.1038/s41572-019-0111-2

31. Ren JX, Gong Y, Ling H, Hu X, Shao ZM. Racial/ethnic differences in the outcomes of patients with metastatic breast cancer: contributions of demographic, socioeconomic, tumor and metastatic characteristics. *Breast Cancer Res Treat*. 2019;173(1):225-237. doi:10.1007/s10549-018-4956-y

32. Kohler BA, Sherman RL, Howlader N, et al. Annual Report to the Nation on the Status of Cancer, 1975-2011, Featuring Incidence of Breast Cancer Subtypes by Race/Ethnicity, Poverty, and State. *JNCI J Natl Cancer Inst*. 2015;107(6). doi:10.1093/jnci/djv048

33. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. *Ann Oncol*. 2015;26(7):1291-1299. doi:10.1093/annonc/mdv022

34. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer

and 101,986 women without the disease. *Lancet Lond Engl*. 2001;358(9291):1389-1399. doi:10.1016/S0140-6736(01)06524-2

35. Ellis MJ, Ding L, Shen D, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature*. 2012;486(7403):353-360. doi:10.1038/nature11143

36. Lopez-Garcia MA, Geyer FC, Lacroix-Triki M, Marchió C, Reis-Filho JS. Breast cancer precursors revisited: molecular features and progression pathways. *Histopathology*. 2010;57(2):171-192. doi:10.1111/j.1365-2559.2010.03568.x

37. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016;534(7605):47-54. doi:10.1038/nature17676

38. Heitzer E, Haque IS, Roberts CES, Speicher MR. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nat Rev Genet*. 2019;20(2):71-88. doi:10.1038/s41576-018-0071-5

39. Ediriweera MK, Tennekoon KH, Samarakoon SR. Emerging role of histone deacetylase inhibitors as anti-breast-cancer agents. *Drug Discov Today*. 2019;24(3):685-702. doi:10.1016/j.drudis.2019.02.003

40. Zhou Y, Wang Y, Zhang K, Zhu J, Ning Z. Reverse effect of chidamide on endocrine resistance in estrogen receptor-positive breast cancer. *J Shenzhen Univ Sci Eng*. 2018;35(4):339. doi:10.3724/SP.J.1249.2018.04339

41. Munster PN, Thurn KT, Thomas S, et al. A phase II study of the histone deacetylase inhibitor vorinostat combined with tamoxifen for the treatment of patients with hormone therapy-resistant breast cancer. *Br J Cancer*. 2011;104(12):1828-1835. doi:10.1038/bjc.2011.156

42. Durak MG, Canda T, Yilmaz B, et al. Prognostic Importance of Tumor Deposits in the Ipsilateral Axillary Region of Breast Cancer Patients. *Pathol Oncol Res POR*. 2019;25(2):577-583. doi:10.1007/s12253-018-0515-4

43. Yates LR, Knappskog S, Wedge D, et al. Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*. 2017;32(2):169-184.e7. doi:10.1016/j.ccell.2017.07.005

44. Aurilio G, Disalvatore D, Pruneri G, et al. A meta-analysis of oestrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 discordance between primary breast cancer and metastases. *Eur J Cancer Oxf Engl 1990*. 2014;50(2):277-289. doi:10.1016/j.ejca.2013.10.004

45. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples | Nature Biotechnology. Accessed February 24, 2022. https://www.nature.com/articles/nbt.2514

46. Jones D, Raine KM, Davies H, et al. cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to Detect Somatic Single Nucleotide Variants in NGS Data. *Curr Protoc Bioinforma*. 2016;56:15.10.1-15.10.18. doi:10.1002/cpbi.20

47. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15(8):591-594. doi:10.1038/s41592-018-0051-x

48. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016;44(11):e108. doi:10.1093/nar/gkw227

49. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol*. 2016;17(1):178. doi:10.1186/s13059-016-1029-6

50. Kalatskaya I, Trinh QM, Spears M, McPherson JD, Bartlett JMS, Stein L. ISOWN: accurate somatic mutation identification in the absence of normal tissue controls. *Genome Med*. 2017;9(1):59. doi:10.1186/s13073-017-0446-9

51. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res*. 2017;45(Database issue):D777-D783. doi:10.1093/nar/gkw1121

52. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311. doi:10.1093/nar/29.1.308

53. Mutational heterogeneity in cancer and the search for new cancer-associated genes | Nature. Accessed February 24, 2022. https://www.nature.com/articles/nature12213

54. Martincorena I, Raine KM, Gerstung M, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*. 2017;171(5):1029-1041.e21. doi:10.1016/j.cell.2017.09.042

55. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

56. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695

57. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603

58. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res*. 2014;42(Database issue):D980-D985. doi:10.1093/nar/gkt1113

59. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol*. 2017;2017. doi:10.1200/PO.17.00011

60. Amarasinghe KC, Li J, Hunter SM, et al. Inferring copy number and genotype in tumour exome data. *BMC Genomics*. 2014;15(1):732. doi:10.1186/1471-2164-15-732

61. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinforma Oxf Engl*. 2012;28(3):423-425. doi:10.1093/bioinformatics/btr670

62. EXCAVATOR: detecting copy number variants from whole-exome sequencing data | Genome Biology | Full Text. Accessed February 24, 2022. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-10-r120

63. Sathirapongsasuti JF, Lee H, Horst BAJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinforma Oxf Engl*. 2011;27(19):2648-2654. doi:10.1093/bioinformatics/btr462

64. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics*. 2014;15:244. doi:10.1186/1471-2164-15-244

65. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinforma Oxf Engl*. 2012;28(10):1307-1313. doi:10.1093/bioinformatics/bts146

66. Bao L, Pu M, Messer K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinforma Oxf Engl*. 2014;30(8):1056-1063.

doi:10.1093/bioinformatics/btt759

67. Zehir A, Benayed R, Shah RH, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med*. 2017;23(6):703-713. doi:10.1038/nm.4333

68. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res*. 2016;44(16):e131. doi:10.1093/nar/gkw520

69. Raine KM, Van Loo P, Wedge DC, et al. ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Curr Protoc Bioinforma*. 2016;56:15.9.1-15.9.17. doi:10.1002/cpbi.17

70. Xi R, Lee S, Xia Y, Kim TM, Park PJ. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res*. 2016;44(13):6274-6286. doi:10.1093/nar/gkw491

71. Kuilman T, Velds A, Kemper K, et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol*. 2015;16(1):49. doi:10.1186/s13059-015-0617-1

72. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLOS Comput Biol*. 2016;12(4):e1004873. doi:10.1371/journal.pcbi.1004873

73. Favero F, Joshi T, Marquard AM, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol*. 2015;26(1):64-70. doi:10.1093/annonc/mdu479

74. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12(4):R41. doi:10.1186/gb-2011-12-4-r41