Computational Investigation of Cancer Genomes

Amjad Alkodsi

Research Program in Systems Oncology
Research Program Unit
Faculty of Medicine
University of Helsinki
Finland

Academic dissertation

To be publicly discussed, with the permission of
the Faculty of Medicine of the University of Helsinki,
in Biomedicum Helsinki 1, Lecture Hall 3, Haartmaninkatu 8, Helsinki,
on 29 May 2019, at 12 o'clock noon.

Helsinki 2019

ILS
DOCTORAL PROGRAM
INTEGRATIVE
LIFE SCIENCE

Supervisor

Sampsa Hautaniemi, DTech, Professor
Research Program in Systems Oncology
Research Program Unit
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Reviewers appointed by the Faculty

Merja Heinäniemi, PhD, Docent
School of Medicine
University of Eastern Finland
Kuopio, Finland

Laura Elo, PhD, Docent
Turku Centre for Biotechnology
University of Turku
Turku, Finland

Opponent appointed by the Faculty

Fran Supek, PhD, ICREA Research Professor
Institute for Research in Biomedicine (IRB Barcelona)
The Barcelona Institute of Science and Technology
Barcelona, Spain

The Faculty of Medicine uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

ISBN 978-951-51-5190-2 (paperback)
ISBN 978-951-51-5191-9 (PDF)
http://ethesis.helsinki.fi
Unigrafia Oy
Helsinki 2019

Abstract

Cancer is a leading cause of death worldwide, and its incidence is increasing due to modern lifestyle that prolonged human life. All cancers originate from a single cell that had acquired genetic aberrations enabling uncontrolled proliferation. Each cancer is unique in its aberrant genetic makeup, which defines, to large extent, its biology, aggressiveness, and vulnerabilities to different treatments. Furthermore, the genetic makeup of each cancer is heterogeneous among its constituent cancer cells, and dynamic with the ability to evolve in order to preserve the survival of cancer cells. Sequencing technologies are currently producing massive amounts of data that, with the help of specialized computational methods, can revolutionize our knowledge on cancer.

A key question in cancer research is how to personalize the treatment of cancer patients, so that each cancer is treated according to its molecular characteristics. The first study in this thesis takes a step in that direction through a proposed novel molecular classification system of diffuse large B-cell lymphoma (DLBCL), which is the most common hematological malignancy in adults. The suggested classification, derived from the integrative analysis of gene expression and DNA mutations, stratifies DLBCL into four groups with distinct biology, genetic landscapes, and clinical outcome. These subtypes could help identify patients at high risk who may benefit from an altered treatment plan.

Understanding the genomic evolution of cancer that transforms a typically curable primary tumor into an incurable drug-resistant metastasis is another aspect of cancer research under intensive investigation. The second study in this thesis investigates the spreading patterns of metastasis in breast cancer, which is the most common cancer in women. Using phylogenetic analysis of somatic mutations from longitudinal breast cancer samples, the metastasis routes were uncovered. The study revealed that breast cancer spreads either in parallel from primary tumor to multiple distant sites, or linearly from primary tumor to a distant site, and then from that to another. However, in all cases, axillary lymph nodes did not mediate the spreading to distant sites. This provided a genetic-based evidence on the redundancy of lymph node dissection in breast cancer management.

Towards a genetic-based diagnostics in cancer, the computational methods used to detect genetic aberrations need to be evaluated for their accuracy. The third study in this thesis performs a comparison of methods for detecting somatic copy number alterations from cancer samples. The study evaluated several commonly used methods for two different sequencing platforms using simulated and real cancer data. The results provided an overview of the weaknesses of the different methods that could be methodologically improved.

Altogether, this thesis gives an overview on the field of computational cancer genomics and presents three studies that exemplify the clinical relevance of computational research.

Contents

Abbreviations vi

Publications and author's contributions vii

1 Introduction 1

2 Review of the literature 3

2.1 Biology of cancer . . . 3

2.1.1 Cancer hallmarks . . . 3

2.1.2 Cancer initiation . . . 3

2.1.3 Intratumor heterogeneity and clonal evolution . . . 5

2.1.4 Inter-individual variability and precision medicine . . . 6

2.1.5 Metastasis . . . 6

2.2 Cancer genomics . . . 8

2.2.1 Genetic alterations in cancer . . . 8

2.2.2 Etiology of genetic alterations . . . 10

2.2.3 Cancer genes . . . 11

2.3 Lymphoma . . . 12

2.3.1 B-cell development and lymphomagenesis . . . 12

2.3.2 Somatic hypermutation . . . 14

2.3.3 Diffuse large B-cell lymphoma . . . 15

2.4 Breast cancer . . . 16

2.5 Analysis of cancer genome data . . . 18

2.5.1 High-throughput sequencing technologies . . . 18

2.5.2 Alignment to the reference genome . . . 18

2.5.3 Detection of somatic mutations . . . 19

2.5.4 Detection of somatic copy number alterations . . . 20

2.5.5 Detection and quantification of mutational signatures . . . 21

3 Aims of the study 23

4 Materials and methods 24

4.1 Biological sample material . . . 24

4.2 Processing of sequencing data (I – II) . . . 25

4.3 Mutational signatures analysis (I – II) . . . 25

4.4 Phylogenetic analysis (II) . . . 26

4.5 Identification of SHM subtypes in DLBCL (I) . . . 26

4.6 Benchmarking SCNA detection algorithms (III) . . . 27

4.7 Statistical analysis (I – III) . . . 27

5 Results 28

5.1 Exposure to SHM mutational processes delineate the cell of origin in DLBCL (I) . . . 28

5.2 Distinct transcriptional subtypes of DLBCL defined by hypermutated genes (I) 28

5.3 The SHM subtypes of DLBCL are associated with distinct genetic alterations (I) . . . 30

5.4 Linear and parallel spreading of breast cancer distant metastasis (II) . . . 31

5.5 No involvement of synchronous axillary lymph node metastasis in distant metastasis seeding (II) 31
5.6 Activity of mutational processes during breast cancer progression (II) . . . 32
5.7 Assessment of SCNA detection algorithms from deep sequencing data (III) 33

6 Discussion **35**

Acknowledgements **38**

Bibliography **40**

Abbreviations

ABC	Activated B-cell
AID	Activation-induced cytidine deaminase
BAF	B-allele frequency
CGCI	Cancer Genome Characterization Initiative
cnLOH	Copy-neutral loss of heterozygosity
COO	Cell of origin
DLBCL	Diffuse large B-cell lymphoma
DSB	Double strand break
ER	Estrogen receptor
FDR	False discovery rate
FFPE	Formalin-fixed paraffin-embedded
HER2	Human epidermal growth factor receptor 2
IPI	International prognostic index
GATK	Genome Analysis ToolKit
GCB	Germinal center B-cell
GEO	Gene Expression Omnibus
HR	Homologous recombination
HTS	High-throughput seqeuncing
ITH	Intratumor heterogeneity
LOH	Loss of heterozygosity
NHL	Non-Hodgkin lymphoma
NMF	Non-negative matrix factorization
PR	Progesterone receptor
R-CHOP	Rituximab, cyclophosphamide, doxorubicin, vincristine and prednison
RNA-Seq	RNA sequencing
SCNA	Somatic copy number alteration
SHM	Somatic hypermutation
SNP	Single nucleotide polymorphism
SV	Structural variation
TCGA	The Cancer Genome Atlas
TSG	Tumor suppressor gene
UV	Ultraviolet
VAF	Variant allele frequency
WGS	Whole-genome sequencing

Publications and author's contributions

Publication I **Amjad Alkodsi**, Alejandra Cervera, Kaiyang Zhang, Riku Louhimo, Leo Meriranta, Annika Pasanen, Suvi-Katri Leivonen, Harald Holte, Sirpa Leppä, Rainer Lehtonen, Sampsa Hautaniemi.
Distinct subtypes of diffuse large B-cell lymphoma defined by hypermutated genes
Submitted.

Publication II Ikram Ullah*, Govindasamy-Muralidharan Karthik*, **Amjad Alkodsi***, Una Kjällquist, Gustav Stålhammar, John Lövrot, Nelson-Fuentes Martinez, Jens Lagergren, Sampsa Hautaniemi, Johan Hartman#, Jonas Bergh#.
Evolutionary history of metastatic breast cancer reveals minimal seeding from axillary lymph nodes.
The Journal of Clinical Investigation, 2018, 128(4).

Publication III **Amjad Alkodsi**, Riku Louhimo, Sampsa Hautaniemi.
Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data.
Briefings in Bioinformatics, 2014, 16(2), 242-254.

* co-first authors
# co-senior authors

Author's contributions

Publication I Conceptualized the study, designed the methodology, processed sequencing data, performed bioinformatics and statistical analyses, made the figures and wrote the paper.

Publication II Processed sequencing data, performed variants, copy number alterations and mutational signatures analyses, contributed to study conceptualization, other bioinformatics analyses, making the figures and writing the paper.

Publication III Conceptualized the study, performed data analysis, made the figures and wrote the paper.

1 Introduction

Cancer incidence was estimated at 18 million new cases worldwide in 2018, and is projected to increase to 29.5 million in 2040 [1, 2]. Cancer is the second leading cause of death, and was responsible for about 9.6 million deaths worldwide in 2018 [1, 2]. Due to its increasing social and economic impact, cancer research has been continuously growing through the years, aiming at understanding cancer and exploring new effective treatment strategies.

Cancer is a genetic disease. Although the implication of the genome in cancer development was discovered in the late 19th and early 20th centuries [3, 4], it was not until 1982 when the first DNA point mutation in cancer genome was discovered [5]. This discovery marked the beginning of a quest for finding cancer genes. The completion of the first draft of the human genome sequence in 2001 [6], and the launch of The Cancer Genome Atlas (TCGA) project [7] were two major milestones in the high-throughput cancer genomics field. The TCGA project aimed at generating comprehensive maps of the genomic changes in 33 types of cancer producing data from more than 10,000 cancer patients.

Since the completion of the human genome, the costs of sequencing have been continuously decreasing, and the amounts of generated sequence data are rapidly growing. The storage, analysis and interpretation of the vast amounts of generated data became the new bottleneck in modern genome research [8]. With these challenges, the role of bioinformatics became increasingly important as a multidisciplinary field that uses computational and statistical approaches to derive insights from biological data. This thesis covers a background in cancer genomics and the computational approaches used in analyzing cancer genome data, and discusses three studies on computational cancer genomics.

The first study in this thesis used computational methods to classify diffuse large B-cell lymphoma (DLBCL) into four subtypes with distinct clinical outcome. DLBCL is an aggressive and the most common type of non-Hodgkin lymphoma (NHL). NHL incidence was estimated at about 509,000 cases, and accounted for 248,000 cancer-related deaths globally in 2018 [2]. DLBCL is curable in 60% of all cases [9, 10], and therefore stratifying DLBCL patients into groups with unique clinical outcomes is crucially important.

The second study investigated the genomic evolutionary path of breast cancer metastasis. Breast cancer is the most common malignancy in women accounting for more than two million new cases and 626,000 deaths worldwide in 2018 [2].

Metastatic breast cancer remains incurable, and is almost exclusively responsible for mortality in breast cancer patients [11]. Therefore, understanding the evolutionary origin of breast cancer metastasis is of high importance.

DNA sequencing experiments can inform on all kinds of genetic alterations in the genome, and genetic testing is increasingly becoming clinically relevant. However, in order for the derived information to be meaningful, the employed computational tools need to be accurate. The third study compared the performance of several computational methods for detecting somatic copy number alterations (SCNAs) from DNA sequencing data. SCNAs constitute an important class of genetic alterations affecting the dosage of up to thousands of genes simultaneously, and their accurate detection is a major step in analyzing cancer genomes.

Both Publication I and Publication II in this thesis investigate highly heterogeneous cancers. Both DLBCL and breast cancer are divided into different subtypes with significantly different outcome. The first study takes the approach of discerning unexplained heterogeneity in the primary tumor in order to enhance the therapeutic choices of the patients at risk, and thereby reducing recurrence rate. The second study aims at understanding the genomic evolution of recurrence (in the form of metastasis) in order to inform the therapeutic choices after metastasis. The analysis of genomic evolution in Publication II depends on the detected SCNAs, which are heavily studied in Publication III.

The thesis proceeds with a review of the literature on cancer genomics with particular focus on lymphoma and breast cancer. Additionally, some of the main computational approaches used in investigating cancer genomes are discussed and reviewed.

2 Review of the literature

2.1 Biology of cancer

2.1.1 Cancer hallmarks

Cancer encompasses a group of diseases characterized by uncontrolled proliferation of cells that can cross the normal tissue boundaries and metastasize to distant sites [12]. Human cancer cells can originate from most cell types in human body, and the cell of origin constitutes the main way of cancer classification. Malignant transformation of normal cells occurs by acquiring special capabilities that give selective advantage to cancer cells over non-cancerous ones. These special capabilities are called the hallmarks of cancer [13, 14]. Although all types of cancer share the same hallmarks, the mechanisms through which individual cancers acquire those hallmarks are highly heterogeneous [13, 14]. Furthermore, the chronological order of acquiring cancer hallmarks also varies among cancers [13, 14]. Figure 1 shows a list of cancer hallmarks observed in different cancers.

2.1.2 Cancer initiation

All cancers develop as a result of genetic alterations in the genomes of cancer cells [12]. Cancer capabilities of growth and invasion are acquired through a multi-step process that resembles Darwinian evolution among cell populations [15, 16]. Normal cells continuously acquire somatic mutations during mitotic cell divisions due to exogenous or endogenous exposure to various mutagens or mutational processes [17, 18]. The random acquisition of mutations is followed by selection. Cells carrying mutations that confer a survival and proliferation advantage are selected over their neighbour cells [19]. Cancer initiates as a result of clonal expansion of a single cell that have acquired a set of advantageous mutations sufficient for malignant transformation [12]. These mutations are called driver mutations.

The number of driver mutations required for malignant transformation varies among different cancers, and was estimated to be fewer than ten mutations for the majority of cancers (considering only base substitutions) [20]. Given the nature of ongoing mutagenesis in normal cells, the likelihood of reaching malignant transformation increases with age, which is the most prominent risk factor [21]. However, there are other factors that increase cancer risk. Environmental exposure to mutagens such

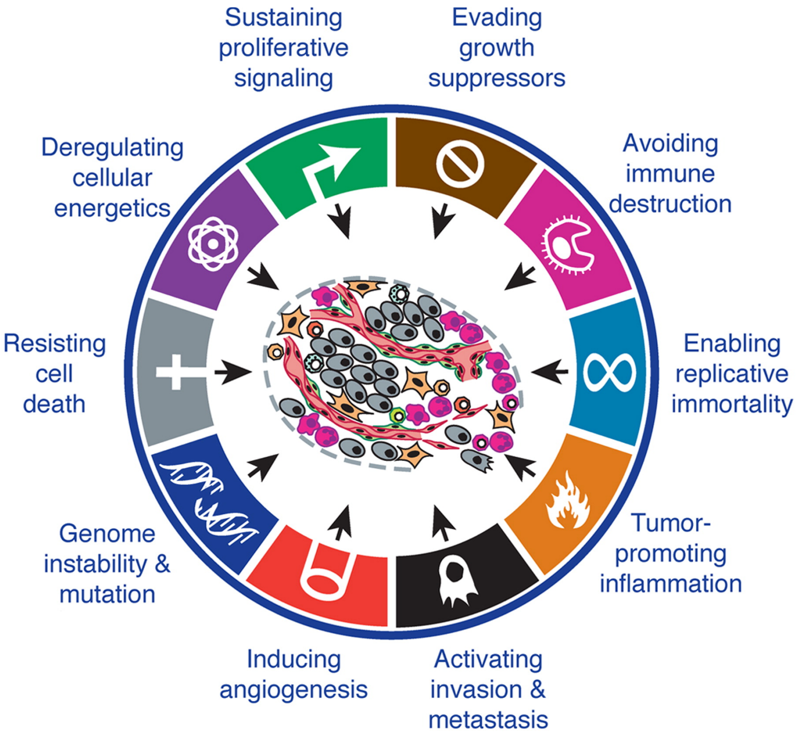

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1: The ten hallmarks of cancer. Edited from [14] with permission from Elsevier Inc.

as tobacco smoking and UV light increases the rate of mutagenesis in the exposed tissue contributing to higher cancer risk. Viral infection inserts completely new DNA sequences in the genome and contribute to the genesis of several cancer types [22]. Hereditary factors explain cancer incidence in a sizable fraction of cancers [23]. Inherited deleterious germline mutations can be drivers of cancer making their harboring cells one step ahead towards malignant transformation, or they can disrupt the DNA repair machinery accelerating endogenous mutagenesis.

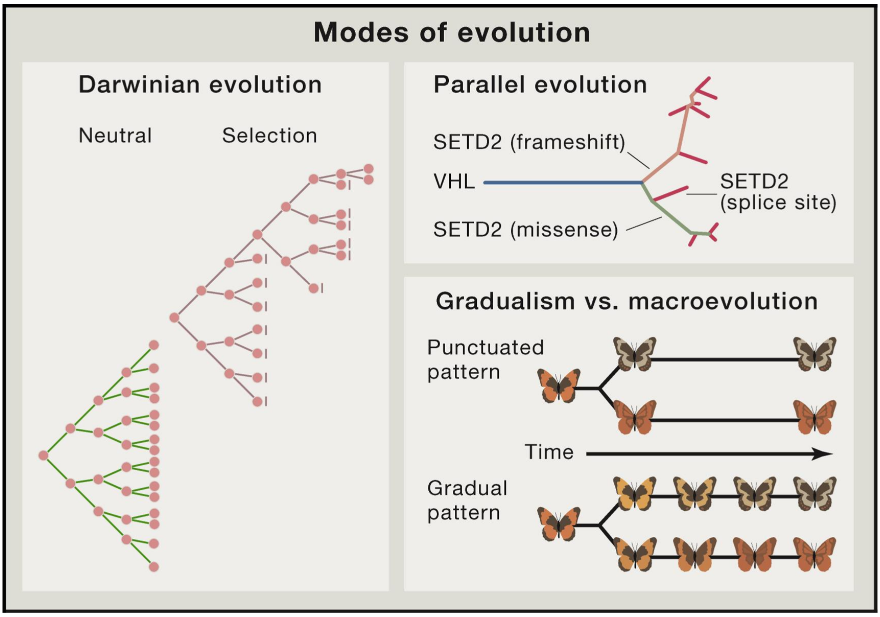


Figure 2: Illustration of several modes of evolution in cancer. Edited from [27] with permission from Elsevier Inc.

2.1.3 Intratumor heterogeneity and clonal evolution

Cancer cells continue to evolve after clonal expansion of the malignant progenitor cell by the same mechanism: mutagenesis and selection. However, evolution of malignant cells is further fueled by their genomic instability and by activating additional mutational processes that accelerate acquisition of mutations [24, 25]. The ongoing evolution contributes to the genetic and phenotypic diversification of cancer cells within the same tumor. The phenomenon of diversified cancer cells is called intratumor heterogeneity (ITH). Although genomic alterations are the major source of ITH, epigenetic mechanisms including DNA methylation, histone modifications and chromatin remodeling also contribute to ITH [26].

ITH manifests as spatial heterogeneity between genetically distinct cell populations, or temporal heterogeneity reflecting dynamic clonal composition over time. Temporal evolution is particularly apparent under selective pressure imposed, for example, by treatment [24]. Genetically distinct cancer cell populations can have differential response to treatment, and therefore, their survival and selective advantage may change after the treatment bottleneck. Changing microenvironment is another bottleneck that shakes the fitness of different cancer cell populations. For example, the growth advantage of different cancer cell populations can change after metastasis to distant sites with different microenvironment [28].

Several patterns of evolution govern the dynamics of clonal diversification in cancer (Figure 2) [29]. In linear evolution, subclones arise sequentially in a step-wise manner corresponding to higher clonal fitness at each step that may or may not lead to full clonal sweep of previous cell populations [30, 29]. Branching evolution describes the case when two or more subclones diverge from a common ancestor and continue to expand and co-exist in the tumor [29]. Neutral evolution assumes that there is no selection during most of the lifetime of the tumor leading to neutral growth of clones rather than clonal expansion [29, 27]. The patterns of evolution vary among individual cancers, and may operate simultaneously or at different stages of cancer progression within the same tumor [29, 27].

2.1.4 Inter-individual variability and precision medicine

The nature of cancer evolution leads to pronounced diversity among individual cancers such that each individual tumor is genetically unique. The genetic diversity reflects into unique aggressiveness and response to different treatments for each cancer. Precision medicine attempts to match the right treatment to the right patient based on the unique characteristics of each tumor [31]. The identification of prognostic and predictive biomarkers promotes personalized treatment of cancer. Prognostic biomarkers inform on the aggressiveness of the disease and evaluate the overall patient outcome after standard treatment, whereas predictive biomarkers predict the response to a specific cancer treatment [32]. Molecular classification of cancer divides a certain type of cancer into biologically distinct subgroups that may have unique prognosis and response to treatment [33, 34].

2.1.5 Metastasis

The vast majority of cancer-related deaths is due to metastasis [35]. Metastasis occurs as a multi-stage process that sequentially involves local invasion, intravasation, survival in the circulation, arrest at a distant organ site, extravasation, micrometastasis formation and metastatic colonization [36]. The genetic makeup of metastasis differs from that of its respective primary tumor, which further complicates the treatment of metastatic cancers. Metastatic dissemination can be described by two models: the linear progression model and the parallel progression

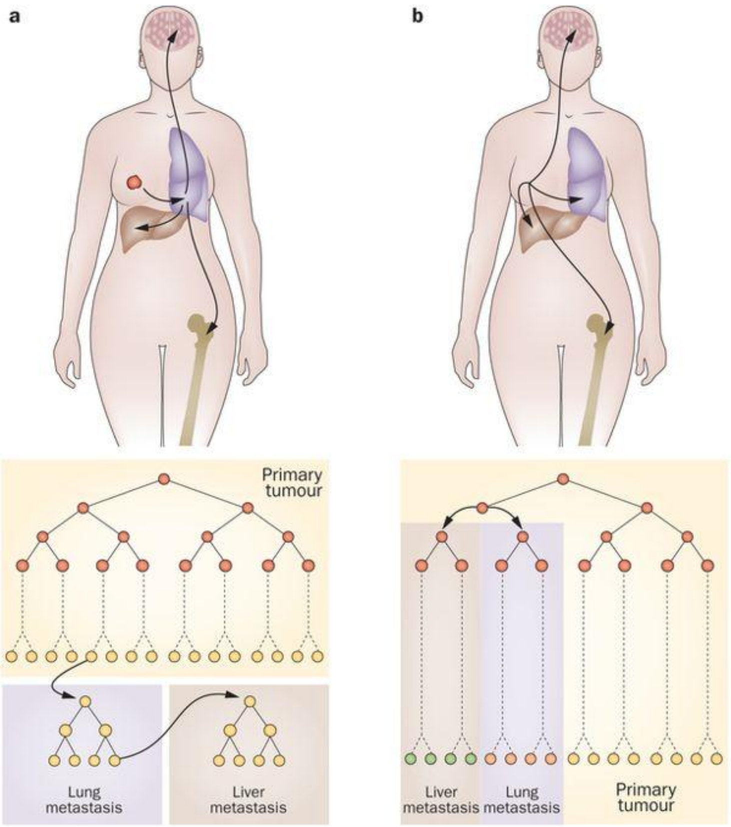


Figure 3: Illustration of the linear progression model (a) and parallel progression model (b) of breast cancer metastasis, together with corresponding possible phylogenetic trees. Edited from [39] with permission from Springer Nature.

model [37, 38]. The linear progression model assumes that metastatic dissemination occurs late in tumorigenesis leading to a small degree of genetic divergence between metastasis and its respective primary tumor [37, 38]. In parallel progression, the metastatic clone leaves the primary tumor early and continues to evolve independently in parallel with the primary tumor leading to a substantial genetic divergence [37, 38].

Metastatic spreading to multiple different distant sites could also be described by linear or parallel spreading patterns [37, 39]. In linear spreading, metastases spread from one site to another sequentially accumulating further genetic alterations at each step. Alternatively, metastatic clones can spread directly from the primary tumor to different metastasis sites, and each evolves independently in parallel (Figure 3). Metastatic seeding could be polyclonal in which more than one clone seed the metastasis, in contrast to monoclonal seeding, where seeding occurs by only one clone [37]. The patterns of metastatic progression, spreading and seeding varies among different cancers and even within individual cancers [37].

2.2 Cancer genomics

2.2.1 Genetic alterations in cancer

Unlike inherited germline variations, somatic mutations arise in somatic cells and pass to the progeny of a mutated cell by mitotic cell divisions [12]. Somatic mutations include all kinds of genetic alterations in the DNA sequence ranging in size from a single nucleotide to a whole chromosome or even the whole genome [40]. Single nucleotide substitutions or point mutations change a single nucleotide to another, and can be either transitions (purine to purine or pyrimidine to pyrimidine) or transversions (purine to pyrimidine or vice versa). Small insertions or deletions (indels) affect few bases of DNA sequence. Somatic copy number alterations (SCNAs) include amplifications or deletions of a chromosomal segment that range in size from small focal alterations to arm-level chromosomal events. Structural variations (SVs) include insertions, deletions, inversions, tandem duplications, intra- and inter-chromosomal translocations. SVs may or may not be balanced in preserving the copy number after alterations. Similarly, loss of heterozygosity (LOH) may or may not change the copy number while leading to allelic loss of a genomic segment. Whole-genome doubling is a single event that leads to the duplication of a complete set of chromosomes [41]. Figure 4 shows several types of genetic alterations and their manifestation in sequencing reads.

Genetic alterations can have various effects depending on their type, size and location. Point mutations and indels in the exons of protein-coding genes may alter the amino acid sequence of a protein or lead to an immature stop codon [40]. These changes can cause a gain or loss of function for the disrupted protein. Small alterations in the non-coding regions of intragenic or intergenic regions may alter the splicing of a gene or disrupt DNA elements such as promoters and enhancers.

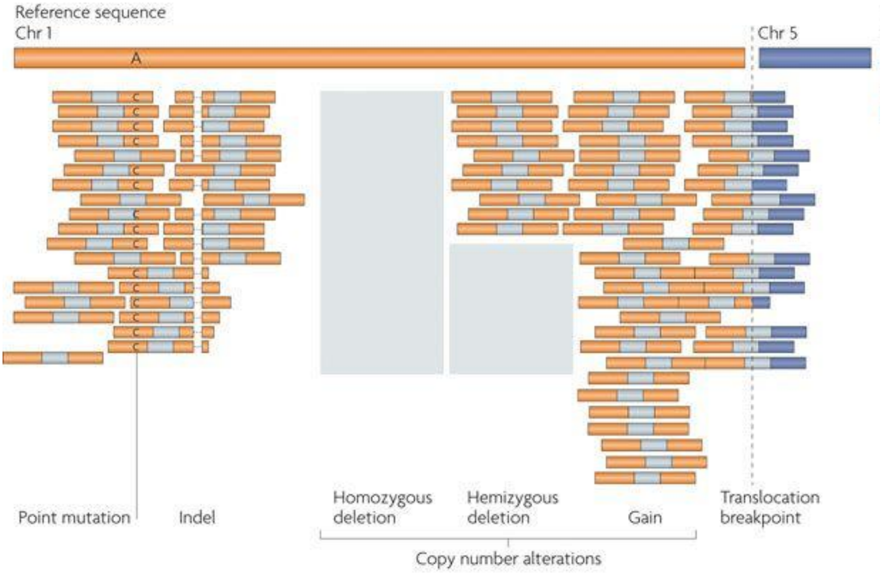


Figure 4: Illustration of the manifestation of several types of genetic alterations in sequencing reads. Edited from [42] with permission from Springer Nature.

SCNAs lead to changes in expression of their target genes that range from one in focal alterations to thousands in chromosomal alterations [43]. SVs can juxtapose parts of two genes, leading to gene fusion products with altered oncogenic functions.

From evolutionary perspective, accumulation of genetic alterations occurs gradually or as bursts of punctuated evolution [29]. In punctuated evolution, several alterations arise in one single event. SCNAs have been shown to follow punctuated evolution in triple-negative breast cancer [44]. Other punctuated evolution events include chromothripsis (from Greek for chromosome shattering), chromoplexy and kataegis (from Greek for thunderstorm). Chromothripsis is a single catastrophic event by which up to thousands of localized chromosomal rearrangements occur simultaneously [45]. Chromoplexy is a series of complex rearrangements occurring as closed chains without genomic loss [46]. Kataegis is a phenomenon of localized cytosine-to-thymine hypermutation attributed to activity of the APOBEC family of cytidine deaminases [47].

2.2.2 Etiology of genetic alterations

Genetic alterations are caused by various exogenous and endogenous factors. Exposure to environmental mutagens such as radiation, tobacco smoking, and ultraviolet (UV) light can cause DNA damage, and ultimately leads to mutations. Mutagenesis can also arise due to endogenous factors. DNA is replicated during each cell division by DNA polymerases. Replication is estimated to generate a constant rate of mutations, especially through increased transient exposure of single-strand DNA to spontaneous deamination of methylated cytosines that can ultimately lead to C>T substitutions at CpG context [48]. C>T transitions in CpG sites constitute the most prevalent mutational pattern in cancer, and their rate significantly correlate with age in many cancer types [49]. Single-strand DNA is also a substrate for enzymatic editing by the family of cytidine deaminases such as APOBEC enzymes that constitute the second most prevalent source of mutations in cancer [48, 49]. The increased transient exposure of single-strand DNA to mutagenesis is also apparent during transcription within the transcription bubble formed by RNA polymerase [50]. Transcription-coupled mutagenesis typically exhibit transcriptional strand asymmetry since the DNA repair machinery is more likely to be invoked by a damage that stalls the RNA polymerase on the transcribed strand [51].

DNA is continuously exposed to endogenous DNA damage sources such as hydrolysis and oxidation [52]. DNA damage can also occur due to an oncogene-induced replication stress that ultimately leads to double-strand breaks (DSBs), the most dangerous form of DNA damage [53]. DSBs and chromosomal instability have been also attributed to other mechanisms such as telomere attrition and centrosome abnormalities [54, 55]. It has been estimated that each human cell encounters approximately 70,000 lesions per day [48]. Luckily, the vast majority of the encountered damage is counteracted by effective DNA repair mechanisms, or other protective mechanisms that lead to cell-cycle arrest or apoptosis [56].

Different kinds of DNA damage are repaired by distinct repair mechanisms. Defects in DNA repair mechanisms through inherited germline or somatic mutations can lead to increased mutational load, cancer and other diseases. The type of genetic alterations induced by DNA repair deficiencies depends on the defective mechanism and the type of defect. For example, defective homologous recombination (HR) repair leads to a distinct signature of mutations and structural variations depending on the type of defect in the pathway [57]. Other unique mutational signatures have been attributed to deficiencies in mismatch repair and nucleotide excision repair pathways [58].

Repair of DNA damage can be error-prone introducing genetic alterations. For example, DSBs can be repaired by either the high-fidelity homologous recombination (HR) pathway or the error-prone non-homolougous end joining that leaves rearrangements in the genome [59]. DNA polymerases are involved in DNA repair and recruited by different pathways [60]. The fidelity of repair also varies depending on the recruited polymerase. Error-prone polymerases allow the completion of replication and avoidance of replication fork collapse at the cost of introducing errors [58, 60]. The genetic alterations left in the genome by error-prone mechanisms typically have distinct signatures. For example, processing with the non-canonical non-homolougous end joining leaves rearrangements with microhomology at breakpoint junctions, and recruitment of the error-prone polymerase η leads to thymine mutations at $\underline{\text{T}}$W context (W is A or T) [58, 59].

2.2.3 Cancer genes

Cancer genes are those causally implicated in cancer development [61]. Driver mutations in cancer genes confer a selective advantage to the tumor. Therefore, the number or pattern of mutations in a gene that cannot be expected by chance is indicative of positive selection and a driver role of that gene [61]. Driver genes are distinguished from other passengers whose mutations do not promote growth advantage of cancer cells. Driver genes have been identified based on the frequency [62, 63], the positional clustering [64], and the predicted deleterious function [65] of their mutations. However, not all mutations in driver genes are driver mutations. In fact, identification of driver mutations is more challenging than finding driver genes [40]. Although protein coding genes constitute a very small fraction of the genome, the identified driver mutations in the non-coding genome are so far not common [66]. However, there are established examples where the non-coding genome is implicated in tumorigenesis such as *TERT* promoter mutations [67], translocations involving the immunoglobulin locus in B-cell lymphomas [68], and oncogene activation by enhancer hijacking [69].

Cancer genes can be classified into oncogenes or tumor suppressor genes (TSGs) [40]. Activation of oncogenes via activating mutations, copy number amplifications or gene-fusions promotes cancer growth. In contrast, loss-of-function mutations, deletions or promoter hypermethylation inactivate TSGs and promote cancer. Unlike oncogenes, TSGs were thought to be recessive; that is biallelic inactivation is needed to promote cancer [70]. This two-hit model of tumorigenesis have been reevaluated after observations of haploinsufficient and gain of function phenotypes of TSGs [71]. TSGs can be categorized into gatekeepers or caretakers. Gatekeepers directly inhibit tumor growth or promote tumor death, and their inactivation directly contributes to cancer initiation or progression [72]. In contrast, inactivation of caretakers leads to increased accumulation of genomic alterations indirectly promoting cancer development [72].

2.3 Lymphoma

Lymphoma encompasses a group of malignancies that arise from lymphocytes that are at various stages of development. Lymphomas are traditionally divided into Hodgkin's lymphoma and non-Hodgkin lymphoma (NHL). NHL accounts for about 90% of all lymphomas [73], and more likely affects B-cells than T-cells or natural killer cells, owing to the unique biology of B-cells. In this section, I review non-Hodgkin B-cell lymphomas with particular focus on diffuse large B-cell lymphoma (DLBCL), the most common type of NHL.

2.3.1 B-cell development and lymphomagenesis

B lymphocytes in mammals develop from hematopoietic precursor cells in the bone marrow [74]. Rearrangements of the immunoglobulin heavy chain and the immunoglobulin light chain gene segments occur in the bone marrow generating a B-cell repertoire capable of recognizing more than 5×10^{13} different antigens [74]. Immature B-cells that have survived positive selection for affinity and negative selection for autoreactivity leave the bone marrow and migrate to the spleen, where they differentiate into naive, follicular or marginal zone B cells [74]. Activation of B-cells occurs in the secondary lymphoid organs after antigen encounters, with or without the help of T-cells. Germinal centers are transient structures of fast proliferating B-cells that form within secondary lymphoid organs during T-cell dependent activation [75]. A process termed affinity maturation takes place in the germinal center [76]. Affinity maturation is a result of iterative rounds of somatic hypermutation (SHM) and affinity-based selection [76]. SHM randomly introduces nucleotide substitutions in the variable region of immunoglobulin genes, which acts as a fine-tuning step of the already rearranged immunoglobulin to enhance antigen affinity [76]. Class switching also occurs in the germinal center by a process called class switch recombination, where antibodies of classes M and D switch to classes G, A or E [77]. Ultimately, B-cells in the germinal center become long-lived

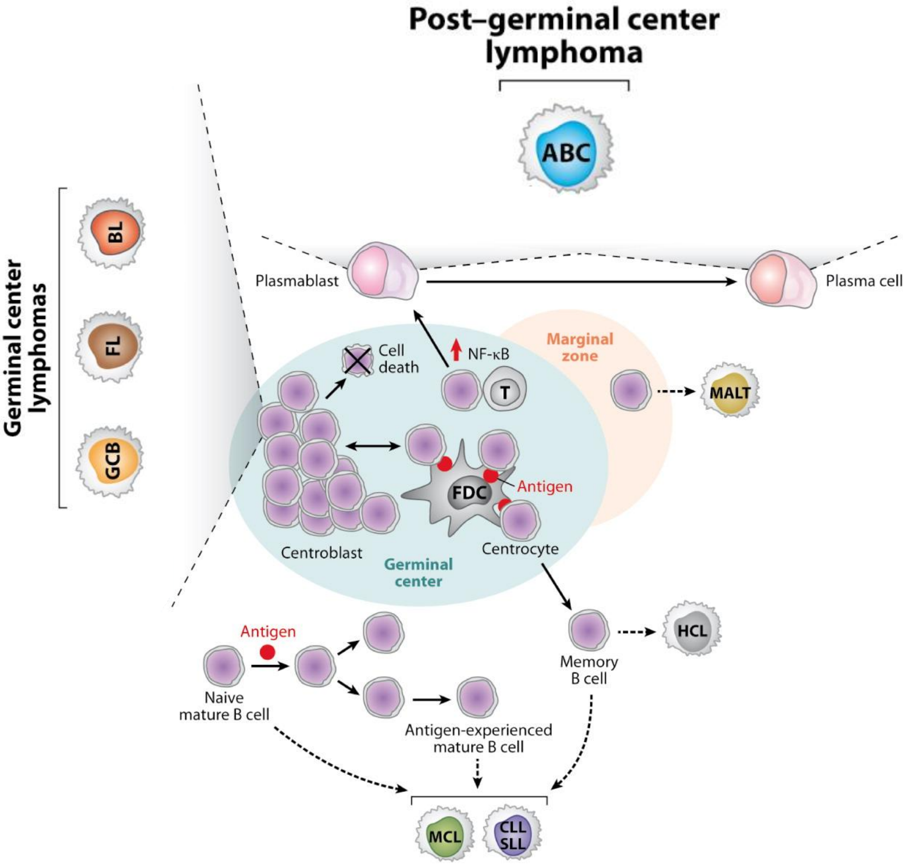


Figure 5: The cell of origin for various B-cell neoplasms. Edited from [78] with permission from Annual Reviews, Inc.

memory B-cells or plasma cells residing in secondary lymphoid organs or in the bone marrow [74].

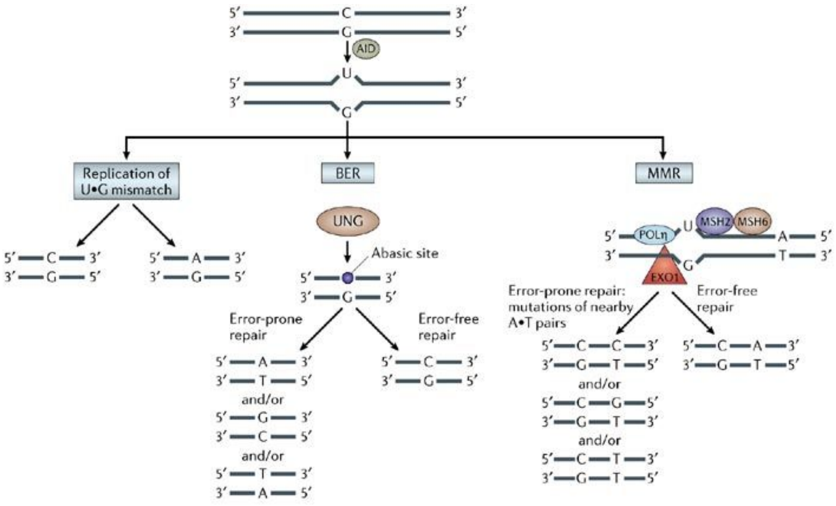
Lymphomagenesis can occur at various stages of B-cell differentiation resulting in different types of lymphomas with distinct pathological and clinical features (Figure 5) [78, 79]. Malignant B-cells bear a phenotypic resemblance in gene expression and immunoglobulin hypermutation to normal B-cells at a certain stage of differentiation [78]. This resemblance constitutes the basis for determining the cell of origin (COO) for different B-cell malignancies, and for subtype classification of histologically indistinguishable lymphomas [78]. For example, most mantle cell lymphomas show resemblance in gene expression profile to pre-germinal center B-cells, whereas most follicular lymphomas are similar to germinal center B-cells. About 95% of lymphomas originate from B-cells despite the similar frequency of B and T cells in the human body [80]. This is because the same mechanisms used by B-cells for antibody diversity likely underlie lymphomagenesis. The frequent chromosomal translocations in B-cell lymphomas are thought to be mediated by aberrant rearrangements of the immunoglobulin gene, SHM or class switch recombination [81]. Aberrant SHM is also responsible for oncogene activation by off-targeting non-immunoglobulin genes [82].

2.3.2 Somatic hypermutation

Somatic hypermutation (SHM) mediates affinity maturation by introducing point mutations into the variable region of the immunoglobulin genes in germinal center B-cells. SHM mutations have several properties: (1) they accumulate up to 1.5 - 2.0 kbp downstream of transcription start sites [83], (2) they require and correlate with transcription although transcription alone is not sufficient for SHM [84, 85], and (3) they are preferentially targeted to to cytosines within WR$\underline{C}$H motif or thymines within the $\underline{T}$W motif (W is A or T; R is A or G; H is A, C, or T) [86].

SHM is initiated by the activation-induced cytidine deaminase (AID), which deaminates cytosines creating uracil mismatches. Three possible pathways can repair the uracil mismatch: (1) proceeding with replication that produces C $>$ T transitions, (2) removal of uracils resulting in abasic sites, that either lead after replication to any of the four bases, or proceed with further processing by error-prone polymerases such as polymerase η, and (3) recognition by MSH2 and MSH6 mismatch repair enzymes and further processing by error-prone polymerases (Figure 6) [87, 88, 89]. Recruitment of error-prone polymerases may also introduce additional mutations in the vicinity of the site of deamination.

SHM mistargeting was seen in several B-cell lymphomas leading to mutations downstream of transcription start sites of several proto-oncogenes [82, 90, 91]. Off-target mutations by SHM are not random, but rather occur recurrently in specific regions in the genome [90]. The specificity of AID targeting is still not well understood. However, several features significantly associate with AID targets. Transcription levels are significantly higher in AID target genes than non-target genes [92]. The binding of RNAPolII and the stalling factor Spt5 is higher within AID mutational targets [93, 92]. Finally, AID mutations associate



Copyright © 2006 Nature Publishing Group
Nature Reviews | Immunology

Figure 6: Three possible pathways for the repair of uracil mismatches upon AID-mediated deamination. Reprinted from [89] with permission from Springer Nature.

with genomic regions of active enhancers and transcriptional elongation, the vicinity of superenhancers and regions of convergent transcription [92, 94, 95].

2.3.3 Diffuse large B-cell lymphoma

Diffuse large B-cell lymphoma (DLBCL) is the most common type of NHL accounting for 30 – 40% of all new NHL cases. Although DLBCL is an aggressive lymphoma, the current standard immunochemotherapy regimen consisting of the CD20 antibody rituximab in addition to cyclophosphamide, doxorubicin, vincristine and prednisone (R-CHOP) cures approximately 60% of patients [9, 10]. Patients who fail R-CHOP treatment present with a refractory or relapsed disease with dismal outcome [96]. The International Prognostic Index (IPI) is a clinical tool developed to predict the outcome in patients with DLBCL [97]. The IPI is based on five criteria that associate with poor prognosis: (1) age >60, (2) stage III/IV, (3) elevated serum lactate dehydrogenase level, (4) the patient performance status $\geq$2, and (5) more than one extranodal disease sites [97].

DLBCL is classified using gene expression profiling into germinal center B-cell-like (GCB), activated B-cell-like (ABC) and a third minor subtype that includes unclassified cases [98]. GCB DLBCLs are believed, based on similarity in gene expression profile, to originate from normal B-cells in the germinal center, whereas ABC DLBCLs are thought to originate from activated plasmablasts. This cell of origin (COO) classification is associated with survival after standard treatment, where ABC cases have significantly worse outcome [99, 100].

DLBCL is a genetically heterogeneous disease with a multitude of low frequency driver alterations [101, 102, 103, 104]. Several of the recurrent driver alterations are preferentially specific to either the GCB or ABC subtype, and some are associated with survival [105]. Translocations of *BCL2* and mutations in the chromatin remodeling pathway are characteristic of a fraction of the GCB subtype [79]. Constitutive NF-κB signaling and chronic active B-cell receptor signaling are hallmarks of a fraction of ABC DLBCLs [79]. Co-occurrence of genomic alterations in DLBCL was exploited to identify genetic subtypes with distinct oncogenic signaling and clinical outcome [106, 107].

Most relapse cases of DLBCL occur within the first three years following treatment [108]. The genetic landscape of relapsed DLBCL shows considerable differences to that of the primary tumor [109]. Clonal evolution of DLBCL relapse was shown to follow either late or early branching patterns reflecting various genetic similarity to the primary tumor [110, 111]. The outcome of patients with relapsed DLBCL is typically very poor, especially for those with early relapse [112]. Treatment of relapsed or refractory DLBCL may involve high-dose therapy and autologous stem cell transplantation, depending on patient age and performance status [96]. Patients who cannot tolerate autologous stem cell transplantation are usually treated for palliative purposes [96].

2.4 Breast cancer

Breast cancer is the most common malignancy and the first leading cause of cancer death in women worldwide [2]. Breast cancer is clinically heterogeneous, where early stage tumors are highly curable as opposed to metastatic cancers that remain incurable [113]. Clinical and pathological variables are highly informative for patient prognosis and treatment strategy selection. Lymph node metastasis, tumor size, grade, and proliferation rate are all established prognostic markers [11]. The estrogen receptor (ER) status, progesterone receptor (PR) status, and human epidermal growth factor receptor 2 (HER2) status are three essential biomarkers that guide treatment strategies in breast cancer [114].

Treatment of breast cancer involves one or more of the following strategies [113]. Surgical removal of the tumor by entire breast removal (mastectomy) or by a breast-conserving surgery (lumpectomy). Axillary lymph node dissection can be a part of the main surgery or as a separate operation. Radiation therapy may follow surgery in some cases of breast cancer. Hormone therapy can be used in cancers with positive hormone receptors. Chemotherapy is the main treatment for hormone receptor-negative cancers, and can be administered after surgery (adjuvant), before surgery (neoadjuvant) or both. HER2-positive cancers can be treated with trastuzumab, a targeted therapy that inhibits the HER2 protein [115]. A fraction of breast cancers with BRCAness phenotype are eligible for treatment with PARP inhibitors [116]. Despite of the plethora of clinical prognostic and predictive biomarkers in breast cancer, a considerable fraction of patients are considered over-treated [11].

Several gene expression profiling strategies were used to stratify breast cancer. Analysis of gene expression identified five intrinsic molecular subtypes associated with survival in breast cancer: luminal A, luminal B, basal, HER2-like and normal-like subtypes [117, 118]. The majority of luminal breast cancers are positive for estrogen receptor, and luminal B tumors have worse prognosis than luminal A [119]. Basal breast cancers are mainly triple-negative (ER-, PR-, HER2-) with bad prognosis [119, 120]. The intrinsic subtypes were also used to build a supervised risk predictor of breast cancer recurrence [121]. Two other gene expression signatures were developed to predict the risk of distant metastasis and were translated into diagnostic tests [122, 123].

Only three driver genes in breast cancer are mutated in more than 10% of patients: *PIK3CA*, *TP53* and *GATA3* [120]. However, a long tail of low-frequency drivers illustrate the high level of heterogeneity in breast cancer [57]. Breast cancer is in fact considered to belong to a class of cancers dominated by copy number alterations [124]. Several oncogenes are amplified in more than 20% of breast cancers including *MYC*, *ERBB2* (also known as HER2), *ZNF703* and *CCND1* [125]. Integrative clustering of both gene expression and copy number alterations identified 10 groups of patients with distinct survival association [125]. Predictive scores for chemotherapy sensitivity were developed using patterns of loss of heterozygosity and allelic imbalance in triple-negative breast cancers [126].

Breast cancers have variable extents of intra-tumor heterogeneity and subclonal driver mutations within the primary tumor [127]. Distant metastasis, as well as locoregional relapse, were found to be seeded by subclones in the primary tumor [127, 128]. However, distant metastases had a wider repertoire of driver genes than locoregional relapses and primary tumors [128, 129]. Treatment can have a dramatic influence on the clonal architecture of breast cancer, which can lead to drug resistance [130]. In particular, acquired mutations in estrogen receptor α encoded by the *ESR1* gene were found to drive resistance to hormonal therapy antagonizing estrogen receptor in ER+ cancers [131]. On the other hand, chemoresistance in triple-negative breast cancers was shown to occur via adaptive selection of pre-existing genotypes and acquired reprogramming of transcriptional profiles [132].

2.5 Analysis of cancer genome data

2.5.1 High-throughput sequencing technologies

High-throughput sequencing (HTS), also known as next generation sequencing, is a sequencing technology that allows the examination of billions of DNA templates in a single instrument run [133]. Several HTS instruments are available from different manufacturers such as Illumina, Oxford Nanopore, Pacific Biosciences and ThermoFisher Ion Torrent. These instruments vary in the technologies used, the number of reads produced each run and the length of sequencing reads [133]. Illumina instruments, used in the studies of this thesis, allow sequencing a very high number of reads per run but at a short read length. To compensate for the short length of the reads, Illumina sequencers support paired-end sequencing, in which both ends of the same DNA fragment is sequenced to the same read length. Paired-end sequencing enhances the alignment to the reference genome and also allows a better detection of structural variations [42]. Therefore, Illumina sequencing has been the most popular choice for resequencing experiments such as exome sequencing, whole-genome sequencing (WGS) and other read counting applications such as RNA sequencing (RNA-Seq) and chromatin immunoprecipitation sequencing. Exome sequencing targets the exons of protein coding genes providing higher sequencing coverage (the average genome-wide or exome-wide sequencing depth) at a reduced cost in comparison to WGS [134].

2.5.2 Alignment to the reference genome

Alignment involves finding the chromosomal locations from which the reads have most likely originated. Considering the massive number of reads produced by HTS experiments, the large size of the human genome, and the large genomic regions with repetitive sequences, alignment is the most computationally intensive task when analyzing cancer genomes. Alignment algorithms need not only to be computationally effective in mapping sequencing reads, but also accurate in doing so. The accuracy of the alignment has a crucial effect on the quality of detecting point mutations and structural variations from aligned sequencing data [42, 135]. The most commonly used aligners are the Burrows Wheeler Aligner (BWA) and Bowtie2 [136, 137]. Both BWA and Bowtie2 support fast gapped alignment with mismatches to account for possible variations in the genome. Other preprocessing steps typically follows alignment such as sorting reads by coordinate, and marking or removing duplicates (reads that have likely originated from the same original DNA fragments).

2.5.3 Detection of somatic mutations

Accurate detection of somatic mutations is crucial in the era of precision medicine. The statistical methods used for detection of somatic mutations, also called somatic variant callers, typically requires matched tumor-normal sample pairs that have been aligned to a reference genome. The matched normal sample, which can be derived from blood or adjacent normal tissue, is used to eliminate inherited germline variations. Somatic variant callers scan the entire genome, or exome, for mismatching basepairs or small insertions and deletions. Read count statistics in both the normal and tumor samples are collected together with various other informative metrics for each recorded mismatch. Finally, each recorded mismatch is either accepted as a somatic mutation or rejected as an artifact or germline variation based on a specified statistical model. The candidate variants could also undergo a set of hard filters defined by the user. This general workflow is followed by several variant callers such as MuTect, Varscan and Strelka [138, 139, 140].

Several biological and technical factors challenge the seemingly simple workflow of variant callers, leading to relatively low concordance between different analysis pipelines [135]. Intra-tumor heterogeneity (ITH), aneuploidy and tumor purity deviate the expected variant allele frequencies between different samples [141]. ITH and aneuploidy can also lead to deviations in variant allele frequencies between different regions in the genome of the same sample. Mutations with exceedingly low variant allele frequencies ultimately become indistinguishable from technical artifacts. Artifacts can arise as a result of sequencing errors, alignment errors or DNA damage during tissue fixation (formalin-fixed paraffin-embedded [FFPE] induced artifacts) or during library preparation (8-oxoguanine artifacts) [142, 143]. The accuracy of variant callers is largely dependent on local sequencing depth, which is the number of reads mapping to the local genomic region. Higher sequencing depth increases the power for detection of somatic mutations especially in cancer samples with low purity [144].

2.5.4 Detection of somatic copy number alterations

Somatic copy number alterations (SCNA) affect a large fraction of the genome disrupting up to thousands of genes [145, 146]. SCNAs are detected from sequencing data using depth of coverage methods. The number of copies for a chromosomal segment is proportional to the local sequencing depth for that segment. Therefore, the workflow for SCNA callers starts from computing the ratio of read counts (local depth of coverage) between the tumor and matched normal samples at non-overlapping windows, exons, or single nucleotide polymorphisms (SNPs). Normalization typically follows to account for inherent sequencing biases in the depth of coverage such as GC-content, mappability, library size or hybridization affinity (exome sequencing) biases. Next, chromosomal segments with similar values of the log-ratios are identified using a segmentation algorithm such as the circular binary segmentation algorithm [147]. Chromosomal segments with high and low log-ratios are then called amplifications and deletions, respectively.

Detection of copy-neutral loss of heterozygosity (cnLOH) requires the quantification of B-allele frequencies (BAFs). BAF represents the fraction of reads exhibiting the alternative allele in germline SNPs. The expected value for BAF is 0.5 for a heterozygous SNP in the normal sample. This value will deviate from 0.5 in the tumor sample at chromosomal regions with imbalanced amplification or deletion of a single allele. BAFs can be incorporated in the segmentation step or overlaid on an existing segmentation. Segments with neutral log-ratios but deviated average BAF values are called cnLOH.

To convert the log-ratios to the corresponding absolute number of copies, estimation of ploidy and purity of the sample is needed. Ploidy and purity can be estimated jointly from the segmented copy number profile, such that the converted absolute copy number for all chromosomal segments in the genome is the closest possible to integer values [148, 144]. Estimating the purity-adjusted absolute copy number profile is advantageous for interpretation and analysis of cancer genomes. First, defining thresholds for amplifications and deletions using purity-adjusted copy number leads to better sensitivity in samples with low tumor content. Second, detection of genome doubling events, which are common in cancer and associate with poor prognosis [41], becomes possible. Finally, the multiplicity (or cancer cell fraction) of somatic mutations can be computed from variant allele frequencies when purity and absolute copy number values are known [144]. Mutation multiplicities are especially useful in deciphering the evolutionary history and clonal ordering of cancer subpopulations [149].

2.5.5 Detection and quantification of mutational signatures

The catalog of somatic mutations in cancer genomes carries imprints of mutational processes that have been active throughout the cellular lineage of cancer cells [150]. It has been known even before the era of high-throughput sequencing that mutagens such as UV light and tobacco smoking induce unique spectra of mutations [151, 152]. With high throughput sequencing, examination of the full catalog of somatic mutations in cancer genomes became possible. Each mutational catalog is a representative record of a mixture of unique signatures of different mutational processes. The computational challenge is to isolate the spectra of the mutational signatures and quantify their exposures in crafting the observed mutational catalog. This resembles a blind source separation problem that involves the isolation and quantification of hidden signals from a set of mixtures of these signals [153]. Non-negative matrix factorization (NMF) is an algorithm for learning part-based representation of objects such as images or text [154]. NMF is the first and most commonly used algorithm for deciphering the signatures of mutational processes in cancer genomes [155, 49]. NMF corresponds to factorizing a matrix A into two non-negative matrices: A ~WH. The most commonly used approximation employs simple multiplicative updates [156, 157].

The spectrum of point mutations is often represented by the type of substitution (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, T:A>G:C) and the trinucleotide sequence context of each mutated base leading to 96 possible classes (6 substitution types × 16 possible trinucleotide contexts). A matrix of counts for each class of mutation in each cancer genome (or exome) is used as an input to the NMF algorithm. NMF decomposes the input matrix into two matrices under the non-negativity constraint. The first output matrix includes the mutational spectra for N signatures, and the second output matrix contains the exposure of each of the N signatures in each cancer genome (or exome). The number of signatures N is pre-specified to the NMF algorithm. A suitable number of signatures can be found by running the NMF analysis on multiple values for N and choosing the value that minimizes the reconstruction error and maximizes the cophenetic correlation (similarity between extracted signatures at different random runs) [155]. Several factors limit the accurate extraction of mutational signatures such as low number of mutations per sample, low number of samples and the high similarity between signatures [155]. To overcome the number of samples limitation, alternative approaches have been developed to estimate the exposure of reliably known signatures in each single sample [158].

3 Aims of the study

1. To enhance the molecular stratification of diffuse large B-cell lymphoma thereby guiding clinical decision making. (Publication I)
2. To characterize and understand the metastatic spreading patterns in breast cancer. (Publication II)
3. To benchmark and compare the performance of somatic copy number alteration detection methods from whole-genome and exome sequencing data. (Publication III)

4 Materials and methods

4.1 Biological sample material

Table 1 lists the cancer sample datasets used in each publication, the measurement technologies and sources of the data.

Publication	Samples	Technology	Source
Publication I	Matched primary-relapse sample pairs from seven patients	WGS, RNA-Seq	In house
	Primary DLBCL samples from 97 patients	WGS (39 samples), RNA-Seq (97 samples)	Cancer Genome Characterization Initiative (CGCI) (phs000532) [159]
	Primary DLBCL samples from 414 patients	Affymetrix gene expression microarrays	Gene Expression Omnibus (GEO) (GSE10846) [99]
	Primary DLBCL samples from 470 patients	Affymetrix gene expression microarrays	GEO (GSE31312) [160]
	Primary DLBCL samples from 521 patients	RNA-Seq (234 samples) and exome sequencing (521 samples)	Genomic Data Common [106]
	Primary DLBCL samples from 101 patients	Affymetrix gene expression microarrays	GEO (GSE98588) [107]
	Primary DLBCL samples from 383 patients	RNA-seq (383 samples) and WGS (153 samples)	Study in reference [161]
	Primary DLBCL samples from 604 patients	RNA-seq and exome sequencing	Study in reference [104]
Publication II	99 samples from 20 breast cancer patients. The samples included primary tumors, axillary lymph node metastasis and distant metastasis samples.	Exome sequencing	In house
Publication III	Primary breast cancer samples from 4 patients	WGS, Exome sequencing and SNP arrays	The Cancer Genome Atlas (TCGA) [120]

Table 1: Cancer sample materials used in Publications I-III

4.2 Processing of sequencing data (I – II)

Raw paired-end sequencing reads for WGS of DLBCL samples (Publication I) and exome sequencing of breast cancer samples (Publication II) were mapped to the human reference genome hg19 using BWA [136]. Following alignment, several preprocessing steps were performed including sorting by coordinates and marking duplicates using PICARD tools [162], base quality score recalibration and realignment around indels using the Genome Analysis ToolKit (GATK) [163]. Raw paired-end RNA-Seq reads in Publication I were processed using the Sepia pipeline [164] as the following. Trimmomatic [165] was used to trim Illumina adapter, when present, and low quality bases at either end, keeping only reads with at least 25 bp of length after trimming. Alignment to human reference genome was done using STAR 2-pass mapping with regenerated genome [166]. Expression quantification was done using eXpress [167]. All the processing workflows were built with the Anduril framework for scientific data analysis [168, 169].

MuTect/MuTect2 [138] was used to identify variants from WGS data (Publication I) and exome data (Publication II). In cases where multiple cancer samples per patient were sequenced, a forced calling approach was used. In forced calling, each detected variant in at least one sample undergoes counting reads supporting the reference and the detected variant allele in all other samples from the same patient. Specialized variant filtering criteria were applied on called variants to account for possible sources of artifacts in each study design. For example, to account for FFPE artifacts that manifest as low frequency C>T substitutions, sample-specific C>T substitutions were required to have a variant allele frequency of at least 0.15 in Publication II. The final variants were then annotated for functional prediction using Annovar [170]. SCNAs were detected using Ascat following the AscatNGS workflow [148, 171].

4.3 Mutational signatures analysis (I – II)

In Publication I, a clustered mutation signature approach was used in order to allow the isolation of mutational processes with localized mutagenesis [172]. Each mutation was annotated with the distance to the adjacent mutation that has the same substitution and strand orientation. Based on the distance, point mutations were classified into clustered (distance $<$ 1000 basepairs) or unclustered (distance $\geq$ 1000 basepairs). Incorporating the clustering classification expanded the classes of substitutions from 96 to 192. Clustered mutation signatures were then extracted using the NMF algorithm. In Publication II, NMF was used to extract mutational signatures using the standard 96 classes of substitutions. NMF was executed using the *NMF* and *SomaticSignatures* R packages [173, 174]. The number of signatures was decided by running NMF with several values of N (number of signatures) and choosing a suitable N based on cophenetic correlation and reconstruction error.

4.4 Phylogenetic analysis (II)

Somatic single nucleotide substitutions were used to construct phylogenetic trees using the Dollo parsimony [175], which has the main assumption that the same mutation can only be gained once in an evolutionary trajectory. The confidence of each constructed tree was estimated using 1000 boostraps. The constructed trees were further validated using an orthogonal method called LICHeE [176]. LICHeE uses variant allele frequencies (VAFs) to build the trees under the following constraints. First, a mutation shared by a group of samples cannot be a successor to a mutation shared by a smaller subset of the same group. Second, the VAF of a given mutation cannot be higher than the VAF of its predecessor. Last, the sum of VAF values of mutations disjointly present in distinct subclones cannot exceed the VAF value of a common predecessor mutation [176]. The linear and parallel progression modes were inferred from the topography of the trees.

4.5 Identification of SHM subtypes in DLBCL (I)

The first step was to identify the genes targeted by somatic hypermutation (SHM). Motivated by the detected clustered mutation signatures, two mutational patterns (mutations in R$\underline{\text{C}}$H or $\underline{\text{T}}$W context) associated with SHM were used separately in the analysis. We restricted the analysis to genomic regions up to 2500bp downstream of transcription start sites of protein coding genes. Our choice was based on previous literature on SHM targeting [82, 90, 89] and based on our exploratory analysis. A one-sided binomial test was used to identify genes whose target regions had an unexpectedly high number of R$\underline{\text{C}}$H or $\underline{\text{T}}$W mutations in comparison to the background mutation rate. Genes with p-values < 0.1 after adjustment for multiple hypotheses testing using the false discovery rate (FDR) method by Benjamini & Hochberg [177] were considered significant. In the second step, gene expression of the identified target genes in 97 primary DLBCL samples were clustered using consensus hierarchical clustering as implemented in the *ConsensusClusterPlus* R package [178, 179]. Cluster centroids were computed as the average Z-score normalized gene expression of samples belonging to each cluster. Classification of new samples in validation sets was done by assigning each sample to the cluster of the nearest centroid.

4.6 Benchmarking SCNA detection algorithms (III)

Comparison of 10 methods for SCNA detection was performed using simulated sequencing data and cancer samples from TCGA. The compared SCNA detection methods included BICseq [180], HMMcopy [181], CNAnorm [182], SegSeq [183], COPS [184] and CNAseg [185] for WGS data, ExomeCNV [186], VarScan2 [139] and ADTEx [187] for exome data, and ControlFreeC [188, 189] for both WGS and exome data. Simulated sequencing data used chromosome 22 as a template to which SCNAs with different types and sizes were introduced by RSVSim [190]. To create simulated sequencing reads from the altered chromosome 22, wgsim from the samtools package was used [191]. Simulated reads were mapped to the reference genome by BWA and Bowtie2 [136, 137]. SCNA calls from the different algorithms were compared with the ground truth created by RSVSim for evaluation. To evaluate the methods in cancer sequencing data, the concordance between SCNA calls for each algorithm and SNP array results was evaluated.

4.7 Statistical analysis (I – III)

Statistical analysis in all the three publications was done in R. Statistical tests were two-sided unless otherwise specified. Correction for multiple hypotheses testing was performed whenever needed. In Publication I, survival analysis was done using the *survival* R package and visualized using the *survminer* R package. Overall survival was defined from the date of diagnosis to last followup or death from any cause. In Publication III, sensitivity and specificity were used to evaluate SCNA methods in simulated data. Sensitivity was computed as as the proportion of the ground truth regions called correctly by an SCNA method. Specificity was calculated as the compliment to false positive rate, and was computed as the length of non-overlapping genomic regions between the ground truth and called SCNAs divided by the length of genomic regions not in the ground truth. The Jaccard index was used to measure the concordance with SNP arrays in cancer data.

5 Results

5.1 Exposure to SHM mutational processes delineate the cell of origin in DLBCL (I)

Using WGS data from 53 samples, we extracted four signatures of mutational processes in DLBCL. Two signatures were similar to the previously identified Signature 5 and Signature 17 (both with unknown etiology) [49]. The third signature (labeled Signature TW) was characterized by clustered mutations at $\underline{T}$W context (T is A or T) attributed to error-prone repair by polymerase η. The fourth signature (labeled Signature RCH) was characterized by clustered mutations at R$\underline{C}$H context (R is A or G; H is C, T or A) attributed to AID mediated mutagenesis [192]. Mutations at R$\underline{C}$H and $\underline{T}$W contexts are characteristic of somatic hypermutation (SHM) [86].

Examination of the exposure of mutational processes in DLBCL indicated a striking difference between the ABC and GCB subtypes. Contribution of Signature TW was significantly higher in the GCB subtype, whereas the contribution of Signature RCH was significantly higher in the ABC subtype. The high contrast in the activity of SHM-related mutational processes between the ABC and GCB subtypes was sufficient to correctly classify 30 out of 35 DLBCL samples (86%) with either of the two subtypes. The results were validated in an extended set of 153 DLBCL samples with WGS, where the exposure of SHM mutational processes determined the ABC/GCB subtype correctly in 71 out of 91 DLBCL samples (78%).

5.2 Distinct transcriptional subtypes of DLBCL defined by hypermutated genes (I)

We examined the rate of clustered $\underline{T}$W and R$\underline{C}$H mutations within a region of 2500 basepairs downstream of transcription start site of each gene, and identified 38 target genes of R$\underline{C}$H mutations and 16 target genes of $\underline{T}$W mutations. Since SHM mutational processes are distinctive of the cell of origin in DLBCL and given the transcriptional dependence of SHM, we hypothesized that the expression of SHM target genes may identify biologically meaningful patterns in DLBCL. To verify our hypothesis, we conducted consensus hierarchical clustering of 36 SHM target genes in 97 DLBCL samples with RNA sequencing data. Four clusters (denoted SHM subtypes [SHM1-4]) with distinct association to COO subtypes

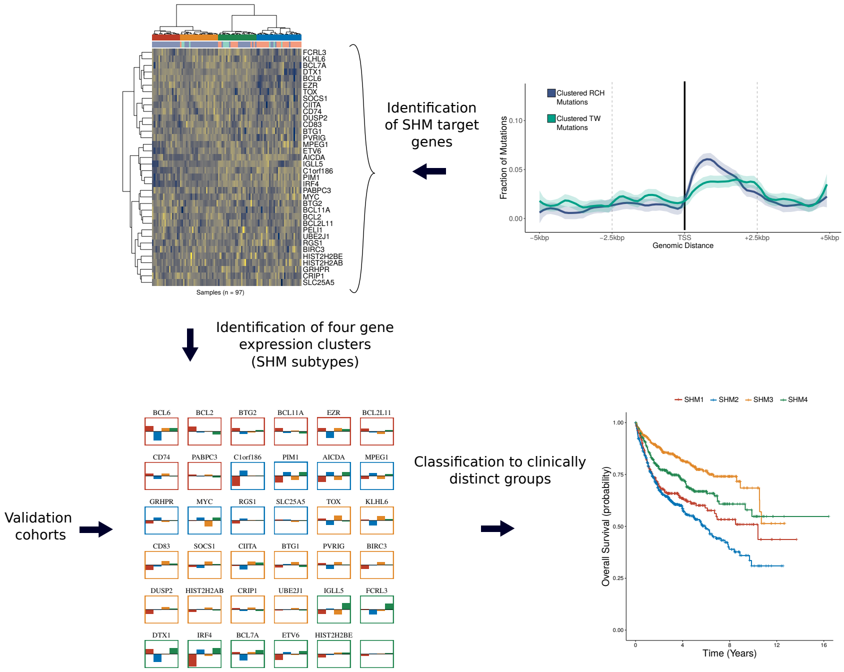


Figure 7: Schematic of the study in Publication I. Clustering of 36 SHM target genes identified four SHM subtypes used for classification of DLBCL samples. The figure is modified from Alkodsi et al., unpublished.

were identified. SHM1 and SHM3 were predominated with GCB cases, whereas SHM2 and SHM4 had a majority of ABC cases. Overall survival was different between the SHM subtypes, and especially between each two groups of the same COO subtype majority. In particular, among GCB enriched subtypes, SHM1 had a worse prognosis than SHM3, and among ABC enriched subtypes, SHM2 had a worse prognosis than SHM4. A schematic illustration of the discovery of the SHM subtypes is shown in Figure 7.

The SHM subtype assignment was performed by nearest centroid classification in the Lenz et al. CHOP and R-CHOP treated cohorts [99], Visco et al. [160], Schmitz et al. [106], Chapuy et al. [107], Reddy et al. [104], and Arthur et al. [161] cohorts. The proportions of ABC and GCB cases in each SHM subtype were similar in all tested cohorts. The SHM subtypes were significantly associated with overall survival in Lenz et al. cohorts (CHOP cohort: $p = 0.0006$; R-CHOP cohort: $p < 0.0001$), Visco et al. ($p = 0.0005$), Schmitz et al. ($p < 0.0001$), Reddy et al. (p

= 0.001), and Arthur et al. (p = 0.0059) cohorts. Progression-free survival was significantly different between the SHM subtypes in Visco et al. ($p < 0.0001$) and Schmitz et al. ($p < 0.0001$) cohorts. Time-to-progression differed significantly between the SHM subtypes in the Arthur et al. cohort (p = 0.0018).

Meta-analysis of DLBCL patients treated with immunochemotherapy (R-CHOP like treatment) indicated a significant association of the SHM subtypes with overall survival (1,642 patients from five cohorts; $p < 0.0001$) and progression-free survival (795 patients from three cohorts; $p < 0.0001$). The SHM subtypes remained significant in multivariate analyses of overall and progression-free survival using Cox proportional hazard regression model that included the IPI score and COO subtypes. In a COO subtype specific analysis, the SHM subtypes had a significantly different overall and progression free survival within the ABC and GCB subtypes. Surprisingly, overall survival was significantly different between the SHM subtypes in the obscure minor unclassified subtype of DLBCL. These results indicate that the SHM subtypes confer an additional prognostic value beyond the IPI scores and COO subtypes.

5.3 The SHM subtypes of DLBCL are associated with distinct genetic alterations (I)

To characterize the genomic landscape of the SHM subtypes, we used the genomic alterations data from Schmitz et al. (521 samples) and Reddy et al. (604 samples) cohorts. The alterations attributed to each of the SHM subtypes were highly similar between the two cohorts, indicating the robustness of the SHM subtype classification.

SHM1 (GCB majority with bad prognosis) was characterized by increased frequency of *BCL2* and *MYC* translocations that were not necessarily concurrent, alterations in the chromatin remodeling and histone modification pathways such as mutations in *KMT2D* and *EZH2*, alterations targeting the G-protein signaling pathway such as mutations in *GNA13* and *GNAI2*, and chromosomal duplication of chromosomes 7 and 12. SHM2 (ABC majority with bad prognosis) had a high frequency of *MYD88* (L265P) and *CD79B* mutations in addition to other alterations that characterize the ABC subtype such as *PIM1*, *ETV6*, *IRF4* mutations and *CDKN2A* deletions. Chromosomal duplication of chromosomes 3 and 18 were particularly common in SHM2. SHM3 (GCB majority with good prognosis) had a high frequency of alterations in the JAK-STAT pathway such as *SOCS1*, *STAT3* and *STAT6* mutations in addition to mutations in *SGK1*, *IRF8* and *TNFAIP3*. Finally, SHM4 had the highest frequency of *BCL6* fusions as well as *CD70*, *BCL10* and *SPEN* mutations. Collectively, each of the SHM subtypes had a distinct genetic profile that could potentially be targeted by novel therapies in DLBCL.

5.4 Linear and parallel spreading of breast cancer distant metastasis (II)

Among 20 studied breast cancer patients, five patients with exome sequencing of primary tumor and multiple distant metastasis sites were used to study the spreading of distant metastasis in breast cancer (patients 1, 4, 5, 8 and 19). Phylogenetic analysis of mutation data indicated that tumors from four patients (1, 5, 8 and 19) followed a linear progression model of successive metastasis-to-metastasis spreading [39]. In all the four cases, distant metastases from different anatomical sites (liver and lung metastases from patient 1, two bone relapses from patient 5, skin and bone metastases from patient 8, and two brain relapses from patient 19) were genetically more similar to each others than to their primary tumor. Interestingly, the seeding clone shared among different metastases harbored putative driver alterations that were not detected in the primary tumor.

In contrast, tumor in patient 4, from whom six regions from the primary tumor and three distant metastasis sites were exome sequenced, followed the parallel progression model [39]. Phylogenetic analysis indicated that the three distant metastasis sites (brain, colon and uterus) were seeded directly by the primary tumor (likely from different subclones in the primary tumor) and evolved independently in parallel further acquiring new unique genetic alterations. These results show that both of the debated progression models of metastasis dissemination (linear and parallel) occur in breast cancer, but in different cases. The prevalence of each model, however, remains unclear.

5.5 No involvement of synchronous axillary lymph node metastasis in distant metastasis seeding (II)

We investigated the role of axillary lymph node (ALN) metastasis in seeding distant metastasis in eight breast cancer patients from whom primary tumor, ALN metastasis and distant metastasis were exome sequenced (patients 2, 3, 8, 10, 14, 15, 17, and 18). The phylogenetic analysis of point mutations from these patients indicated that distant metastasis seeding from ALN metastasis was unlikely (maximum probability 0.23 in patient 8 while all other patients had a near zero probability). In three of these patients (patients 2, 10, and 1), the sequenced ALN represented the only positive ALN from these patients, excluding the possibility of distant metastasis seeded by unsequenced ALN. However, the possibility of distant metastasis seeded by an unsequenced region of the ALN cannot be excluded despite being unlikely to repeatedly be the case in all the eight patients.

5.6 Activity of mutational processes during breast cancer progression (II)

We extracted four signatures of mutational processes (labeled S1-4) that have been operational in the breast cancer cohort. Three of the four extracted signatures were mapped to previously identified signatures with known etiologies (S1, S2 and S4). Signature S1 was characterized mainly by C>T transitions in CpG context resulting from spontaneous deamination of 5-methyl-cytosines, and was associated with age at diagnosis [49]. Signature S2 had an excess of C>T and C>G mutations in the TpCpN context attributed to the activity of the APOBEC family of cytidine deaminases [193]. Signature S4 was mapped to the known Signature 3 that has been attributed to deficient homologous recombination (HR) in double-strand break repair.

We evaluated the exposure of the mutational processes in different categories including truncal, primary-specific, ALN-specific, local recurrence specific and distant metastasis-specific mutations. All the four signatures had a significantly different contribution between the different categories of mutations ($p < 0.01$; Kruskal-Wallis test). The contributions of signatures S1 (APOBEC), S3 (Unknown etiology) and S4 (HR deficiency) were significantly higher in mutations specific to distant metastasis in comparison to the ones specific to primary tumors ($p < 0.05$; Mann-Whitney U test with FDR correction). On individual level, 15 patients had an increased activity of one of the three (S1, S3 and S4) signatures in metastasis in comparison with the corresponding primary tumor ($p < 0.05$; Fisher's exact test with FDR correction). In patient 4, the significant increase in Signature S4 contribution accompanied two *BRCA2* mutation gained in two different distant metastasis sites. The increase of exposure of a specific mutational process did not associate with the molecular subtype of breast cancer or the treatment of the patient.

5.7 Assessment of SCNA detection algorithms from deep sequencing data (III)

The performance of ten SCNA detection algorithms was performed using simulated sequencing data as well as real breast cancer sequencing data. Various aspects of the algorithms' performance were evaluated using four simulation groups. Each group consisted of 10 simulated WGS and exome samples having a known set of SCNAs with specific characteristics.

The sensitivity of SCNA detection algorithms in WGS data (BICseq, HMMcopy, CNAnorm, SegSeq, COPS and ControlFreeC) was evaluated using the simulation group 1 that included various SCNAs of different classes of size ($<$ 1 Kbp, 1 – 10 Kbp, 10 – 100 Kbp, 100 kbp – 1 Mbp and $>$1 Mbp) and type (homozygous deletions, heterozygous deletions, gains of one copy, gains of two copies and high-level amplifications). SCNAs with small size were more challenging to detect as exemplified by the lower sensitivity of their detection by all algorithms. BICseq, SegSeq and COPS were specifically better at detecting smaller SCNAs than the other algorithms. The types of SCNAs did not have a large influence on the detection power, except for gains of one copy, which were more difficult to detect than other types.

The sensitivity of SCNA breakpoint localization was assessed also using the simulation group 1. The accuracy was determined as the distance in basepairs between the actual (ground truth) and the detected breakpoints. The breakpoints were considered in sensitivity assessment only when the overlap between the detected SCNA and the ground truth exceeded 70% of the size of both detected and ground truth SCNA. BICseq and SegSeq had the best accuracy for breakpoint localization, owing to their unique approaches that differ from the window approach utilized by the other algorithms. The performance of SCNA detection algorithm from exome data was similar between the algorithms (Control-FreeC, ExomeCNV, ADTEx and Varscan), since it depends on the inter-exonic distance rather than the methodology of the algorithms. As expected, the accuracy of WGS algorithms in breakpoint localization was significantly higher than exome algorithms.

The effect of complex copy number events (adjacent SCNAs sharing common breakpoints) on the performance of the algorithms was assessed in three simulation groups. Samples in simulation group 3 had the possibility of having adjacent SCNAs, in contrast to simulation group 2, in which adjacent SCNAs were not allowed. Samples in simulation group 4 had complex SCNA events spanning the whole chromosome. The presence of complex events did not significantly affect the performance of the algorithms as exemplified by the similar sensitivity and specificity of the different algorithms between simulation groups 2 and 3. However, the performance heavily deteriorated in simulation group 4 when the complex events spanned the whole chromosome. The observed drop in performance was due to the inconsistency of estimating the baseline ploidy between the different algorithms.

To evaluate the SCNA detection algorithms in real sequencing data, we used four breast cancer samples from TCGA characterized by WGS, exome sequencing and SNP arrays. We used SCNAs called from SNP arrays as a reference for evaluation. The Jaccard index was used to measure the concordance between SCNAs detected by the different algorithms in WGS or exome sequencing and those found by SNP arrays. The average concordance in WGS algorithms ranged between 0.645 (HMMcopy) and 0.451 (CNAnorm). In exome sequencing algorithms, the concordance with SNP arrays was significantly lower, ranging between 0.405 (ExomeCNV) and 0.235 (ADTEx).

6 Discussion

The three studies presented in this thesis provided various kinds of clinical, biological and technical insights in the field of cancer genomics. In Publication I, the proposed classification system informs on the pathogenesis and the biology of four unique disease entities in DLBCL, and may potentially have clinical applications. Publication II sheds light on the evolutionary origin of metastasis in breast cancer, and provides answers on previously debated progression models of the disease. Publication III gives a technical overview of the computational methods used for detection of somatic copy number alterations from cancer genomes. The findings of the three studies and their implications exemplify the wide range of insights that can be obtained by the computational analysis of cancer genomes.

The major implication of the first study is the ability to stratify DLBCL patients into biologically distinct groups that may differentially benefit from altered treatment options. Both SHM1 and SHM3 have a majority of cases with GCB cell of origin, but they significantly differ in their clinical outcome after R-CHOP standard therapy. The SHM1 group, which has a dismal outcome, may benefit from a modified treatment strategy unlike the SHM3 group, which has excellent outcome with standard treatment. The genetic landscape of the SHM1 group was characterized by alterations in *BCL2* and *MYC* that were not necessarily concurrent, in addition to mutations in histone-modifying genes such as *EZH2* and *KMT2D*. These alterations pinpoint the specific vulnerabilities in this subtype that can be targeted by novel treatments. The SHM2 and SHM4 subtypes include mainly ABC DLBCLs, but also significantly differ in their survival after R-CHOP treatment. SHM2 has the most dismal outcome among the four SHM subtypes, and therefore, has the highest priority for adjustment in the treatment. Targeting the B-cell receptor in the SHM2 subtype may be beneficial given the enrichment of genetic alteration in the B-cell receptor signaling pathway [194].

Publication I also sheds light on the mutational processes that have been active throughout the lifetime of lymphoma progenitor cells. Two mutational processes related to somatic hypermutation were particularly interesting, as their activity segregated with the cell of origin of DLBCL. A process attributed to error-prone repair by polymerase η was highly active in the GCB subtype, unlike the second process characterized by mutations attributed to AID mutagenesis, which was highly active in the ABC subtype. This highlights the distinct repair pathways utilized by each subtype upon cytosine deamination by AID, and motivates further functional studies on DNA repair in DLBCL. Additionally, the high contrast between the ABC and GCB subtypes suggests that the activity of mutational processes could be an important component in a classifier that determines the cell of origin in DLBCL using non-invasive genetic testing.

The major finding in Publication II was the lack of lymphatic seeding of distant metastasis in breast cancer, and thereby providing the first genetically-derived evidence of the redundancy of lymph node dissection in breast cancer. Lymph node involvement is undoubtedly an important prognostic factor that predicts worse outcome and an aggressive phenotype in breast cancer [195]. However, our results in Publication II support the notion that lymph node involvement may reflect the metastatic capability of the tumor, rather than mechanistically contributing to metastasis. Indeed several clinical trials have shown that lymph node dissection had little to no effect on patient survival [196, 197, 198], which further support our findings. A similar study investigating the lymphatic seeding of distant metastasis in colorectal cancer concluded that lymphatic and distant metastases arose independently from primary tumor in 65% of the cases, while in the rest, an evidence of lymphatic seeding was found [199]. Therefore, the prevalence of lymphatic seeding may vary between different types of cancers, and further studies in other cancers are needed to understand the factors that affect lymphatic seeding.

The results of Publication II showed that the spreading patterns of breast cancer vary in different cases. Linear progression or the successive spreading from a distant metastasis site to another was detected in several cases, and was often accompanied with acquisition of new putative driver mutations. Parallel progression from primary tumor to multiple distant sites was also found in one of the studied cases. The factors that influence the type of progression remain unclear. In both scenarios, the heterogeneous nature of distant metastases complicates and limits clinical genetic testing, as each biopsy could be distinct in the genetic makeup and targetable mutations. Liquid biopsies and the genetic screening of circulating tumor DNA provide a promising avenue for tackling the heterogeneity and predicting recurrence in breast cancer [200], as well as in other cancers [201].

The change in the landscape of active mutational processes in breast cancer metastasis was apparent in the majority of studied cases. A landscape characterized by aging signature in the primary tumor fades in metastasis giving rise to other mutational processes. The exposure to a mutational process attributed to activity of APOBEC family of cytidine deaminases showed a significant increase in distant metastasis. The increased activity of APOBEC-mediated mutagenesis upon recurrence was observed in several cancers [25], and linked to the acquisition of subclonal driver mutations [202]. Two treatment strategies were suggested to tackle APOBEC mutagenesis. The first is by increasing mutation rate to an unsustainable level of DNA damage that eventually leads to cell death [203], and the second is by decreasing mutation rate via the inhibition of APOBEC gene expression [204].

Both Publication I and Publication II highlighted the important role of hypermutation either in the form of AID mutagenesis in DLBCL or APOBEC mutagenesis in breast cancer. Both AID and APOBEC belong to the same family of cytidine deaminases inducing localized mutations in the genome. Localized mutations in breast cancer, known as kataegis, were linked to a gene expression signature that was associated with prognosis and Her2 status [205]. The patterns of clustered mutations were found to be widespread in cancer [172]. Investigation of the genetic-transcriptomic interplay of these mutational processes may lead to important discoveries in several types of cancer.

Publication III highlights the technical difficulties resulting from analyzing cancer genome data from different platforms. The computational methods that operate on whole-genome sequencing showed a superior performance in comparison to exome sequencing. Another remarkable obstacle was the difficulty of estimating the baseline level for copy number estimation in the cases where a large proportion of the genome is altered. However, since the time of publication, several recent methods for analyzing SCNAs from cancer genomes and exomes have tackled these difficulties. Several recent methods operating on exome sequencing successfully utilized panels of normal controls and/or the off-target aligned sequencing reads to eliminate hybridization bias often seen in exome sequencing [206, 207]. Estimation of ploidy and purity is currently a main analysis performed by several SCNA methods utilizing allele-specific information inherent in heterozygous SNPs [208, 209, 210]. Indeed, a choice of a method that is able to estimate ploidy and purity was necessary in Publication II. Benchmarking the performance of different bioinformatics tools is therefore a crucial task towards applied clinical genomics.

The last decade has witnessed an explosive growth of data collection in several fields and industries. Artificial intelligence and machine learning algorithms have already become a backbone in many data-driven industries. As biological and sequencing data accumulate, the performance of machine learning algorithms will surge, and more applications of artificial intelligence in health care, medicine and genomics will emerge [211]. The increasingly important role of bioinformatics and computational biology will continue the rise in a next era of cancer genomics.

Acknowledgements

This work was carried out in the Systems Biology of Drug Resistance in Cancer Laboratory at the Faculty of Medicine, University of Helsinki during 2013 – 2018. During this time, I had the privilege to work under the supervision of Professor Sampsa Hautaniemi. I thank Sampsa for the continuous support and guidance throughout these years, and for providing all the necessary resources to conduct research at the highest level. Sampsa leads the lab in a professional manner and allows individual research freedom, and for that I am deeply grateful.

My PhD work was generously funded by the Integrative Life Science (ILS) graduate program at the University of Helsinki and the Biomedicum Helsinki Foundation for whom I am sincerely grateful. I would like also to express my gratitude for the travel grants provided by the graduate programs at the university, which allowed me to attend several international conferences helping my scientific growth. I thank my thesis committee members Professor Satu Mustjoki and Professor Samuli Ripatti for their guidance and feedback in our annual meetings. Many thanks to Dr. Merja Heinäniemi and Dr. Laura Elo for the constructive review of this thesis.

The published research work in this thesis would not have been possible without the wonderful collaboration with many intelligent people. I would like to thank Professor Sirpa Leppä and her research group including Annika, Leo and Suvi-Katri for their collaboration, discussions and scientific inputs in the DLBCL project. I thank Professor Jonas Bergh, Dr. Johan Hartman and their team for the collaboration and for allowing me to be a part of their research projects. Special thanks to Ikram and Karthik for the tremendous efforts and collaboration in the breast cancer project. Several people in Sampsa's group are natural everyday collaborators. Many thanks to Riku, Alejandra, Kaiyang and Rainer for their help in various projects in the lab. Finally, I would like to acknowledge the fruitful collaboration with the research groups of Professor Lauri Aaltonen and Dr. Liisa Kauppi that produced several publications not part of this thesis.

The atmosphere at Sampsa's lab has been always friendly, positive and collaborative owing to the wonderful and professional people working there. I thank the current and past lab members with whom it was a privilege to work including Alejandra, Anna-Maria, Antti, Chengyu, Chiara, Emilia, Erkka, Inga-Maria, Ingrid, Jaana, Juha, Julia, Kari, Katherine, Kaiyang, Kristian, Lilli, Marko, Miikka, Oskari, Ping, Rainer, Riku, Sanaz, Sirkku, Tiia, Valeria, Veli-Matti, Ville and Yilin. Special thanks to Ville Rantanen for being the technical admin of the lab and

for providing unlimited technical help whenever needed, and to Tiia Pelkonen for all the administrative help and consultation.

Keeping the high spirit and motivation in a long PhD is not easy. Without the help of great friends, it would have been impossible. I sincerely thank many friends who stayed in touch despite being countries or even continents apart. I particularly thank Adnan and Wessam for their continuous reinforcement and for the nice times we spend each vacation.

My achievements are indebted to my family who have always supported me throughout my life. I thank my mom, sisters, brothers, nieces and nephews for being there.

Amjad Alkodsi
Helsinki, 2019

References

Page(s)

[1] Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* **68**, 394–424 (2018). 1

[2] Ferlay, J. *et al.* Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. `https://gco.iarc.fr/today` (2019). Accessed: January, 2019. 1, 16

[3] Hansemann, D. Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung. *Archiv für pathologische Anatomie und Physiologie und für klinische Medicin* **119**, 299–326 (1890). 1

[4] Calkins, G. N. Zur frage der entstehung maligner tumoren (1914). 1

[5] Reddy, E. P., Reynolds, R. K., Santos, E. & Barbacid, M. A point mutation is responsible for the acquisition of transforming properties by the t24 human bladder carcinoma oncogene. *Nature* **300**, 149 (1982). 1

[6] Consortium, I. H. G. S. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). 1

[7] McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008). 1

[8] Muir, P. *et al.* The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology* **17**, 53 (2016). 1

[9] Coiffier, B. *et al.* CHOP chemotherapy plus rituximab compared with CHOP alone in elderly patients with diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 235–242 (2002). 1, 15

[10] Pfreundschuh, M. *et al.* CHOP-like chemotherapy plus rituximab versus CHOP-like chemotherapy alone in young patients with good-prognosis diffuse large-B-cell lymphoma: a randomised controlled trial by the MabThera International Trial (MInT) Group. *The lancet oncology* **7**, 379–391 (2006). 1, 15

[11] Weigelt, B., Peterse, J. L. & Van't Veer, L. J. Breast cancer metastasis: markers and models. *Nature reviews cancer* **5**, 591 (2005). 2, 16, 17

[12] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719 (2009). 3, 8

[13] Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *cell* **100**, 57–70 (2000). 3

[14] Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *cell* **144**, 646–674 (2011). 3, 4

[15] Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197 (1975). 3

[16] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976). 3

[17] Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome medicine* **2**, 54 (2010). 3

[18] Ames, B. N. & Gold, L. S. Endogenous mutagens and the causes of aging and cancer. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **250**, 3–16 (1991). 3

[19] Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature Reviews Cancer* **6**, 924 (2006). 3

[20] Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017). 3

[21] De Magalhães, J. P. How ageing processes influence cancer. *Nature Reviews Cancer* **13**, 357 (2013). 3

[22] Talbot, S. J. & Crawford, D. H. Viruses and tumours–an update. *European Journal of Cancer* **40**, 1998–2005 (2004). 4

[23] Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *New England journal of medicine* **343**, 78–85 (2000). 4

[24] Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology* **15**, 81 (2018). 5

[25] Swanton, C., McGranahan, N., Starrett, G. J. & Harris, R. S. APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. *Cancer discovery* (2015). 5, 36

[26] Mazor, T., Pankov, A., Song, J. S. & Costello, J. F. Intratumoral heterogeneity of the epigenome. *Cancer cell* **29**, 440–451 (2016). 5

[27] McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017). 5, 6

[28] Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338 (2013). 6

[29] Davis, A., Gao, R. & Navin, N. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* **1867**, 151–161 (2017). 6, 9

[30] Burrell, R. A. & Swanton, C. Tumour heterogeneity and the evolution of polyclonal drug resistance. *Molecular oncology* **8**, 1095–1111 (2014). 6

[31] Friedman, A. A., Letai, A., Fisher, D. E. & Flaherty, K. T. Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer* **15**, 747 (2015). 6

[32] Mehta, S. *et al.* Predictive and prognostic molecular markers for cancer medicine. *Therapeutic advances in medical oncology* **2**, 125–148 (2010). 6

[33] Rouzier, R. *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical cancer research* **11**, 5678–5685 (2005). 6

[34] Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* **17**, 79 (2017). 6

[35] Mehlen, P. & Puisieux, A. Metastasis: a question of life or death. *Nature Reviews Cancer* **6**, 449 (2006). 6

[36] Valastyan, S. & Weinberg, R. A. Tumor metastasis: molecular insights and evolving paradigms. *Cell* **147**, 275–292 (2011). 6

[37] Turajlic, S. & Swanton, C. Metastasis as an evolutionary process. *Science* **352**, 169–175 (2016). 7, 8

[38] Klein, C. A. Parallel progression of primary tumours and metastases. *Nature Reviews Cancer* **9**, 302 (2009). 7

[39] Naxerova, K. & Jain, R. K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nature reviews Clinical oncology* **12**, 258 (2015). 7, 8, 31

[40] Vogelstein, B. *et al.* Cancer genome landscapes. *science* **339**, 1546–1558 (2013). 8, 11

[41] Bielski, C. M. *et al.* Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature genetics* 1 (2018). 8, 20

[42] Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* **11**, 685 (2010). 9, 18, 19

[43] Fehrmann, R. S. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature genetics* **47**, 115 (2015). 9

[44] Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nature genetics* **48**, 1119 (2016). 9

[45] Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *cell* **144**, 27–40 (2011). 9

[46] Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214 (2011). 9

[47] Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012). 9

[48] Tubbs, A. & Nussenzweig, A. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* **168**, 644–656 (2017). 10

[49] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013). 10, 21, 28, 32

[50] Jinks-Robertson, S. & Bhagwat, A. S. Transcription-associated mutagenesis. *Annual review of genetics* **48**, 341–359 (2014). 10

[51] Haradhvala, N. J. *et al.* Mutational strand asymmetries in cancer genomes reveal mechanisms of dna damage and repair. *Cell* **164**, 538–549 (2016). 10

[52] Lindahl, T. Instability and decay of the primary structure of DNA. *nature* **362**, 709 (1993). 10

[53] Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced dna damage model for cancer development. *science* **319**, 1352–1355 (2008). 10

[54] Vitre, B. D. & Cleveland, D. W. Centrosomes, chromosome instability (CIN) and aneuploidy. *Current opinion in cell biology* **24**, 809–815 (2012). 10

[55] Artandi, S. E. & DePinho, R. A. Telomeres and telomerase in cancer. *Carcinogenesis* **31**, 9–18 (2009). 10

[56] Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annual review of biochemistry* **73**, 39–85 (2004). 10

[57] Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47 (2016). 10, 17

[58] Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**, 585 (2014). 11

[59] Moynahan, M. E. & Jasin, M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature reviews Molecular cell biology* **11**, 196 (2010). 11

[60] Rattray, A. J. & Strathern, J. N. Error-prone DNA polymerases: when making a mistake is the only way to get ahead. *Annual review of genetics* **37**, 31–66 (2003). 11

[61] Futreal, P. A. *et al.* A census of human cancer genes. *Nature Reviews Cancer* **4**, 177 (2004). 11

[62] Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214 (2013). 11

[63] Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12**, R41 (2011). 11

[64] Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013). 11

[65] Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology* **17**, 128 (2016). 11

[66] Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature genetics* **46**, 1258 (2014). 11

[67] Huang, F. W. *et al.* Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013). 11

[68] Tsujimoto, Y., Finger, L. R., Yunis, J., Nowell, P. C. & Croce, C. M. Cloning of the chromosome breakpoint of neoplastic B cells with the t (14; 18) chromosome translocation. *Science* **226**, 1097–1099 (1984). 11

[69] Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428 (2014). 11

[70] Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971). 11

[71] Payne, S. R. & Kemp, C. J. Tumor suppressor genetics. *Carcinogenesis* **26**, 2031–2045 (2005). 12

[72] Oliveira, A. M., Ross, J. S. & Fletcher, J. A. Tumor suppressor genes in breast cancer: the gatekeepers and the caretakers. *Pathology Patterns Reviews* **124**, S16–S28 (2005). 12

[73] Armitage, J. O., Gascoyne, R. D., Lunning, M. A. & Cavalli, F. Non-Hodgkin lymphoma. *The Lancet* **390**, 298–310 (2017). 12

[74] Pieper, K., Grimbacher, B. & Eibel, H. B-cell biology and development. *Journal of Allergy and Clinical Immunology* **131**, 959–971 (2013). 12, 13

[75] De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nature reviews immunology* **15**, 137 (2015). 12

[76] Victora, G. D. & Nussenzweig, M. C. Germinal centers. *Annual review of immunology* **30**, 429–457 (2012). 12

[77] Rajewsky, K. Clonal selection and learning in the antibody system. *Nature* **381**, 751 (1996). 12

[78] Shaffer III, A. L., Young, R. M. & Staudt, L. M. Pathogenesis of human B cell lymphomas. *Annual review of immunology* **30**, 565–610 (2012). 13, 14

[79] Basso, K. & Dalla-Favera, R. Germinal centres and B cell lymphomagenesis. *Nature reviews Immunology* **15**, 172 (2015). 13, 16

[80] Küppers, R. Mechanisms of B-cell lymphoma pathogenesis. *Nature Reviews Cancer* **5**, 251 (2005). 14

[81] Nussenzweig, A. & Nussenzweig, M. C. Origin of chromosomal translocations in lymphoid cancer. *Cell* **141**, 27–38 (2010). 14

[82] Pasqualucci, L. *et al.* Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature* **412**, 341–346 (2001). 14, 26

[83] Both, G. W., Taylor, L., Pollard, J. & Steele, E. Distribution of mutations around rearranged heavy-chain antibody variable-region genes. *Molecular and cellular biology* **10**, 5187–5196 (1990). 14

[84] Fukita, Y., Jacobs, H. & Rajewsky, K. Somatic hypermutation in the heavy chain locus correlates with transcription. *Immunity* **9**, 105–114 (1998). 14

[85] Bachl, J., Carlson, C., Gray-Schopfer, V., Dessing, M. & Olsson, C. Increased transcription levels induce higher mutation rates in a hypermutating cell line. *The Journal of Immunology* **166**, 5051–5057 (2001). 14

[86] Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T. & Kunkel, T. A. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nature immunology* **2**, 530 (2001). 14, 28

[87] Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007). 14

[88] Peled, J. U. *et al.* The biochemistry of somatic hypermutation. *Annu. Rev. Immunol.* **26**, 481–511 (2008). 14

[89] Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nature reviews Immunology* **6**, 573 (2006). 14, 15, 26

[90] Khodabakhshi, A. H. *et al.* Recurrent targets of aberrant somatic hypermutation in lymphoma. *Oncotarget* **3**, 1308 (2012). 14, 26

[91] Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature communications* **6**, 8866 (2015). 14

[92] Álvarez-Prado, Á. F. *et al.* A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets. *Journal of Experimental Medicine* **215**, 761–771 (2018). 14, 15

[93] Pavri, R. *et al.* Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**, 122–133 (2010). 14

[94] Meng, F.-L. *et al.* Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell* **159**, 1538–1548 (2014). 15

[95] Qian, J. *et al.* B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell* **159**, 1524–1537 (2014). 15

[96] Friedberg, J. W. Relapsed/refractory diffuse large B-cell lymphoma. *ASH Education Program Book* **2011**, 498–505 (2011). 15, 16

[97] Shipp, M. A predictive model for aggressive non-Hodgkin's lymphoma. the international non-Hodgkin's lymphoma prognostic factors project. *N Engl j Med* **329**, 987–994 (1993). 15

[98] Alizadeh, A. A., Elsen, M. B., Davis, R. E., Ma, C. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503 (2000). 15

[99] Lenz, G. *et al.* Stromal gene signatures in large-B-cell lymphomas. *New England Journal of Medicine* **359**, 2313–2323 (2008). 16, 24, 29

[100] Rosenwald, A. *et al.* The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine* **346**, 1937–1947 (2002). 16

[101] Pasqualucci, L. *et al.* Analysis of the coding genome of diffuse large B-cell lymphoma. *Nature genetics* **43**, 830 (2011). 16

[102] Morin, R. D. *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* **476**, 298 (2011). 16

[103] Lohr, J. G. *et al.* Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proceedings of the National Academy of Sciences* **109**, 3879–3884 (2012). 16

[104] Reddy, A. *et al.* Genetic and functional drivers of diffuse large B cell lymphoma. *Cell* **171**, 481–494 (2017). 16, 24, 29

[105] Meriranta, L. *et al.* Deltex-1 mutations predict poor survival in diffuse large B-cell lymphoma. *haematologica* **102**, e195–e198 (2017). 16

[106] Schmitz, R. *et al.* Genetics and pathogenesis of diffuse large B-cell lymphoma. *New England Journal of Medicine* **378**, 1396–1407 (2018). 16, 24, 29

[107] Chapuy, B. *et al.* Molecular subtypes of diffuse large B cell lymphoma are associated with distinct pathogenic mechanisms and outcomes. *Nature medicine* (2018). 16, 24, 29

[108] Larouche, J.-F. *et al.* Lymphoma recurrence 5 years or later following diffuse large B-cell lymphoma: clinical characteristics and outcome. *Journal of Clinical Oncology* **28**, 2094–2100 (2010). 16

[109] Morin, R. D. *et al.* Genetic landscapes of relapsed and refractory diffuse large B cell lymphomas. *Clinical cancer research* clincanres–2123 (2015). 16

[110] Jiang, Y. *et al.* Deep sequencing reveals clonal evolution patterns and mutation events associated with relapse in B-cell lymphomas. *Genome biology* **15**, 432 (2014). 16

[111] Juskevicius, D. *et al.* Distinct genetic evolution patterns of relapsing diffuse large B-cell lymphoma revealed by genome-wide copy number aberration and targeted sequencing analysis. *Leukemia* **30**, 2385–2395 (2016). 16

[112] Gisselbrecht, C. *et al.* Salvage regimens with autologous transplantation for relapsed large B-cell lymphoma in the rituximab era. *Journal of Clinical Oncology* **28**, 4184–4190 (2010). 16

[113] Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019). 16

[114] Cao, S.-S. & Lu, C.-T. Recent perspectives of breast cancer prognosis and predictive factors. *Oncology letters* **12**, 3674–3678 (2016). 16

[115] Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *New England Journal of Medicine* **353**, 1673–1684 (2005). 17

[116] Lord, C. J. & Ashworth, A. BRCAness revisited. *Nature Reviews Cancer* **16**, 110 (2016). 17

[117] Perou, C. M. *et al.* Molecular portraits of human breast tumours. *nature* **406**, 747 (2000). 17

[118] Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98**, 10869–10874 (2001). 17

[119] Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *American journal of cancer research* **5**, 2929 (2015). 17

[120] Network, C. G. A. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61 (2012). 17, 24

[121] Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology* **27**, 1160 (2009). 17

[122] Van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *nature* **415**, 530 (2002). 17

[123] Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine* **351**, 2817–2826 (2004). 17

[124] Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127 (2013). 17

[125] Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346 (2012). 17

[126] Telli, M. L. *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). 17

[127] Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nature medicine* **21**, 751 (2015). 17

[128] Yates, L. R. *et al.* Genomic evolution of breast cancer metastasis and relapse. *Cancer Cell* **32**, 169–184 (2017). 17

[129] Kjällquist, U. *et al.* Exome sequencing of primary breast cancers with paired metastatic lesions reveals metastasis-enriched mutations in the A-kinase anchoring protein family (akaps). *BMC cancer* **18**, 174 (2018). 17

[130] Miller, C. A. *et al.* Aromatase inhibition remodels the clonal architecture of estrogen-receptor-positive breast cancers. *Nature communications* **7**, 12498 (2016). 18

[131] Robinson, D. R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature genetics* **45**, 1446 (2013). 18

[132] Kim, C. *et al.* Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893 (2018). 18

[133] Levy, S. E. & Myers, R. M. Advancements in next-generation sequencing. *Annual review of genomics and human genetics* **17**, 95–115 (2016). 18

[134] Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nature genetics* **39**, 1522 (2007). 18

[135] Alioto, T. S. *et al.* A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature communications* **6**, 10001 (2015). 19

[136] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009). 19, 25, 27

[137] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357 (2012). 19, 27

[138] Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213 (2013). 19, 25

[139] Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* (2012). 19, 27

[140] Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012). 19

[141] Shin, H.-T. *et al.* Prevalence and detection of low-allele-fraction variants in clinical cancer samples. *Nature communications* **8**, 1377 (2017). 19

[142] Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41**, e67–e67 (2013). 19

[143] Do, H. & Dobrovic, A. Sequence artifacts in dna from formalin-fixed tissues: causes and strategies for minimization. *Clinical chemistry* clinchem–2014 (2014). 19

[144] Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413 (2012). 20, 21

[145] Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899 (2010). 20

[146] Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134 (2013). 20

[147] Olshen, A. B., Venkatraman, E., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics* **5**, 557–572 (2004). 20

[148] Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* **107**, 16910–16915 (2010). 20, 25

[149] Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012). 21

[150] Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *science* **331**, 1553–1558 (2011). 21

[151] Miller, J. H. Mutagenic specificity of ultraviolet light. *Journal of molecular biology* **182**, 45–65 (1985). 21

[152] Hernandez-Boussard, T. M. & Hainaut, P. A specific spectrum of p53 mutations in lung cancer from smokers: review of mutations compiled in the IARC p53 database. *Environmental Health Perspectives* **106**, 385 (1998). 21

[153] Comon, P. & Jutten, C. *Handbook of Blind Source Separation: Independent component analysis and applications* (Academic press, 2010). 21

[154] Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788 (1999). 21

[155] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell reports* **3**, 246–259 (2013). 21, 22

[156] Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* **101**, 4164–4169 (2004). 21

[157] Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562 (2001). 21

[158] Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome biology* **17**, 31 (2016). 22

[159] Morin, R. D. *et al.* Somatic mutations altering EZH2 (Tyr641) in follicular and diffuse large B-cell lymphomas of germinal-center origin. *Nature genetics* **42**, 181 (2010). 24

[160] Visco, C. *et al.* Comprehensive gene expression profiling and immunohistochemical studies support application of immunophenotypic algorithm for molecular subtype classification in diffuse large B-cell lymphoma: a report from the International DLBCL Rituximab-CHOP Consortium Program Study. *Leukemia* **26**, 2103–2113 (2012). 24, 29

[161] Arthur, S. E. *et al.* Genome-wide discovery of somatic regulatory variants in diffuse large B-cell lymphoma. *Nature Communications* **9**, 4001 (2018). URL `http://www.nature.com/articles/s41467-018-06354-3`. 24, 29

[162] Broad Institute. Picard Tools. `http://broadinstitute.github.io/picard/` (-). 25

[163] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**, 491 (2011). 25

[164] Icay, K. *et al.* SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData mining* **9**, 20 (2016). 25

[165] Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014). 25

[166] Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013). 25

[167] Roberts, A. & Pachter, L. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods* **10**, 71 (2013). 25

[168] Ovaska, K. *et al.* Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome medicine* **2**, 65 (2010). 25

[169] Cervera, A. *et al.* Anduril 2: Upgraded large-scale data integration framework. *Bioinformatics* (2019). 25

[170] Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164–e164 (2010). 25

[171] Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics* **56**, 15–9 (2016). 25

[172] Supek, F. & Lehner, B. Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell* **170**, 534–547 (2017). 25, 37

[173] Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**, 367 (2010). 26

[174] Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015). 26

[175] Farris, J. S. Phylogenetic analysis under Dollo's law. *Systematic Biology* **26**, 77–88 (1977). 26

[176] Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome biology* **16**, 91 (2015). 26

[177] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300 (1995). 26

[178] Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**, 91–118 (2003). 26

[179] Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573 (2010). 26

[180] Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proceedings of the National Academy of Sciences* (2011). 27

[181] Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome research* (2012). 27

[182] Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P. & Berri, S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28**, 40–47 (2011). 27

[183] Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nature methods* **6**, 99 (2008). 27

[184] Krishnan, N. M., Gaur, P., Chaudhary, R., Rao, A. A. & Panda, B. COPS: a sensitive and accurate tool for detecting somatic copy number alterations using short-read sequence data from paired samples. *PloS one* **7**, e47812 (2012). 27

[185] Ivakhno, S. *et al.* CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. *Bioinformatics* **26**, 3051–3058 (2010). 27

[186] Sathirapongsasuti, J. F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**, 2648–2654 (2011). 27

[187] Amarasinghe, K. C., Li, J. & Halgamuge, S. K. CoNVEX: copy number variation estimation in exome sequencing data using hmm. In *BMC bioinformatics*, vol. 14, S2 (BioMed Central, 2013). 27

[188] Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27**, 268–269 (2010). 27

[189] Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2011). 27

[190] Bartenhagen, C. & Dugas, M. RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics* **29**, 1679–1681 (2013). 27

[191] Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009). 27

[192] Rogozin, I. B. & Diaz, M. Cutting edge: DGYW/WRCH is a better predictor of mutability at G: C bases in ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *The Journal of Immunology* **172**, 3382–3384 (2004). 28

[193] Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nature genetics* **45**, 970 (2013). 32

[194] Wilson, W. H. *et al.* Targeting B cell receptor signaling with ibrutinib in diffuse large B cell lymphoma. *Nature medicine* **21**, 922 (2015). 35

[195] Jatoi, I., Hilsenbeck, S. G., Clark, G. M. & Osborne, C. K. Significance of axillary lymph node metastasis in primary breast cancer. *Journal of clinical oncology* **17**, 2334–2334 (1999). 36

[196] Giuliano, A. E. *et al.* Axillary dissection vs no axillary dissection in women with invasive breast cancer and sentinel node metastasis: a randomized clinical trial. *Jama* **305**, 569–575 (2011). 36

[197] Fisher, B. *et al.* Twenty-five-year follow-up of a randomized trial comparing radical mastectomy, total mastectomy, and total mastectomy followed by irradiation. *New England Journal of Medicine* **347**, 567–575 (2002). 36

[198] Galimberti, V. *et al.* Axillary dissection versus no axillary dissection in patients with sentinel-node micrometastases (IBCSG 23–01): a phase 3 randomised controlled trial. *The lancet oncology* **14**, 297–305 (2013). 36

[199] Naxerova, K. *et al.* Origins of lymphatic and distant metastases in human colorectal cancer. *Science* **357**, 55–60 (2017). 36

[200] Garcia-Murillas, I. *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. *Science translational medicine* **7**, 302ra133–302ra133 (2015). 36

[201] Wan, J. C. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer* **17**, 223 (2017). 36

[202] McGranahan, N. *et al.* Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science translational medicine* **7**, 283ra54–283ra54 (2015). 37

[203] Nikkilä, J. *et al.* Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. *British journal of cancer* **117**, 113 (2017). 37

[204] Li, M. *et al.* First-in-class small molecule inhibitors of the single-strand DNA cytosine deaminase APOBEC3G. *ACS chemical biology* **7**, 506–517 (2012). 37

[205] D'Antonio, M., Tamayo, P., Mesirov, J. P. & Frazer, K. A. Kataegis expression signature in breast cancer is associated with late onset, better prognosis, and higher HER2 levels. *Cell reports* **16**, 672–683 (2016). 37

[206] Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS computational biology* **12**, e1004873 (2016). 37

[207] Kuilman, T. *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome biology* **16**, 49 (2015). 37

[208] Luo, Z., Fan, X., Su, Y. & Huang, Y. S. Accurity: accurate tumor purity and ploidy inference from tumor-normal WGS data by jointly modelling somatic copy number alterations and heterozygous germline single-nucleotide-variants. *Bioinformatics* **34**, 2004–2011 (2018). 37

[209] Riester, M. *et al.* PureCN: copy number calling and SNV classification using targeted short read sequencing. *Source code for biology and medicine* **11**, 13 (2016). 37

[210] Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic acids research* **44**, e131–e131 (2016). 37

[211] Esteva, A. *et al.* A guide to deep learning in healthcare. *Nature Medicine* **25**, 24–29 (2019). URL `http://www.nature.com/articles/s41591-018-0316-z`. 37