The hunt for cancer genes

Statistical inference of cancer risk and driver genes using next generation sequencing data

Hana Sušak

DOCTORAL THESIS / YEAR 2017

THESIS SUPERVISOR
Dr. Stephan Ossowski
Department Genomic and Epigenomic Variation in Disease Group
Bioinformatics and Genomics Department

Universitat Pompeu Fabra
Barcelona

I would like to dedicate this thesis to my grandmother and uncle who did not live long to see me passing through the finishing line. Bako, uskoro ću početi da radim “za stvarno”.

Acknowledgments

The last four years have been a long journey but some people made it easier and more pleasant. First, I would like to express my gratitude to friends and colleagues of mine, Luis, with whom I've done the work presented in chapter two of this thesis, and Georgia, whit who I developed the project presented in chapter three of my thesis. I would like to thank Stephan, my supervisor, for giving me the chance to be part of a great lab and for all the support, guidance, and patience he showed as a person and as a boss. I am also thankful to my thesis committee members for all the valuable input they provided, to CRG for all the support, and to the CRG's secretaries and HR personnel for their help, especially Rut, Romina, Gloria, and Imma. Also many thanks to every past and present member of the Ossowski lab. It was a wonderful place to work!

I would like to give special thanks to my friends Freza, Cebolla and Batti with whom I've started my Ph.D. Big thank you for being such a rich source of conversations, education, and entertainment. Also to all my other friends I've met during the thesis progression: Bambino, Azul, Linux, Tigi.

I will remember and want to mention several people outside the science bubble who greatly helped me to finish my thesis. I am grateful to my mother and father, whose support is the greatest gift I've got in my life. They planted great values in my head since childhood, that education and knowledge were always of the most importance. Besides my parents, my greatest source of support and love is my sister Žana. Her daughter Asja is my source of joy.

Maybe not sisters by blood, but by everything else: Joja, Mima, Ema, and Una. I am so happy to have you in my life. I have to thank two little ones, Sofi and Luka for being so proud of me. What can be more satisfying and motivating than being a child's role model?! And I can not forget to mention and thank my local "non–science" friends, Kala, Miljana, Ciki, and Bane who kept my social life going and for being great supportive friends whenever I needed them!

Abstract

International cancer sequencing projects have generated comprehensive catalogs of alterations found in tumor genomes, as well as germline variant data for thousands of individuals. In this thesis we describe two statistical methods exploiting these rich datasets in order to better understand tumor initiation, tumor progression and the contribution of genetic variants to the lifetime risk of developing cancer. The first method, a Bayesian inference model named cDriver, utilizes multiple signatures of positive selection acting on tumor genomes to predict cancer driver genes. Cancer cell fraction is introduced as a novel signature of positive selection on a cellular level, based on the hypothesis that cells obtaining additional advantageous driver mutations will undergo rapid proliferation and clonal expansion. We benchmarked cDriver against state of the art driver prediction methods on three cancer datasets demonstrating equal or better performance than the best competing tool. The second method, termed REWAS is a comprehensive framework for rare-variant association studies (RVAS) aiming at improving identification of cancer predisposition genes. Nonetheless, REWAS is readily applicable to any case-control study of complex diseases. Besides integrating well-established RVAS methods, we developed a novel Bayesian inference RVAS method (BATI) based on Integrated Nested Laplace Approximation (INLA). We demonstrate that BATI outperforms other methods on realistic simulated datasets, especially when meaningful biological context (e.g. functional impact of variants) is available or when risk variants in sum explain low phenotypic variance. Both methods developed during my thesis have the potential to facilitate personalized medicine and oncology through identification of novel therapeutic targets and identification of genetic predisposition facilitating prevention and early diagnosis of cancer.

Resum

Els distints projectes internacionals de seqüenciació de càncer duts a terme en els últims anys han generat catàlegs complets d'alteracions trobades en els genomes tumorals, així com informació de variants germinals per a milers d'individus. En aquesta tesi descrivim dos mètodes estadístics aprofitant aquestes bases de dades per tal d'entendre millor la iniciació i la progressió dels tumors, i la contribució de variants genètiques al risc de desenvolupar càncer al llarg de la vida. El primer mètode, anomenat cDriver, es basa en un model d'inferència Bayesià que utilitza múltiples senyals de la selecció positiva que ocorre en els genomes tumorals per tal de predir els gens *driver* del càncer. En aquest mètode, hem inclòs la fracció de cèl·lules tumorals com a nova senyal de la selecció positiva a nivell cel·lular. Aquesta es basa en la hipòtesi que les cèl·lules que adquireixen mutacions ventajoses addicionals proliferaran i s'expandiran clonalment més ràpidament. Per avaluar cDriver, aquest es va comparar amb els mètodes més utilitzats per a la predicció de gens *driver* actuals. L'anàlisi es va dur a terme amb conjunts de dades de tres càncer diferents i els resultats van ser iguals o millors que els obtinguts per les eines més competitives en el tema. El segon mètode, anomenat REWAS, és un marc de treball per l'estudi d'associació de variants rares (RVAS) amb l'objectiu de millorar la identificació dels gens de predisposició al càncer. Tot i això, REWAS es pot aplicar a qualsevol estudi cas-control de malalties complexes. Per una altra part, a més d'integrar mètodes RVAS ben establerts, hem desenvolupat un nou mètode d'inferència Bayesiana RVAS (BATI) basat en *Integrated Nested Laplace Approximation* (INLA). També demostrem que BATI mostra millors resultats que altres mètodes en dades simulades amb soroll de fons real, especialment quan el context biològic (p.e. variants amb impacte funcional) està disponible or quan les variants de risc expliquen en total poca variància fenotípica. Tots dos mètodes desenvolupats durant la meva tesi tenen el potencial de facilitar la medicina i l'oncologia personalitzada mitjançant la identificació de noves dianes terapèutiques i la predisposició genètica que faciliti la prevenció i el diagnòstic precoç del càncer.

Preface

Though I had great interest in genetics since high school, I never had the chance to practice and learn more about it until my masters studies. Some would say that my journey to the Ossowski lab was unconventional, from computer science to machine learning, and ending up analyzing cancer genomics data, while not knowing at the time what is an epigenome or a Golgi apparatus. My first experience in bioinformatics started quite unexpectedly, as a master thesis project at CRG, that continued with 4 years of Ph.D.

Since then I believe my experience in basic biology, and especially in cancer genomic has grown exponentially (easy when you start from zero). For this I have to thank all the lab members for their forbearance with all my questions and confused facial expressions, and other colleagues outside the lab who were also very understanding and helpful whenever I needed. I can still remember my excitement once my colleagues and dear friends Andrea, Reza, and Sebastian laughed at my first successful biological joke, which if I recall well was “he is as promiscuous as CTCF”. I especially would like to mention Luis, who explained me everything I know today about evolution, Daniela who was always volunteering to help me understand wet-lab experimental work, Oliver and Fran who were always available for all my basic biological questions, and Mattia who is the ‘God of annotation’. Still the one with most patience and the one who took the biggest risk when hiring me is my PI, is Stephan Ossowski. I hope he never regretted it. From my side these last 4.5 years were my best working experience ever. I had the opportunity to learn, work on topics and projects I like, to meet incredibly intelligent people, grow as a person, and also to pass on some of my knowledge to others. For all this, I am in debt to Stephan, and to this amazing place few steps from the beach - CRG. Now when I look back, the toughest part of my PhD will be leaving the lab.

This thesis was a natural continuation of my masters thesis work “Analysis of genetic variant enrichment in PPI networks” in which I analyzed germline and somatic point mutations in chronic lymphocytic leukemia (CLL). Drawing on the experience from my past work, we have continued with two projects in which we developed software to address predictions of cancer driver and predisposition genes involved in any cancer type. In my opinion, this thesis work can provide help to anyone who wants to analyze NGS data from cancer cohorts or any other complex disease. The first project, implementation of a tool for prediction of cancer driver genes was mostly done together with my colleague and dear friend Luis Zapata. The second project, development of an RVAS framework for inferring cancer predisposition genes was done for the most part together with Georgia Escaramis Babiano with whom I had the pleasure to work. I would be delighted to have the opportunity to work with her again.

I hope you enjoy reading the thesis! (When you see a typo please consider that most of the thesis was written between 8PM and 8AM, over several months.)

Hana Sušak,
Barcelona, 29 September 2017

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Discovering Cancer through history

One of the first links between cancer and genetics was found as early as 1914, when Theodor Boveri proposed and tested the hypothesis that tumors develop as a result of abnormal chromosome segregation [Boveri, 1914, Balmain, 2001, Stratton et al., 2009]. He observed that abnormal cell division in sea urchins often led to the death of daughter cells or, less often, to an aberrant development of daughter cells [Boveri, 1914, Balmain, 2001]. Based on fundamental discoveries in cancer research achieved during the 20th century, tumors are now defined as clusters of identical cells (clones) having a fitness advantage over their neighboring “normal” cells. Abnormally increased growth and proliferation of cells are typically used as measures of increased fitness. Moreover, Boveri hypothesized that tumors arise from a single genetically altered daughter cell when he observed the same characteristic chromosome abnormalities in all the cells that were surviving. This concept was at first known as “the stem line concept”, but today we refer to it as the “clonal evolution” model of tumor development. Clonal tumor evolution can be compared to Darwinian evolution theory where we have two ongoing processes: (*i*) random acquisition of mutations in cells and (*ii*) natural selection acting on phenotypes (traits) of cells [Stratton et al., 2009]. Half a century after the work of Boveri, technologies became available that allowed to prove Boveri’s hypothesis. In 1960 a chromosome aberration, today known as ‘Philadelphia chromosome’, was discovered in patients diagnosed with chronic myelogenous leukemia (CML) using cytogenetic techniques [Nowell and Hungerford, 1960, Balmain, 2001]. In the following years, it was discovered that the Philadelphia chromosome was formed due to a translocation between chromosomes 9 and 22 [Rowley, 1973], leading to a fusion between the genes *BCR* and *ABL1*, and ultimately to the increased activity of *ABL1* [Heisterkamp et al., 1983]. Thus, *ABL1* was defined as an oncogene whose over-expression causes CML.

Our knowledge about cancer has been increasing whenever new technologies for genomics, proteomics, and molecular genetics became available. Today we observe that cancer forms and develops using different strategies. Besides the large chromosome abnormalities, it has been shown that accumulation of point mutations (substitution of one base) can also initiate tumor development [Tabin et al., 1982]. The majority of causal point mutations have been found in protein coding genes. Mutations giving a fitness advantage to cancer cells, e.g. allowing the cancer cells to proliferate faster, are typically called driver mutations. Genes in which these driver mutations reside are commonly referred as cancer driver genes. In addition to nucleotide substitutions, deletions, insertions and translocations of short or long segments of DNA are commonly found as driver events affecting cancer genes [Yang et al., 2010, Maruvka et al., 2017, Zhang et al., 2010, Hogenbirk et al., 2016]. Due to the prevalence of whole-exome sequencing (WES), driver mutations affecting coding genes have been the most intensely studied by the cancer genomics community. However, the field is slowly moving towards the use of the more powerful, but also challenging whole genome sequencing technique, which will allow for better interrogation of structural and non coding driver mutations. Other known somatic alterations that can contribute to tumor development are DNA rearrangements, copy gains, copy losses, elongation of telomeres, as well as epigenetic changes, or exogenous sources (e.g. bacteria or viruses) introducing completely new DNA sequences [Forment et al., 2012, Zack et al., 2013, Cox et al., 2005, Meyerson et al., 1997, Bodnar et al., 1998, Kulis and Esteller, 2010, Talbot and Crawford, 2004].

In addition to the accumulation of somatic genetic alterations, we also know that there is a heritable genetic basis for cancer susceptibility. This was first discovered based on the observation that many cancer types are recurrent in some, susceptible families, i.e. familial clustering of cancer cases [Broca, 1866]. Today, there is strong evidence that sporadic cancers also have a significant genetic component [Lichtenstein et al., 2000]. There have been several approaches to identify cancer predisposition genes (CPG), each of which had success for specific types of alterations. Genome-wide linkage analysis in familial cancer clusters was particular successful for identifying rare but highly penetrant mutations in CPGs [Easton et al., 1993]. Analysis of candidate genes present in the same pathways as known CPGs, or interacting with known CPGs, or surrogates of CPGs, revealed new CPGs [Meijers-Heijboer et al., 2002, Seal et al., 2006, Rahman et al., 2007, Loveday et al., 2011, Hanks et al., 2004]. Recently, next-generation sequencing facilitated the discovery of new CPG candidates by whole-exome sequencing (WES) of the DNA from familial cases as well as by WES of sporadic cases and healthy controls [Comino-Mendez et al., 2011, Smith et al., 2013]. However, the cohorts analyzed using WES have often been too small to allow for significant results, and larger scale efforts are necessary in the future. With an extensive effort of literature search and database interrogation, Rahman et al. [Rahman, 2014] compiled a list of 114

CPGs having rare mutations, which is considered a gold standard in the field. Additionally, common mutations in CPGs have been successfully identified with Genome-wide Association Studies (GWAS) [Varghese and Easton, 2010, Stadler et al., 2010, Chang et al., 2014]. From work with animal models of cancer, we further learned that several low-penetrance variants in genes could have tremendous impact when present in combination [Balmain and Nagase, 1998]. As cancer can be seen as a complex disease, and with rapidly increasing exome sequencing datasets of cancer patients, we can hope that standard approaches for rare variant association case-control studies could fill the gap for high-penetrance and low-penetrance rare variants in CPGs.

Based on data gathered trough 2003-2012 it is estimated that the number of new cancer cases per year ranges from 67 to 434 per 100,000 individuals depending on country and gender [Torre et al., 2016]. Unfortunately and despite the broad knowledge about cancer development, for many cases there is no successful treatment available. Hence, we are still challenged with 31 to 236 deaths per 100,000 individuals (again depending on country and gender). Population statistics reveal that today's newborns have an 18.5% chance of developing cancer before the age of 75 and a 10.5% chance of dying from cancer before the age of 75 [Ferlay et al., 2013a]. Although cancer is one of the most prevalent diseases in modern society and has been extensively studied, there are still several big challenges to be addressed, as for example: identifying the whole spectrum of low and intermediate frequency cancer driver genes [Lawrence et al., 2014], interrogating interactions between germline and somatic mutations in cancer genes [Gonzalez-Perez et al., 2013], identifying low-penetrance and/or low-frequency cancer predisposition genes [Balmain, 2001, Bodmer and Tomlinson, 2010], and implementation of longitudinal genomic studies of cancer patients [Weinstein et al., 2013]. Ultimately we want to better understand the causal mutations leading to malignant transformation of cells, such that patients can be diagnosed at an earlier stage and treated with mutation-profile specific drugs, i.e. to advance towards 'precision oncology'.

1.2 Next Generation Sequencing

How and how fast we analyze genomes dramatically changed in the mid 2000s, with the emergence of the first high-throughput sequencing platforms, including 454 Life Sciences' pyrosequencing and Solexa's SBS technologies (now owned and developed by Illumina). This generation of sequencing platforms is often referred to as 2nd or as next generation sequencing (NGS). Prior to the NGS era, Sanger sequencing was dominating the field. Automation of Sanger sequencing over more than two decades ultimately allowed for the completion of the first human genome sequence in 2001 [Lander et al., 2001]. However, finishing a single genome using Sanger required massive resources, money and the work of hundreds of scientists over ten years. But, with the arrival of NGS technologies sequencing cost and hands-on work for sequencing a human genome have decreased exponentially. First, the amount of nucleotides sequenced per time unit has skyrocketed due to massive parallelization. For example, throughput with automated Sanger sequencing can reach 0.166 Mb/hr while NGS reached a throughput of ~20 Mb/hr in 2008 [Sinville and Soper, 2007, Morozova and Marra, 2008] and ~10 Gb/hr today (Illumina NovaSeq specifications). Second, the price of sequencing a human genome to 30 fold coverage came down to ~1,000$ (`https://www.veritasgenetics.com/mygenome`). To complete a human genome with Sanger sequencing technologies required ~24 years since it's development [Sanger et al., 1977], but within ~5 years since NGS technologies became broadly available it is estimated that ~30,000 human genomes have been sequenced [News, 2010]. Reasons for exponential growth of the number of sequenced genomes are both the faster throughput and the brisk decline of cost per nucleotide since the emergences of NGS. Moreover, NGS sequencing technology has many advantages over Sanger sequencing technologies in cancer genome studies. For instance, the analysis of sub-clonal tumor structure of heterogeneous tumor samples and the study of low-purity tumor tissue (extensive contaminated with healthy tissue) benefit from the deep coverage provided by the billions of reads produced in a single sequencing run [Arsenic et al., 2015].

Driven by the rapidly improving NGS technologies new bioinformatics tools emerged that enabled the analysis of massive amounts of sequencing reads for various types of applications. NGS enabled the analysis of genomes, epigenomes, transcriptomes, and even interactomes (e.g. protein-DNA interactions) and hence forced the development of novel algorithms and data structures. Two common approaches are widely used in genetics studies, whole-genome sequencing (WGS) and whole-exome sequencing (WES). WES is based on techniques for enriching specific regions of the genome using oligo probes complementary to the sequence of interest. In case of WES, the targeted regions of the genome are the exons, units within genes which code for the proteins. The entirety of all exons is named the 'exome'. For WGS, no targeted enrichment is used and hence the whole genome is sequenced without enriching for preferential regions. As genomic regions that encode for proteins are only ~1.5% of total human genome, WES is cheaper than WGS. As most of the known disease-causing variants are found in coding genes, WES is a good compromise between cost and benefit. However, as WGS is becoming more and more affordable and because regulatory elements and structural genomic variations are becoming a focus of interest, WGS is expected to completely replace WES in the near future. As an added advantage, WGS is introducing less technological biases to the data, because coverage across the genome is more uniform if no target enrichment methods are applied. Nonetheless, both approaches have shown to be valuable methods for the discovery of the genetic causes of rare and complex diseases [Gonzaga-Jauregui et al., 2012], as well as cancer [TCGANetwork, 2008, TCGANetwork, 2012c, TCGANetwork, 2013].

The essential steps of generating and analyzing NGS data are shown in figure 1.1. The first step is library preparation. The exact wet-lab procedure depends on the sequencing type and differs substantially for applications like WES, WGS, RNA-seq or ChIP-seq. For WES an important step is the enrichment of the targeted coding regions for which several companies (e.g. Agilent) are offering specific kits containing the complementary biotinylated oligos that bind to target DNA and are then 'pulled' using magnetic beads. These target enrichment kits have constantly been updated and improved over the last years to e.g. include more genes or to better cover the targeted region. Specialized kits are available that do not target the whole exome, but only a specific sub-set of genes. However, in this work, if not mentioned otherwise, we have focused on WES. NGS platforms will produce raw sequence data, i.e. files in FASTA format containing the reads and per-base quality information. The analysis of WES or WGS data goes trough five steps (Figure 1.1): (i) quality control of the raw sequence data, (ii) alignment of the reads to a reference genome, (iii) variant identification, (iv) functional annotation of the variants and (v) statistical data analysis (e.g. genotype-phenotype association tests).

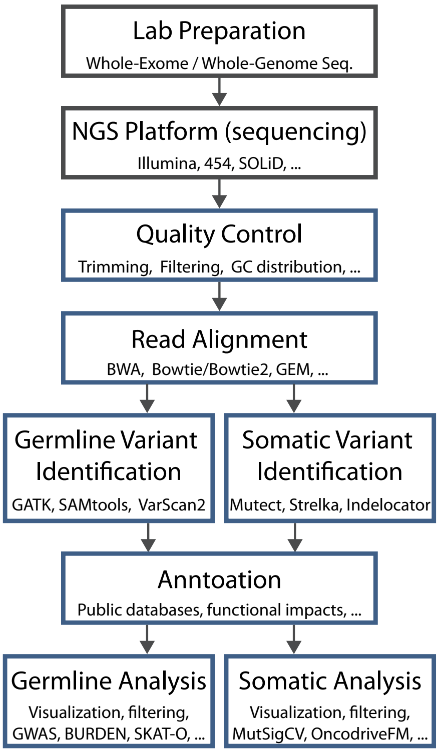
Lab Preparation
Whole-Exome / Whole-Genome Seq.
NGS Platform (sequencing)
Illumina, 454, SOLiD, ...
Quality Control
Trimming, Filtering, GC distribution, ...
Read Alignment
BWA, Bowtie/Bowtie2, GEM, ...
Germline Variant Identification
GATK, SAMtools, VarScan2
Somatic Variant Identification
Mutect, Strelka, Indelocator
Anntoation
Public databases, functional impacts, ...
Germline Analysis
Visualization, filtering, GWAS, BURDEN, SKAT-O, ...
Somatic Analysis
Visualization, filtering, MutSigCV, OncodriveFM, ...

Fig. 1.1: Basic work-flow for whole-exome and whole-genome sequencing projects.

First, quality control (QC) is performed to assure that reads are passing defined standards, as NGS procedures are susceptible to a wide range of chemistry and instrument failures [Dohm et al., 2008]. Reads can be removed or trimmed if they are not passing desired quality thresholds. It is possible to detect if any contamination was present, or if nucleotide distributions are not in the expected range, i.e. if the GC content distribution differs from the expectation. Second, the quality-checked reads have to be aligned to a reference genome. For human there are currently multiple reference genome versions being used such as hg18, hg19, and GRCh38 (listed from earlier to latest version using UCSC Genome Browser nomenclature). The bet reference genome for alignment should be chosen taking into account downstream analysis methods, available data, and compatibility with previous analyses. For example, the choice of tools for functional annotation of variants might depend on the reference genome version. Next, the alignment information is stored into a file, usually in the BAM format, which is the most frequently used alignment format to date. During the past ten years many alignment algorithms have been developed including the two most commonly used which are BWA [Li and Durbin, 2010] and Bowtie/Bowtie2 [Langmead et al., 2009, Langmead and Salzberg, 2012].

The next step is variant identification, which is performed based on the alignment provided in the BAM file. A variant is defined as an observed difference between the genome/exome of the sequenced individual and the human reference genome. There are many types of variants such as single nucleotide variants (SNVs), insertions, deletions, and other complex variants such as large scale rearrangements and copy number alterations. For the detection of germline Single Nucleotide Variants (SNVs) and short insertion-deletion variants (InDels) one of the most commonly used tools is GATK [DePristo et al., 2011]], which offers the 'GATK UnifiedGenotyper' and, more recently, the 'GATK HaplotypeCaller' algorithms. Both methods use a Bayesian inference model that returns a posterior probability for each of three genotypes (homozygous reference, heterozygous alternative or homozygous alternative base) for each position of the genome. Both methods also work with multiple samples at a time (multi-sample variant calling) and return all potentially variable positions for all samples in one 'variant call format' (VCF) file. For the identification of common and medium-rare variants, multi-sample calling is the advised mode, as pooled information from multiple samples can increase the quality of the calls. However, for ultra-rare variants and singletons, it does not provide any advantage compared to single-sample based methods (`software.broadinstitute.org/gatk/documentation/article?id=4150`). GATK furthermore provides quality score recalibration and depth of coverage analysis, as well as local re-alignment and assembly algorithms to improve InDel calling. Other tools for detection of SNVs and InDels are also available and widely used such as SAMtools [Li et al., 2009], VarScan2 [Koboldt et al., 2012b], SNVer [Wei et al., 2011], among others. Somatic SNV and InDel callers are trying to identify mutations which are different between tumor and healthy tissue from the same individual (often referred to as 'tumor and normal pairs'). There are several tools specialized for calling somatic SNVs and InDels. The most commonly used are: Mutect [Cibulskis et al., 2013], Strelka [Saunders et al., 2012], and SomaticSniper [Larson et al., 2012], with Strelka being the only one of the three able to call InDels. In recent years several tools specialized in calling somatic InDels were developed (e.g. Indelocator `http://archive.broadinstitute.org/cancer/cga/indelocator`). However, somatic InDel calling is still prone to high error rates, leading to substantial differences in call sets from different pipelines [Alioto et al., 2014]. Numerous tools are also available for calling structural variants and copy number variations (CNVs), for both somatic and germline variant analysis, but as they are out of the scope of this thesis they will not be elaborated here. Pabinger et al. [Pabinger et al., 2013], Koboldt et al. [Koboldt et al., 2012a], and Nakaga et al. [Nakagawa et al., 2015] have provided excellent reviews on SNV and CNV calling.

Functional annotation of variants is the last mandatory step before performing association and other statistical tests to infer disease causal variants and/or genes. Functional annotation methods aim at classifying each variant with information necessary to understand the impact of the variant on the phenotype of an individual. Besides basic information such as (*i*) amino-acid change, (*ii*) affected gene, and (*iii*) exonic function, most annotation tools can provide information from multiple variant databases. Such information could come in the form of scores estimating evolutionary conservation (e.g. phyloP and phastCons), possible damage caused to the protein's function (e.g. PolyPhen2, SIFT, CADD, etc.) and allele frequencies observed in different human populations (Exome Variant Server (`evs.gs.washington.edu/EVS/`), 1000 Genomes Project, and ExAC (`exac.broadinstitute.org/`)). In addition, prior knowledge could be added from disease or cancer-specific databases such as OMIM, ClinVar, and COSMIC. Annotation is an extremely important step as it can help to reduce the number of variants that could be of interest in disease studies and hence, to significantly reduce the number of variants participating in association tests. Some of the most used annotation pipelines are ANNOVAR [Wang et al., 2010] and Variant effect predictor (VEP) [McLaren et al., 2010], but the choice is not limited to those. In this work, the eDiVa pipeline (`ediva.crg.es/`) is used for annotation, which is an in-house developed annotation database and pipeline with a large collection of disease-specific features.

The cancer genomics field is rapidly evolving, and procedures, formats and tools are in constant development and improvement. Big consortia such as The Cancer Genome Atlas (TCGA; `cancergenome.nih.gov/`) and the International Cancer Genome Consortium (ICGC; `icgc.org/`) have made great efforts to introduce best practice work-flows, which we have followed in this work whenever possible.

1.3 Cancer study design

1.3.1 Germline versus somatic type of mutations - in cancer context

An affordable strategy to identify the genetic component of tumorigenesis is to sequence the cancer tissue and the healthy (cancer free) tissue of diagnosed patients (Figure 1.2). Although it would be optimal to have the same cell lineage for both cancer and normal cells, only blood is often used as the healthy counterpart (or a non-cancer blood cell type in case of hematopoietic cancers). As all cells in any human tissue are descendants of a fertilized egg (through mitotic cell division) sequencing their genome can reveal variants either inherited from the parents ('germline variants') or acquired after the formation of the fertilized egg (Figure 1.3). The mutations acquired after formation of the fertilized egg are referred to as somatic mutations to distinguish them from the inherited germline variants. Also, somatic mutations can accumulate before malignant transformation, a phenomenon often termed 'mosaicism', can accumulate after malignant transformation, or can initiate tumorigenesis. Somatic alterations can be the product of external and internal mechanisms [Stratton et al., 2009, Alexandrov et al., 2013]. Examples of external sources could be the exposure to tobacco smoke carcinogens or to radiation such as ultraviolet light. An example for an internal source of somatic mutations is the accumulation of errors during DNA replication. Although most somatic mutations can be repaired or identified at multiple check points, leading to arrested proliferation [Massague, 2008] or cell death [Iyer et al., 2006, Larrea et al., 2010]), a small fraction does not trigger any of these mechanisms and it is passed on to further generations. Several studies have shown that in cancer cells, DNA repair and check point pathways are often disabled by somatic driver mutations, ultimately allowing cells to proliferate uncontrollably.

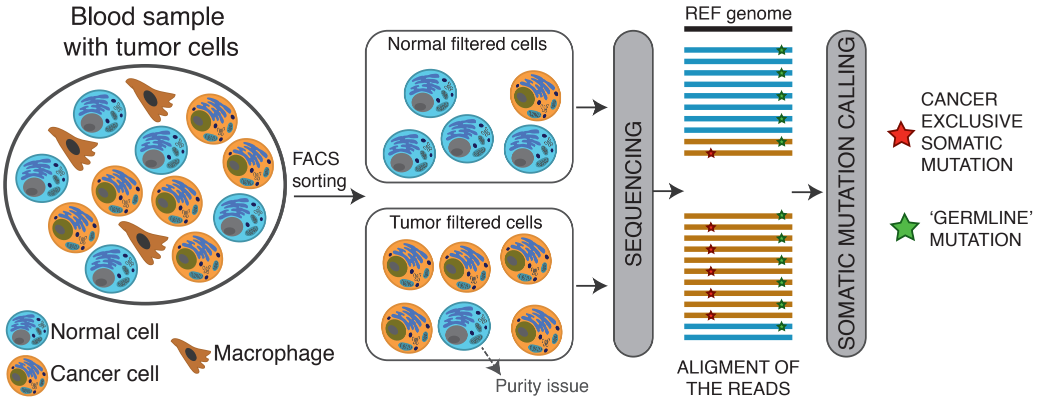

Figure 1.2: Process of inferring cancer somatic mutations from hematologic cancer.

As the first step normal and tumor tissue need to be separated. Then, sequencing is done for both tissues and reads are aligned to a reference genome. Specialized tools (e.g. MuTect) that use BAM files for normal and cancer tissue are used to infer somatic mutations found only in tumor tissue.

Additionally, the difference between variants found in tumor tissue and variants found in healthy tissue (Figure 1.2) can reveal the set of tumor-specific mutations. Such strategy reveals those somatic mutations exclusively found in the tumor tissue and remove all germline and somatic mutations found in the healthy tissue (in Figure 1.3 red mutations). In this thesis, 'tumor exclusive' variants are simply referred as somatic mutations. Though we recognize that both normal and cancer cells carry somatic mutations, mosaicisms in healthy tissue are not a primary interest of this study.

Often, the tumor tissue is contaminated with healthy tissue due to difficulties during the physical separation of the two types. This phenomenon is described as the purity of the tumor tissue or the fraction of the tissue that has cancer properties. Whereas in hematological tumors it is possible to achieve very high purity thanks to the separation of normal cells from tumor cells using FACS, in solid tumors the purity is very heterogeneous across samples and cancer types [Aran et al., 2015], ranging from below 20% to close to 100%. The tumor purity can be estimated by pathologists or by using bioinformatics tools such as ASCAT [Van Loo et al., 2010] or ABSOLUTE [Carter et al., 2012]. The latter tools estimate the purity based on the minor allele frequency (MAF) distribution of predicted somatic SNVs and/or CNVs.

Both, germline and somatic variants, are important for cancer etiology. While some germline variants are thought to influence cancer predisposition, some somatic mutations can drive malignant transformation and cancer development. Germline risk variants often affect DNA damage repair genes [Hodgson, 2008, Ponder, 2001], leading to a reduced repair efficiency and hence to a larger lifetime risk of developing a cancer [Cybulski et al., 2015, Li et al., 2012a]. Somatic mutations driving cancer can affect several important functions of the cell, but most importantly they often impair cell cycle control mechanisms allowing for unhindered cell proliferation. The identification of somatic driver events is seemingly easier than inferring germline risk variants. This is mostly due to the amount of variants observed since tumors accumulate between tens to a few hundreds somatic mutations in coding regions [Martincorena and Campbell, 2015, Alexandrov et al., 2013] while individuals carry more than ~40,000 germline variants in coding regions. Therefore, a smaller pool of candidate variants and less noise reduces the number of false positives when using somatic mutations for identification of causal events. However, prediction methods for somatic mutations typically have a higher error rate and specifically can mistake germline as somatic variants. Thus, methods to identify cancer driver genes and cancer risk genes tackle different types of errors and noise, and require specialized background models and algorithms.

1.3.2 Driver and passenger somatic events in cancer

Somatic mutations contributing to malignant transformation or cancer development are called driver mutations. Genes carrying driver mutations are defined as cancer genes (also termed ‘cancer driver genes’). Driver mutations confer a selective advantage and therefore increase cancer cell fitness [Stratton et al., 2009]. A well-described selective advantage is the acquisition of a higher proliferation rate compared to surrounding cells, which is also considered as one of the hallmarks of cancer [Hanahan and Weinberg, 2011]. Other somatic mutations that are either neutral or negatively selected, are called passenger mutations (Figure 1.3) and they do not confer a selective advantage. Passenger mutations could occur prior to tumor formation or during tumor progression.

During a selective sweep caused by a highly beneficial driver mutation, passenger mutations can hitchhike within the clonal expansion and reach a high tumor fraction. Some somatic events can also impair cell survival and will likely be purified in time (negative selection). However, the impact of negative selection in tumor evolution is still controversial [Ostrow et al., 2014, Pyatnitskiy et al., 2015]. Importantly, the ability of a mutation to drive tumor initiation or progression depends on the tissue, biological context, and/or the environment. For example, a new independent driver event increasing a cell's proliferation rate can make previous somatic mutations neutral or even negatively selected. Also, mutations are not isolated events, and they act together in a given context, so effects of two or more mutations together might differ from their individual effects. This phenomenon is known as epistasis. Importantly, many coding mutation hotspots (e.g. *BRAF* V600 [Davies et al., 2002, Ascierto et al., 2012, Thiel and Ristimaki, 2013]) mainly cause cancer in one or a few tissue types, but have not been reported as cancer drivers for other tissues. In summary, different cancer genotypes are in constant competition, and somatic events contributing to cancer development in a given moment are expected to show signatures of positive selection.

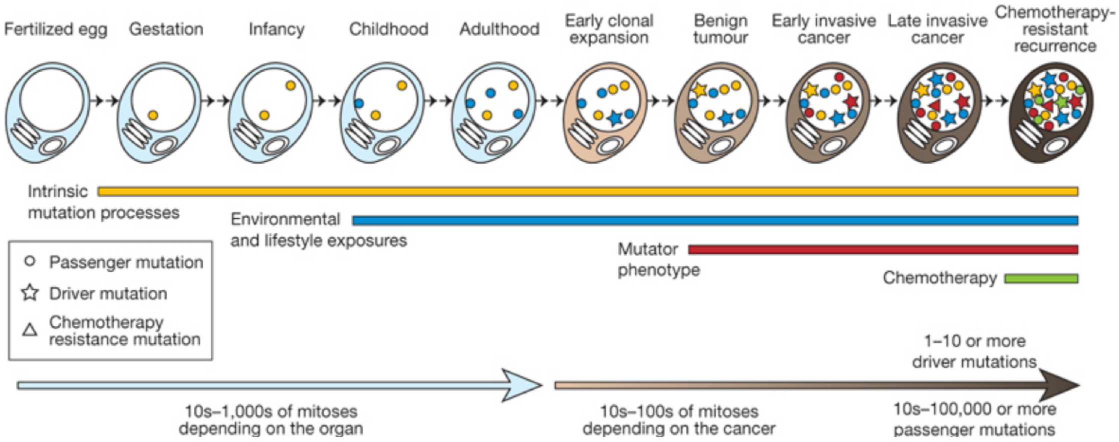
Fertilized egg
Gestation
Infancy
Childhood
Adulthood
Early clonal expansion
Benign tumour
Early invasive cancer
Late invasive cancer
Chemotherapy-resistant recurrence
Intrinsic mutation processes
Environmental and lifestyle exposures
Mutator phenotype
Chemotherapy
○ Passenger mutation
☆ Driver mutation
△ Chemotherapy resistance mutation
1–10 or more driver mutations
10s–1,000s of mitoses depending on the organ
10s–100s of mitoses depending on the cancer
10s–100,000 or more passenger mutations

Figure 1.3: The lineage of mitotic cell divisions from the fertilized egg to a single cell within a cancer mass, showing the timing of somatic mutation events and the processes that contribute to tumorigenesis. Reproduced from [Stratton et al., 2009]

1.3.3 Driver genes - oncogenes and tumor suppressors

There are two major classes of cancer genes, oncogenes and tumor suppressor genes. Oncogenes were discovered in 1989 by J. Michael Bishop and Harold E. Varmus, a discovery which earned these scientists the Nobel prize [Stehelin et al., 1976, Bishop, 1981]. Specifically, genes carrying a mutation or a change of expression that leads to an increased activity and ultimately an increased cell proliferation rate are classified as oncogenes. An example of an oncogene is *HRAS*, a gene which regulates cell division [Tabin et al., 1982, Taparowsky et al., 1982, Sukumar et al., 1983, Cordova-Alarcon et al., 2005]. Often, in a cohort of cancer patients oncogenes accumulate mutations in specific locations, leading to a gain of function phenotype. Therefore, bioinformatics tools like OncodriveClust [Tamborero et al., 2013a] are specialized for predicting driver genes with clusters of mutations, also called mutation hotspots.

Tumor suppressor genes ensure the healthy behavior of cells by either repairing damage or by initiating cell death. If a stop-gain point mutation or a large deletion impairs the tumor suppressor function cells with damaged DNA will still be able to divide. Ultimately, this can lead to uncontrolled cell proliferation and cancer. One of the first discovered examples of tumor suppressor genes is *RB* [Friend et al., 1986]. *RB* prevents excessive cell growth by inhibiting cell cycle progression, and has been described as a cell cycle pacemaker [Weinberg, 1995]. It has an important role during the major G1 checkpoint, namely by blocking S-phase initiation and cell growth. In tumor suppressor genes, mutations causing loss of function or decreased expression allow or promote tumor formation. Opposite to oncogenes, tumor suppressor genes usually follow the "Two-Hit Hypothesis" [Knudson, 1971], where both alleles need to be mutated to allow for malignant transformation. If only one allele is damaged, the second allele would still be able to produce a sufficient amount of the protein to suppress tumor development. However, there are several exceptions from the "Two-Hit" rule, e.g. a heterozygous deletion of *TP53* exhibits a dominant effect and can lead to tumor formation [Willis et al., 2004].

1.3.4 Identification of cancer driver genes

One of the biggest challenges in cancer genomics is to differentiate between driver and passenger mutations, and to narrow down the set of cancer genes. Currently, there are several bioinformatics tools available to differentiate between driver and passenger mutations, most of them using positive selection signatures. The most frequently used signature of positive selection is recurrence, which describes the recurrent observation (occurrence) of mutations in a gene across several patients affected by the same cancer type. One published method, MuSiC, estimates somatic background mutation rates (BMR) and subsequently predicts genes that have significantly more somatic mutations than expected in a given cohort of cancer patients (Figure 1.4) [Dees et al., 2012]. Accurate estimation of somatic mutation rates is difficult, as background mutation rates may differ between genes, tumor types, and patients [Martincorena and Campbell, 2015, Lawrence et al., 2013, Supek and Lehner, 2015]. Along the genome, factors influencing local BMR are chromatin state, replication timing, expression, and transcription factor binding accessibility [Stamatoyannopoulos et al., 2009, Lawrence et al., 2013, Supek and Lehner, 2015, Sabarinathan et al., 2016]. Another tool, MutSigCV, estimates local mutation rates based on several of these features as described in Lawrence et al. [Lawrence et al., 2013].

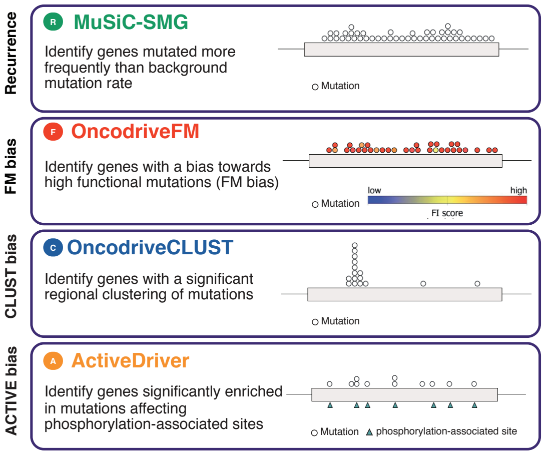
Recurrence
R MuSiC-SMG
Identify genes mutated more frequently than background mutation rate
Mutation
FM bias
F OncodriveFM
Identify genes with a bias towards high functional mutations (FM bias)
Mutation
low
high
FI score
CLUST bias
C OncodriveCLUST
Identify genes with a significant regional clustering of mutations
Mutation
ACTIVE bias
A ActiveDriver
Identify genes significantly enriched in mutations affecting phosphorylation-associated sites
Mutation
phosphorylation-associated site

Fig. 1.4: Signatures of positive selection used to identify driver genes. Reproduced from [Tamborero et al., 2013b]

Illustration of three signatures of positive selection (recurrence, functional impact, clustering) used to identify driver genes and the methods that implement them.

Recently cancer driver identification methods have been developed that do not use mutation recurrence but instead use other signatures of positive selection. For example, OncodriveFM identifies driver genes based on the bias towards high functional impact of mutations in driver genes (Figure 1.4), as mutations contributing to cancer are expected to alter the function of the protein more heavily than mutations not contributing to cancer [Gonzalez-Perez and Lopez-Bigas, 2012]. Another example is OncodriveCLUST [Tamborero et al., 2013a] (Figure 1.4), which identifies driver genes based on the presence of local clusters of mutations, as this indicates that the mutated region has an important function. Studies have shown that clustered mutations often lead to an activation of the gene [Chang et al., 2016]. Similarly, ActiveDriver identifies driver genes significantly enriched for mutations in phosphorylation sites, which are usually associated with important functional domains of a protein (Figure 1.4) [Reimand and Bader, 2013]. All described methods rely on large cohorts of patients sequenced using either WES or WGS to be able to obtain significant results. The TCGA (`https://cancergenome.nih.gov/`) and ICGC (`http://icgc.org/`) consortia have been highly productive during the last 10 years and today provide data of more than 15,000 sequenced tumor and normal samples. TCGA has furthermore published a large bulk of papers describing the landscape of somatic mutations in various cancer types [TCGANetwork, 2008, TCGANetwork, 2011, TCGANetwork, 2012a, TCGANetwork, 2012b, TCGANetwork, 2012c, TCGANetwork, 2013, TCGANetwork, 2015]. Importantly for our study, somatic mutations for all cancer types are available for download in form of MAF files (see Chapter 2 for details), facilitating single or pan-cancer analysis and benchmarking of newly developed algorithms.

In addition to the tools based on point mutations and short InDels, there are other bioinformatics tools exploiting different data types. For example, the HotNet2 method [Leiserson et al., 2015] uses information from pathways or networks of genes to overcome the limitations of single gene testing. Oncodrive-CIS uses copy number changes to infer cancer driver genes [Tamborero et al., 2013c]. Finally, analysis of differential gene expression with e.g. DEseq [Anders and Huber, 2010] or edgeR [Robinson et al., 2010] can be used to identify driver genes. For example, reduced transcription of a tumor suppressor gene caused by a deletion or down-regulation of the gene can increase the fitness of cancer cells, as could the up-regulation of oncogenes. Nonetheless, in recent years many methods have focused on point mutations as their main source of information, mainly due to the rapid increase of freely available exome sequencing data from cancer patients [Weinstein et al., 2013]. But despite the massive efforts of the research community and the availability of several tools we have still not been able to identify the complete catalog of cancer genes and driver alterations [Lawrence et al., 2014]. Therefore, efforts to develop new approaches, combining multiple methods, data sets or cancer signatures, are still in high demand.

Remarkably, all mentioned methods use a frequentist approach to generate p-values for genes. High-throughput sequencing technologies and data coming from big consortium projects like ICGC and TCGA have enabled the community to test all genes or even all regulatory elements, typically reaching above 20,000 tests. With such large test numbers and weak signals (mostly due to cancer intra tumor heterogeneity and high variability between tumor types) frequentist methods are prone to miss true cancer genes after multiple testing correction. Also, frequentist inference does not allow to include prior knowledge in an easy and consistent way, and the combination of multiple signatures is problematic. Moreover, different tools have different stringency. For example, on the TCGA PanCancer dataset MuSic predicts more than 2000 significant genes, while OncodriveClust predicts less than 300 and MutSigCV [Lawrence et al., 2013] only around 100 genes. Hence, in recent years there was an effort to develop methods using Bayesian statistic approaches. Such an example is MADGiC [Korthauer and Kendziorski, 2015], a model based approach that infers cancer driver genes using multiple signatures of positive selection. Although it has been recognized that an ensemble of methods and a combination of cancer signatures improve the recall of cancer genes [Tamborero et al., 2013b] the majority of tools accepted by the cancer community still use a frequentist approach and a single feature (e.g. only recurrence or only clustering etc.) to identify driver genes.

1.3.5 Cancer evolution and heterogeneity

Cancer is a heterogeneous disease. Three main levels of heterogeneity have been described: (*i*) inter-patient heterogeneity, (*ii*) intra-tumor heterogeneity, (*iii*) and metastatic heterogeneity [Vogelstein et al., 2013]. Inter-patient heterogeneity refers to the variability of the same tumor type in different patients and is dependent on multiple factors such as tumor type, age of patients, environmental factors, among others. The tumors of two patients affecting the same tissue, and with a similar phenotype, might still differ substantially on a molecular (genomic) level. These patients may show different driver genes and somatic mutation profiles. Exposure to UV light or tobacco typically results in tumors with large numbers of mutations (high mutation rate) and specific mutational signatures [Alexandrov et al., 2013]. Hence, lung cancer caused by smoking will have a substantially different mutation profile than a lung cancer found in a non-smoker. The mutation profiles found in pediatric cancers (cancer in children) can differ substantially from the respective adult cancer type, and they often exhibit very low somatic mutation rates.

Intra-tumor heterogeneity describes differences within tumor(s) found in the same patient. This type of heterogeneity could also be related to treatment effects, i.e. relapsing tumors are often more aggressive than the original tumor and can have differing mutations. Intra-tumor heterogeneity was described in 1976 when Peter Nowell recognized clonal evolution of cancer [Nowell, 1976]. It has been suggested and later supported that different cancer subpopulations of the same cancer mass have a common ancestral origin, but evolve in a way that they can contain different sub-clonal mutations [Nowell, 1976, Fialkow, 1979, Gerlinger et al., 2012].

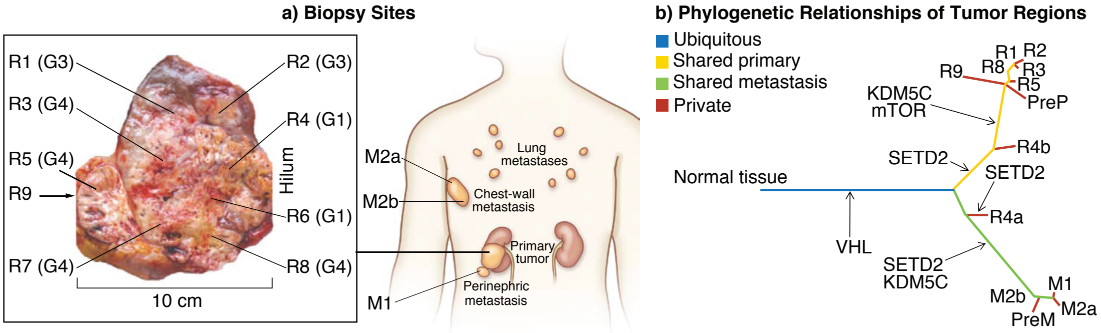

Figure 1.5: Genetic Intra-tumor Heterogeneity and Phylogeny in Patient with renal carcinoma. Reproduced from [Gerlinger et al., 2012].

a) For one patient, multiple sites were sampled and sequenced, including biopsies of the pretreatment primary tumor (PreP) and chest-wall metastasis (PreM), nine primary-tumor regions of the nephrectomy specimen (R1 to R9), three metastasis sites (M1, M2a and M2b), and germline DNA. Regions R6 and R7 were excluded from analyses since only one nonsynonymous variant passed filtering. **b)** Phylogenetic tree of the sequenced tumor regions based on somatic variants showing ancestral relationships of subclones.

Indeed, construction of the phylogenetic tree using DNA sequences obtained from multiple geographical specimens of a solid tumor show common ancestor mutations for the majority of specimens [Gerlinger et al., 2012]. In the example shown in Figure 1.5, mutations in the gene *VHL* were common to all specimens taken from eight locations of one tumor mass. *VHL* is known to regulate apoptosis and was likely the original driver gene leading to malignant transformation. The mutation in *VHL* is called a trunk mutation, with respect to forming the trunk of the phylogenetic tree of the tumor. In addition, all specimens have some private mutations that occurred during tumor development, supporting the hypothesis of intra-tumor heterogeneity and the continuous evolution of tumor cells. The observed data is best described by an asexual evolutionary model similar to bacteria colonies, in which mutations occurring in a single cell of the tumor are purified, if detrimental, or selected, if they increase the fitness of the cell. As private mutations in small subclones are hard to detect, as this would require ultra-deep or large-scale single-cell sequencing, the extent of intra-tumor heterogeneity is usually underestimated. It has been suggested that the number of low-cell-fraction or private mutations in a single cell is exponentially higher than the number of clonal mutations, i.e. the are affecting most cells of the tumor [Williams et al., 2016]. Considering such a large number of low-fraction mutations and hence subclones it is not surprising that tumors can quickly develop a drug resistance after treatment, where one of the many coexisting subclones becomes predominant in the relapsed tumor [Gerlinger and Swanton, 2010, Wu, 2012, Greaves and Maley, 2012]. There is still a debate if the mutations conferring drug resistance typically already exist in a sub-clone prior to treatment, or if they occur after treatment, or if both mechanisms play an important role in the development of treatment resistance. Nonetheless, all evidence supports the hypothesis that tumors are under Darwinian evolution driven primarily by positive selection [Nowell, 1976, Merlo et al., 2006, Pepper et al., 2009].

1.4 Using signatures of selection to identify cancer driver genes

In previous sections we introduced the concept of tumor evolution based on the observation that cancer development shows properties of Darwinian evolution such as signatures of positive selection on mutations beneficial for cancer cell proliferation. Therefore, if we could identify the alterations in cancer genomes that are positively selected we would narrow down list of candidate cancer driver genes. Positive selection in tumor evolution has been measured using multiple approaches such as recurrence, functional impact, and clustering of mutations.

1.4.1 Mutation recurrence as a signature of positive selection

One of the most obvious and widely used signatures of selection is recurrence. Somatic mutations that are recurrently observed in the tumors of a cohort of patients, but cannot be explained by high local mutation rates, can be considered good candidates of causal cancer driver genes for the studied cancer type. Furthermore, to increase statistical power we often look for recurrence of mutations in a genomic locus or functional unit, e.g. a gene, instead of only considering a single position. The biggest challenge for recurrence approaches is to distinguish the true causal mutations from random and non-random noise, technical biases, as well as biological or environmental effects not related to cancer.

When analyzing somatic mutations in tumor tissue we typically do not use a healthy control group for comparison as we only aim to link cancer exclusive mutations to the phenotype. The somatic mutation load of tumors is known to vary strongly depending on tissue type, but also between patients with the same tumor type [Martincorena and Campbell, 2015, Alexandrov et al., 2013]. Furthermore, mutational load as measured in mutations per megabase also varies across the genome [Lawrence et al., 2013, Supek and Lehner, 2015]. Therefore, simple counting of mutations in genes and correcting for the length of genes would result in a large number of false positives, as some genes have a substantially higher background mutation rate than the genome wide average. To infer true driver genes, the standard strategy is to estimate the expected number of mutations in genes with respect to the local background mutation rate (BMR) and then check if the observed number is significantly higher than expected. Obstacles for achieving good estimations of BMR are, for instance, high diversity across patients of a cohort or a low number of somatic mutations in specific genomic regions or for a tumor type in general. Lawrence et al. [Lawrence et al., 2013] have suggested a background mutation rate model, which is used in the tool MutSigCV.

1.4.2 Functional impact bias and mutation clustering as signatures of selection

In addition to recurrence, another successfully exploited signatures of positive selection is the functional impact (FI) bias of variants found in a gene. Various methods have been developed to assess the functional impact of mutations on the protein function [Eilbeck et al., 2017]. FI has been used mostly to (*i*) filter out benign variants, and (*ii*) as measure of unexpected bias. While the former is simply a filter reducing the total number of mutations in consideration, the latter approach uses functional impact bias as a surrogate measure of positive selection. This strategy is based on the hypothesis that the functional impact (damage score) of somatic variants per gene in a cohort follows a specific distribution, when there is no selection. The type of distribution depends on the number of important regions in a gene and the number of possibilities that a variant can significantly change the function of the protein. Any deviation of this expected distribution can be considered a signature of positive or negative selection.

Similarly, the position of variants within a gene can reflect a measure of selection. If there is no selection somatic variants will be distributed uniformly across the gene (assuming that BMR is the same across one gene). Therefore, observing a cluster of mutations in a specific locus of the gene indicates that positive selection of mutations in an important functional site has occurred, unless the predicted variants are false and caused by a systematic calling error. Recently, it has been shown that many tumor types show strong mutational signatures caused by specific environmental effects like smoking or UV light [Alexandrov et al., 2013]. For instance, UV light causes a huge excess of C→T changes due to deamination of methylated cytosines. A strong mutational signature could lead to false recurrence, clustering, or FI bias signatures, i.e. false signatures of selection, if not corrected properly [Martincorena et al., 2017].

1.4.3 Ka/Ks ratio as signature of selection

Ka/Ks is the ratio between the number of nonsynonymous substitutions per nonsynonymous site (Ka) and the number of synonymous substitutions per synonymous site (Ks) that occur during a given time frame in a defined genomic region (usually a gene). Typically, two or more genomes of different species are compared to identify synonymous and nonsynonymous SNPs. The time frame is defined by the last common ancestor. The term Ka/Ks comes from evolutionary biology and it is used to measure rates of evolution [Makalowski and Boguski, 1998]. If a gene has been under purifying selection, it has avoided the accumulation of nonsynonymous mutations and it has conserved his original protein function, therefore the value of Ka/Ks would be less than one. If mutations are happening completely at random and without any selection pressure, the Ka/Ks ratio would be one. Finally, if positive selection acted on variants changing the function of the protein, and leading to a higher fitness, there would be more nonsynonymous mutations than expected, and Ka/Ks would be higher than one. Most genes are subject to purifying selection with selective constraints for nonsynonymous as compared to synonymous mutations (which are considered to be neutral). Therefore, a Ka/Ks ratio below one for most genes when analyzing germline variants has been observed.

Cancer tissues can be seen as a micro-environment where evolution is happening fast, and in an asexual manor. Mutations that are advantageous to the cancer cell will be selected for, while neutral mutations (passengers) are expected to happen following the BMR distribution. Although selective sweeps following the acquisition of a highly advantageous driver mutation will also raise the prevalence of passenger mutations present in the cell, this will not lead to a pattern of linkage disequilibrium, as no cross-over happens during cell divisions. If we generalize the idea of Ka/Ks to a cohort of cancer patients, passengers mutations (both nonsynonymous and synonymous SNVs) are under neutral evolution, while nonsynonymous driver mutations are positively selected. Genes which harbor true driver mutations in a substantial fraction of patients would therefore show a Ka/Ks ratio above one, while most genes not contributing to increased fitness of tumor cells show a Ka/Ks around one.

1.4.4 Mosaicism as a signature of selection

In tumor tissues it is often observed that two or more populations of cells with different genotypes are coexisting (1.5), and we refer to this phenomenon as mosaicism (also termed tumor heterogeneity or sub-clonal tumor structure). This is actually not exclusive to cancer tissues, as we continuously generate cell variants even in healthy (normal) tissue [Fernandez et al., 2016]. In normal tissue, a heterozygous germline mutation is expect to be seen in ~50% of the sequenced reads covering the variant locus. But in case of mosaicism the minor allele frequency (MAF) of the heterozygous mutations diverges more or less from 0.5, depending if it is present in large fraction of the cells ($AF \approx 0.5$) or a small fraction ($AF \ll 0.5$). Identifying mutations that are having very small allele fraction but are not sequencing errors is still challenging and it requires ultra-deep sequencing of tumor tissue. In tumor tissues, if there are no external constraints (e.g. physical barrier), cells with a genotype that gives a selective advantage over other cells will proliferate faster. The mutations that reside in these cells will increase in cancer cell fraction (CCF), i.e. the percent of tumor cells that harbor the mutation increases. Mutations happening very early in tumor development, or that are highly advantageous will be found in the majority of tumor cells and are called clonal mutations. Mutations that are present in smaller fractions of the tumor tissue are called sub-clonal. In diploid loci and at 100% tumor purity (no normal contamination in the tumor sample) the CCF can be easily computed from MAF by multiplying with 2. However, copy number variants or low tumor purity make estimation of CCF from MAF a challenging task, solved by tools as e.g. PyClone [Roth et al., 2014] or Absolute [Carter et al., 2012]. Using low to medium coverage bulk sequencing of tumor tissues will mostly reveal clonal mutations (high MAF), while sub-clonal mutations are missed. However, sub-clonal events are also of great importance during cancer development for several reasons: 1) two or more clones can coexist in symbiosis and one without another would decrease fitness of all sub-clones, 2) after treatment sub-clones can gain (or already have) resistance and become the new dominant clonal population within the changed environment, 3) a sub-clone could become metastatic, 4) high tumor heterogeneity (as measured by the fraction of sub-clones) has been associated with worse treatment outcome and shorter survival. As rates of relapse in cancer are very high, obtaining a fine-grained picture of clonal and sub-clonal tumor structure is an important step on our way to precision oncology. But more importantly for the goals of this study, ignoring sub-clonal mutations results in a loss of statistical power for identification of driver mutations, as the power of basically all available driver prediction tools depends on the number of mutations in the cohort used for prediction.

1.5 Cancer Predisposition

In recent years a wide range of germline mutations that are contributing to the lifetime risk of developing cancer have been identified [Rahman, 2014]. Various strategies have been employed to identify cancer predisposition genes, i.e. genes harboring variants that increase the risk of developing cancer, including the study of families with high cancer prevalence, or Genome-wide association studies (GWAS) using large cohorts of sporadic cases and controls.

Maybe the most widely known examples are germline variants in the genes *BRCA1* or *BRCA2* that dramatically increase the risk of developing breast and ovarian cancer [Miki et al., 1994, Wooster et al., 1995, Ponder, 2001]. It is estimated that 25% of ovarian cancer cases [Stafford et al., 2017] and 15-20% of breast cancer cases [Economopoulou et al., 2015] are due to inherited genetic factors. Mutations in *BRCA1* and *BRCA2* account for about ~25-30% of all familial cases of hereditary breast and ovarian cancer [Nielsen et al., 2016, Siegel et al., 2013]. Although *BRCA1* and *BRCA2* together account for the largest fraction of heritability, more then 25 other genes, most of which are having a function in the same pathways as *BRCA1* and *BRCA2*, have also been implicated with familial breast and/or ovarian cancer susceptibility [Nielsen et al., 2016, Stafford et al., 2017, Economopoulou et al., 2015]. However, even in sum these genes cannot explain all hereditary cases [Couch et al., 2014, Cybulski et al., 2015], a phenomenon termed missing heritability. Therefore, discovery of more genetic factors that can explain the missing heritability for hereditary breast and ovarian cancer is expected in coming years, as well as for other cancer types.

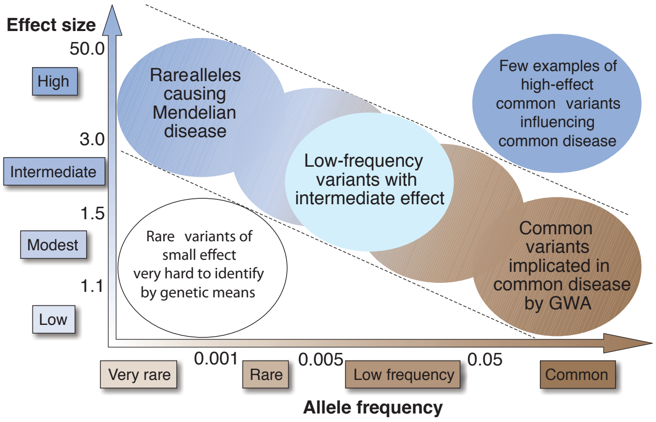
Effect size
50.0
High
3.0
Intermediate
1.5
Modest
1.1
Low
Rarealleles causing Mendelian disease
Low-frequency variants with intermediate effect
Few examples of high-effect common variants influencing common disease
Rare variants of small effect very hard to identify by genetic means
Common variants implicated in common disease by GWA
Very rare
0.001
Rare
0.005
Low frequency
0.05
Common
Allele frequency

Fig. 1.6: Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Reproduced from [Manolio et al., 2009]

Different statistical approaches are developed or adopted to identify risk variants and risk genes, depending on a variants effect size (odds ratio) and allele frequency in the population (Figure 1.6). Very rare variants that segregate in a risk family typically have large effect size (strongly raise the risk to develop cancer) and have most often been identified using linkage analysis or positional cloning [Miki et al., 1994, Kontham et al., 2013]. The 'weaker variants' (or genes) are found less often in familial cases and therefore must be identified using other approaches. Checking for a difference in frequency of candidate genetic variants in large cases-control cohorts is a common approach for detection of variants that have intermediate to low effect size. The most commonly applied method called genome-wide association study (GWAS) facilitated the identification of many common variants that cause a small increase in risk of developing cancer [Varghese and Easton, 2010, Chang et al., 2014]. Similarly, GWAS helped to identify thousands of risk loci for many other genetic diseases (`https://grasp.nhlbi.nih.gov/Overview.aspx` [Eicher et al., 2015]) . However GWAS has limitations, specifically with respect to finding associations using rare and ultra-rare variants. Moreover, GWAS works solely on the level of a single nucleotide and does not support aggregation of variants across functional units such as genes or pathways. In case-control studies is important to have two well-matched groups. Optimally one would use cases and controls from the same population and age group, and with balanced gender. Also, the preparation and sequencing of samples would be done in a single batch using the same technologies, such that technical biases are reduced to a minimum. For all post-sequencing analysis steps (e.g. mutation calling) the same methods and parameters need to be applied. Often, some of these recommendations are not possible to fulfill, and frequently controls are reused for different studies and therefore are not processed using the same technologies and methods as used for cases. Some association tests allow integration of covariates into the model, that can to some extent solve issues with existing biases between cases and controls.

1.5.1 Rare variants and complex diseases

In recent years there have been strong efforts to identify low-frequency and rare cancer risk variants that GWAS is not able to 'pick up'; that at the same time have intermediate or low effect size such that they were also not identified with linkage analysis [Decker et al., 2017, Lin et al., 2017]. Gene-based Rare-Variant Association Studies (RVAS) have proven their potential to identify rare, moderate effect size variants in many complex traits and diseases [Cruchaga et al., 2014, Shtir et al., 2016, Ruiz-Pinto et al., 2017, Nho et al., 2017]. Still, not many cancer risk studies have been exploiting RVAS or RVAS-like methods. Region-based RVAS methods are testing genomic regions or functional units (usually a gene) instead of individual variants to gain statistical power. A variety of RVAS methods has been developed and each has pros and cons depending on the architecture of the disease and the cohort size.

The most straightforward gene-based RVAS tests are burden tests [Li and Leal, 2008, Madsen and Browning, 2009, Morgenthaler and Thilly, 2007]. As the name suggests this is a class of tests that compare number of variants in cases and controls, by collapsing ('aggregating') information for all variants in a focal gene (or defined genomic region) into a single score. This can be done by simply counting the number of minor alleles across all variants that participate in the test. Additionally, variants can be weighted by meaningful biological features (e.g. allele frequency reported in public databases).

The association test statistics is computed between the score and a trait. Numerous burden tests have been proposed differing in how variant aggregation across genes and the association test are performed. Most notable versions of burden tests are collapsing regression models (e.g. CMC [Li and Leal, 2008]), weighted sum methods (e.g. KBAC [Liu and Leal, 2010]) and permutation-based summary count methods (e.g. BURDEN `https://atgu.mgh.harvard.edu/plinkseq/assoc.shtml`). Burden tests are powerful when a large proportion of variants are causal and effects are in the same direction. Therefore, they lose power if the architecture of disease is such that both protective and risk variants are present in a gene, or a large fraction of variants in a gene is neutral.

Contrary, variance component tests (e.g. SKAT [Wu et al., 2011]) do not suffer when variants have different directions (e.g. some variants are protective and some are increasing risk) and/or only a small proportion of the variants in a gene is causal. Rather than collapsing variants, variance component tests evaluate the distribution of the aggregated score test statistics of individual variants. However, variance component tests lose power, compare to burden tests, when the majority of variants have the same direction. Therefore, the SKAT-O test has been proposed that combines burden and variance component tests [Lee et al., 2012b]. SKAT-O computes a weighted linear combination of two tests and weights are automatically estimated from the data. Other new types of association tests have recently been proposed, that can in addition utilize characteristics of variants, such as the MiST test [Sun et al., 2013]). For example, if a gene has no true association with a trait, distributions of functional impact (FI) scores for the variants between cases and controls would be similar. Contrary, if a gene is a true carrier of risk variants, distributions of FI score are expected to be significantly different. Moreover, a risk gene would more likely harbor loss of function (LoF) variants in cases than in controls.

For all RVAS tests it is of great importance to account for possible biases and to have well-matched cases and controls, otherwise tests might fail to reject associations that are not truly related to the trait, but appear due to confounding factors. Most of RVAS methods can account for some biases in the data by having the option to include covariates. Unfortunately, the majority of the available cancer cohorts are lacking well-matched control groups (for instance have not enough healthy controls, controls from a different population, controls sequenced and analyzed with different protocols, etc.).

There is some skepticism in the community if the impact of low-effect risk variants is relevant when compared to risks caused by lifestyle or the environment [Holtzman and Marteau, 2000]. Nonetheless, several successful stories on the hunt for cancer predisposition genes demonstrate that the field is not doomed (yet).

1.6 Objectives

During the last decade big consortia, like The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC), generated rich resources of NGS data for large cohorts of cancer patients from more than 30 tumor types. The trend of sequencing larger and larger cancer cohorts is expected to continue in the near future and one can utilize this valuable data in various ways. The scope of this thesis was to statistically analyze mutations and short insertions and deletions (SNVs and InDels) identified using whole exome and whole genome sequencing of cancer cohorts and healthy controls in order to increase our knowledge about cancer driver genes and cancer predisposition genes. Thus, this thesis work is divided in two parts: (*i*) inference of cancer driver genes using signatures of positive selection, and (*ii*)development of a comprehensive gene-based rare variant association analysis framework for identification of risk genes in case-controls studies.

In the second chapter, we answer how positive selection signatures in tumor evolution can be exploited to obtain a complete landscape of cancer driver genes in various tumor types. We describe and benchmark cDriver, a novel Bayesian inference method developed by the author for identifying cancer driver genes. cDriver identifies and exploits signatures of positive selection in tumor evolution where some of the signatures (e.g. recurrence and functional impact) have been used before by competing methods. However, to the best of our knowledge, cDriver is the first software that exploits mosaicism (tumor heterogeneity) and Ka/Ks as signatures of positive selection to infer cancer driver genes. We have studied several questions, for instance: How to capture mosaicism from NGS data? Are clonal mutations enriched in cancer driver as compared to passenger genes? Are p-values the optimal (or only) way to measure significance? Can we re-purpose ‘strong’ driver genes in one cancer type as evidence (prior knowledge) to increase sensitivity for detection of the same gene as driver in other cancer types? Can we identify gene functions important for tumor development that have previously been under-appreciated? Our novel method and new findings on cancer driver genes have been revised by experts in the field and accepted for publication by Nature Scientific Reports.

In the third chapter, we describe the development of a comprehensive framework for identification of cancer or disease predisposition genes using NGS of case-control cohorts. Taking into account that the majority of known cancer predisposition genes have been found due to common risk variants or variants with high-penetrance effect size, we have focused on rare variant association study (RVAS) approaches, which promise to find genes previously missed. Besides integrating well-established RVAS methods we aimed at implementing and benchmarking a new RVAS approach that can fully utilize variant characteristics (e.g. functional impact), quality control data, population variant allele frequencies and clinical information. Additionally, we addressed problems arising from insufficiently matched cases and controls, or other sources of biases between cases and controls, and provide solutions to some degree. Finally, we addressed the issue of population stratification for studying populations that are not well represented in public variant allele frequency databases using the example of Iberian case-control cohorts. A manuscript describing the REWAS framework, the novel BATI association test methods and benchmarking results for five RVAS methods is in preparation.

Chapter 2

INFERRING CANCER DRIVER GENES

Zapata, L.*, Susak, H.*, Drechsel, O., Friedländer, M., Estivill, X., Ossowski, S. (2016). Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Scientific Reports (accepted)*

Zapata L, Susak H, Drechsel O, Friedländer MR, Estivill X, Ossowski S. Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. Sci Rep. 2017 Dec 13;7(1):13124. DOI: 10.1038/s41598-017-12888-1

Chapter 3

IDENTIFYING CANCER PREDISPOSITION GENES USING RARE VARIANT ASSOCIATION STUDIES

Susak, H.*, Escaramis-Babiano, G.*, Serra-Saurina L., Bosio, M., Rabionet, K., Domenech-Salgado, L., Ozkan, S., Estivill, X., Ossowski S., Bayesian Rare Variant Association Test using Integrated Nested Laplace Approximation. *(In preparation)*

Bayesian Rare Variant Association Test using Integrated Nested Laplace Approximation

Hana Susak*, Georgia Escaramis-Babiano*, Laura Serra-Saurina, Mattia Bosio, Kelly Rabionet, Laura Domenech-Salgado, Selen Ozkan, Xavier Estivill, and Stephan Ossowski

* *These authors contributed equally*

3.1 Background

Rapid growth of next generation sequencing (NGS) technology and the dramatically improved cost-effect ratio are changing the landscape of medical and human genetics research, providing a unique opportunity to study the association of all types of genetic variants with complex diseases at a genome-wide scale. Other than genome-wide association studies (GWAS) that are based on counting of genotypes at predefined genomic positions with alternative alleles of high minor allele frequency in the population (MAF >5 %), new sequencing technologies enable us to study rare genetic variants (RV) across the whole genome. Rare variant association studies promise to identify novel disease genes based on the enrichment of rare variants with low to medium effect size by aggregation of genotypes across genes or other functional units, thereby allowing to fill the gap left by GWAS studies and to address the phenomenon of missing heritability.

It has been previously shown that RVs play an important role in complex genetic disease etiology [Cohen et al., 2004, Chassaing et al., 2016, Priest et al., 2016, Tan et al., 2017]. Furthermore, it has been demonstrated that RVs are more likely than common variants to affect structure, stability and function of proteins [Tennessen et al., 2012, Nelson et al., 2012]. Therefore, statistical analysis of the combined set of rare variants across all genes or regulatory elements promises to reveal new insights into the genetic heritability of complex diseases and cancer.

One of the major difficulties of rare variant association studies (RVAS) is the lack of power when using traditional statistical methods like GWAS. Given that few individuals are carriers of the rare alternative allele, association studies based on single variant positions would require extremely large sample sizes. To overcome this obstacle and in order to increase statistical power, studies of RV consider simultaneously multiple variants within functional biological units, such as genes, promoters or pathways, for association to disease. Different statistical methods have been proposed recently that address the problem of aggregated analysis of rare variants in functional units. For example, score based methods pool minor alleles per unit into a measure of burden, and then this burden score is used for association with a disease or phenotypic trait [Li and Leal, 2008, Price et al., 2010, Madsen and Browning, 2009, Liu and Leal, 2010]. Burden tests are powerful when a high proportion of RVs found in a gene are deleterious or at least their effect on the disease are one-sided, i.e. either protective or deleterious. But this is rarely the case since usually few deleterious variants coexist with many neutral and possibly some protective variants. Hence, other methods have been developed that consider heterogeneous effects among RVs on the disease (or trait), which are mainly based on variance-component tests, e.g. SKAT and C-alpha [Wu et al., 2011, Neale et al., 2011]. These methods are more powerful than burden tests when the hypothesis of unidirectional effects does not hold. More recently, methods have been developed that contemplate the possibility that both types of genetic architectures may coexist throughout the genome, by being constructed as a linear combination between burden and variance-component tests, e.g. SKAT-O [Lee et al., 2012b]. Following this idea [Sun et al., 2013] developed a mixed effects test (MiST) within the framework of a hierarchical model, for which they further considered the possibility of incorporating biological characteristics of the variants into the statistical approach. Thus, the hierarchical method is implemented in the way that individual variants are assumed to be independently distributed, with the mean modeled as a function of variant characteristics and certain variance that accounts for variant heterogeneous effects. The resulting model is a type of generalized linear mixed effects model (GLMM), were variant-specific effects are treated as the random part of the model and patient and variant characteristics as fixed part. The authors claim that under the assumption that associated variants share common characteristics such as similar impact on protein function (e.g. primarily loss of function), using this prior information in the test increases power. [Sun et al., 2013]further argue that attempting to estimate the full model for inference purposes requires multiple integration, such that it becomes computationally intensive in a genome-wide scan. Instead, they propose a score test under the null hypothesis of no association avoiding multiple integration.

Building on the ideas of [Sun et al., 2013], but with the motivation of making inference based on full model estimation, we propose a Bayesian alternative to the GLMM, which is based on Integrated Nested Laplace Approximation (INLA)[Rue et al., 2009]. In a

Bayesian framework, maximizing likelihoods for model estimation becomes unfeasible in complex data structures, hence, the traditional model estimation procedure is Markov Chain Monte Carlo (MCMC) [Sorensen and Gianola, 2007]. MCMC is a very flexible approach that can be used to make inference for any Bayesian model. However, evaluating the algorithm's performance in terms of convergence is not straightforward [Cowles and Carlin, 1996]. Another concern with MCMC is extensive computation time, especially in large-scale analyses such as genome-wide scans. INLA is a non-sampling based numerical approximation procedure, developed to estimate hierarchical latent Gaussian Markov random field models. Being based on numerical approaches instead of simulations makes INLA substantially faster than MCMC. Furthermore, [Rue and Martino, 2007] show for several models that INLA is also more accurate than MCMC when given the same computational resources. The flexibility of modeling within the Bayesian framework combined with rapid inference approaches opens new possibilities for genetic association testing. Here, we present a novel Bayesian rare variant Association Test using INLA (BATI), implemented as part of the REWAS framework. We demonstrate using realistic benchmark tests that BATI outperforms existing methods, including Burden, SKAT-O, KBAC and MiST, if categorical or numerical data on the effect of variants on protein function is available. We further suggest how to use 'difference in deviance information criterion' (DIC) for model selection.
Furthermore we describe the 'Rare variant Exome Wide Association Study' (REWAS) framework, which combines all steps required for RVAS, including quality control (QC), population stratification and functional variant annotation, and integrates four commonly used RVAS test methods (Burden, SKAT-O, KBAC, MiST) as well as the novel BATI test.

3.2 Results

3.2.1 A novel test statistic and a comprehensive framework for RVAS

We developed the 'Rare variant Exome Wide Association Study' (REWAS) framework, an all-in-one tool designed for RVAS analysis using case-control cohorts. REWAS supports rare variant association using genes or any other biological units such as promoters or enhancers. It provides all essential steps and functionalities to perform the complete analysis of whole-exome sequencing (WES) or whole-genome sequencing (WGS) based case-control study designs: (*1*) facilitates comprehensive quality control and filtering, (*2*) enables user created patient-based and/or variant-based characteristics in an easy and intuitive fashion, (*3*) integrates five conceptually different rare-variant association methods, and (*4*) provides a novel approach to address population stratification. It is implemented in a modular way and provides great flexibility, allowing to analyze a wide range of association study designs. Figure 3.1 shows a flowchart of the REWAS framework.

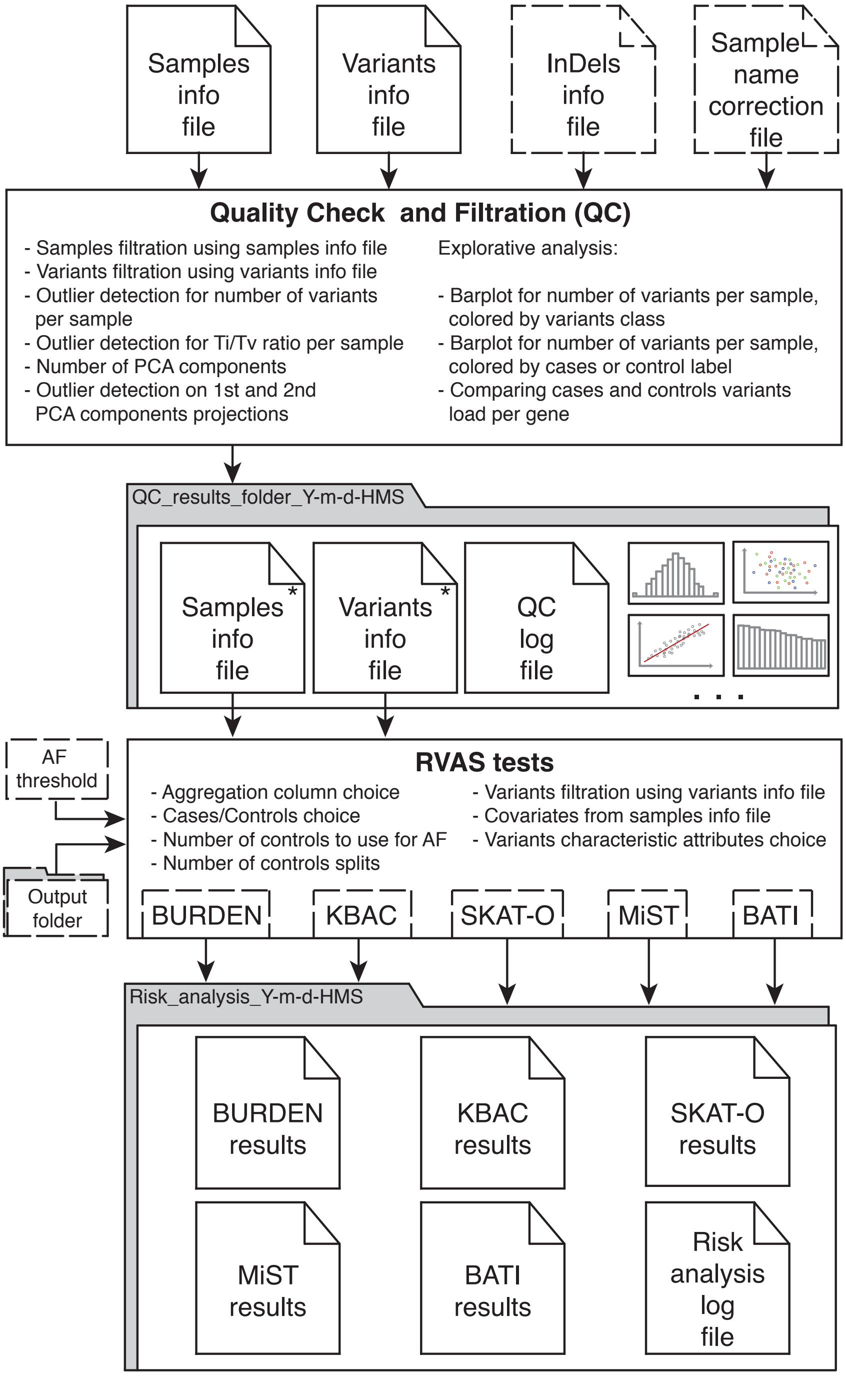
Samples info file
Variants info file
InDels info file
Sample name correction file
Quality Check and Filtration (QC)
- Samples filtration using samples info file
- Variants filtration using variants info file
- Outlier detection for number of variants per sample
- Outlier detection for Ti/Tv ratio per sample
- Number of PCA components
- Outlier detection on 1st and 2nd PCA components projections
Explorative analysis:
- Barplot for number of variants per sample, colored by variants class
- Barplot for number of variants per sample, colored by cases or control label
- Comparing cases and controls variants load per gene
QC_results_folder_Y-m-d-HMS
Samples info file *
Variants info file *
QC log file
. . .
AF threshold
Output folder
RVAS tests
- Aggregation column choice
- Cases/Controls choice
- Number of controls to use for AF
- Number of controls splits
- Variants filtration using variants info file
- Covariates from samples info file
- Variants characteristic attributes choice
BURDEN
KBAC
SKAT-O
MiST
BATI
Risk_analysis_Y-m-d-HMS
BURDEN results
KBAC results
SKAT-O results
MiST results
BATI results
Risk analysis log file

Figure 3.1: REWAS framework summary.

Three methods, KBAC, SKAT-O, and MiST, were chosen to be included in the REWAS framework due to their superior performance compared to eight other RVAS methods in a recent benchmark study by Moutsianas et al.[Moutsianas et al., 2015]. In addition we included the classical Burden test as a baseline and for representing the most simplistic and intuitive form of RVAS tests. Finally, we developed a new association test, Bayesian RVAS Test using INLA (BATI), which better leverages available biological information (numerical or categorical) such as functional impact of variants. Despite being a Bayesian inference method BATI is fast and requires reasonable computational resources. This is achieved by using Integrated Nested Laplace Approximation (INLA) instead of the computationally demanding Markov chain Monte Carlo (MCMC) approach for estimating parameters (see Methods 3.4). MiST and BATI have same theoretical postulation, however, BATI approximates the full model parameters that can further help understanding disease architecture, while MiST obtains score statistics only for the null model. Note that only these two methods (from five included in REWAS) can benefit from biological variant characteristics (e.g. functional impact cores, missense vs. LoF etc.). We have extensively benchmarked Burden, KBAC, SKAT-O, MiST and BATI demonstrating that BATI outperforms other approaches in WES-based RVAS if functional characteristics of variants are available.

3.2.2 Benchmarking RVAS Tests Using Simulated Genetic Disease Architectures

In order to benchmark BATI and four competing RVAS tests, Burden, SKAT-O, KBAC and MiST, we chose a previously published benchmark dataset [Moutsianas et al., 2015] in which 6 genetic disease architectures were simulated (see Methods 3.4). Each of the architectures consist of 24 or 25 genes with 100 sets of simulated disease associated variants. Using our REWAS framework we applied all five RVAS methods to each of the six simulated architectures. As the simulated datasets were devoid of measurable noise, technical biases or population stratification, we did not perform any QC or filtering steps available in REWAS. Furthermore, we did not use variant-specific characteristics such as functional impact, as variants were introduced randomly, and therefore do not have meaningful biological nor evolutionary context.

False Positive Rate estimates

To properly evaluate the statistical power of each method we first estimated appropriate significance thresholds for each method producing comparable type I errors. Candidate genes are evaluated using p-values in BURDEN, KBAC, SKAT-O and MiST, and DIC difference value is used in BATI, further referred simply as DIC (see Methods 3.4). We estimated significance thresholds for p-values and DIC at three levels: (*i*) 5% expected false positives rate (FPR), (*ii*) 0.1% expected FPR, and (*iii*) 0.01% expected

FPR. Genome-wide rare variant association tests consider around 20,000 genes as candidates. Hence, a FPR of 0.01% would result in approximately 2 false positive gene associations (20 false positives (FPs) for 0.1% and 1000 FPs for 5%). To estimate an appropriate threshold we permuted the labels for cases and controls 100 times for each of the six datasets, resulting in 240,000 permutation tests per architecture. (24 genes x 100 simulations x 100 permutations of case and control labels), none of which should be significant. Distributions of obtained p-values and DICs for each of the architectures are shown in Supplementary Figure S3.1. Table 3.1 shows the median of estimated thresholds from 6 architectures per test. Figure 3.2 shows estimated thresholds for each architecture-method pair. As expected, we found that in most cases estimated p-value thresholds are close to the specified FPR we aimed to obtain. To evaluate and compare the power of the different statistical tests we used the thresholds listed in Table 3.1 and labeled a gene as significant according to them in each method.

Method	0.05 FPR	0.001 FPR	1e-04 FPR
BURDEN	4.941e-02	9.124e-04	7.981e-05
KBAC	4.299e-02	7.809e-04	7.574e-05
SKAT-O	5.084e-02	9.319e-04	7.049e-05
MiST	5.265e-02	1.083e-03	9.776e-05
BATI	3.866	14.54	22.47

Table 3.1: AR1–6 P-value and DIC threshold estimates for 3 FPR levels.

FPR stands for False Positive Rate

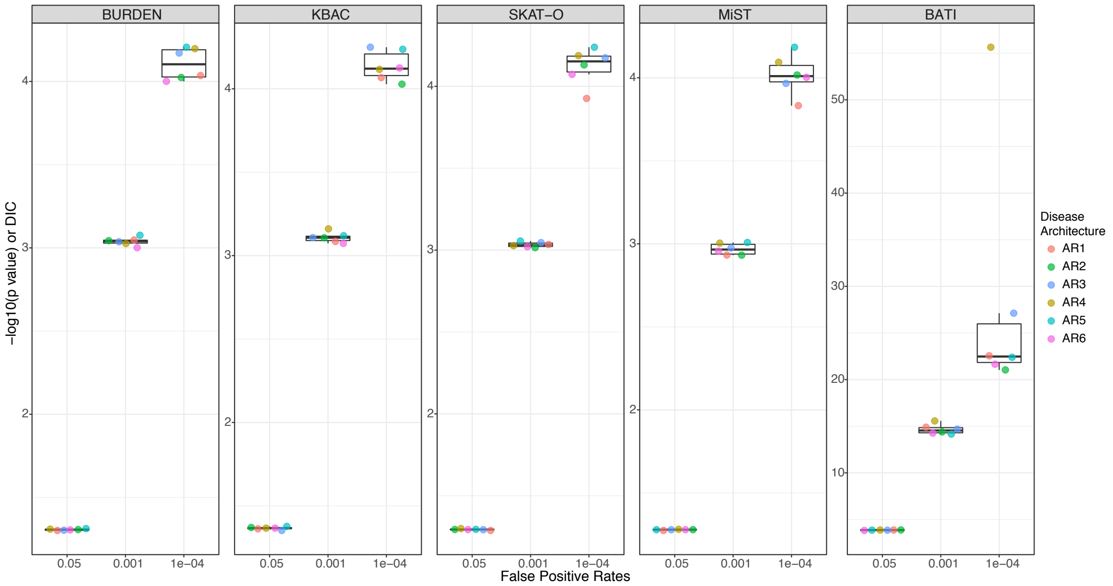

Figure 3.2: Boxplot for P-value and DIC thresholds estimated on 3 FPR levels.

Each boxplot is representing p-value or DIC value estimates for six disease architecturess

Power analysis

We calculated power for each architecture-method pair by counting how many times each of the 24 (or 25) simulated disease genes was significant within the 100 simulations. Genes are considered significant if their p-value generated by the benchmarked method (or DIC value in case of BATI) was bellow (for DIC above) the corresponding threshold listed in Table 3.1.

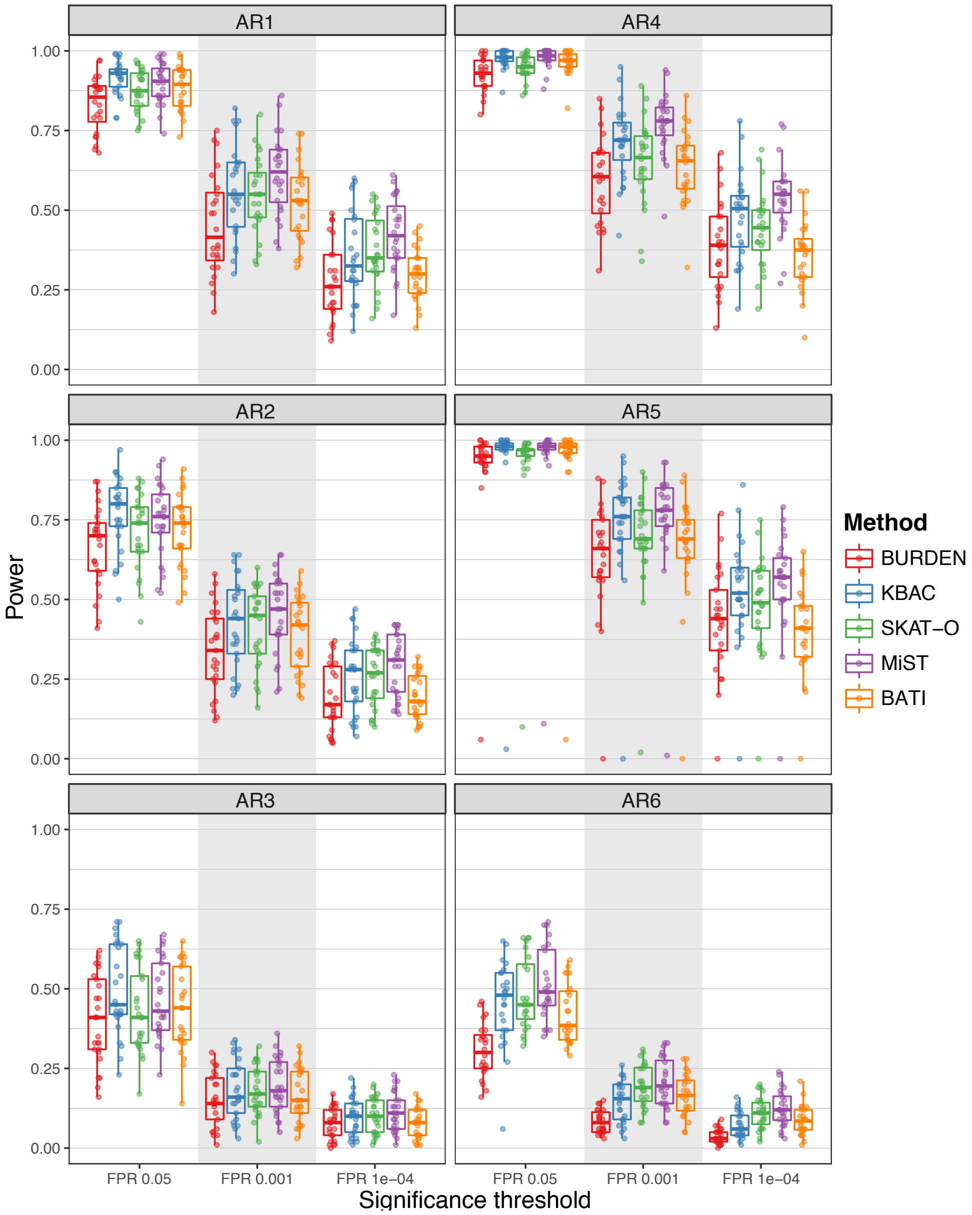

Figure 3.3: Boxplot for power of methods for 3 FPR levels.

Each dot in plots represent a gene, and y axis value is a fraction of the simulations in which the gene was called as significant.

Figure 3.3 shows the estimated statistical power for each method and architecture on three FPR levels. All methods performed similar with MiST being the best in almost all architectures. SKAT-O performed slightly better than MiST on architectures 1, 2, and 3, but only at 0.05 FPR level. As expected, BATI performs comparably, but slightly worse than SKAT-O in this random simulation benchmark, as no biological features (neither categorical nor numerical) were available, eliminating one of the main strengths of the BATI test. We therefore developed a new, more realistic simulation for benchmarking of RVAS tests with the goal to utilize real exome data, real disease variants and biological context information, as described next.

3.2.3 A realistic simulation of whole-exome sequencing based case-control studies

Simulated genetic architectures for benchmarking of disease association tests have been based on randomized variants in small subsets of genes, with population variants sampled from e.g. HapMap. They lack the realistic distribution of variants, background noise and false positive variant calls found in real whole-exome sequencing (WES) data of hundreds of individuals from a typical disease study cohort. But more importantly, the random variants introduced as 'disease-causing variants' lack the biological features of real disease-causing variants, e.g. they do not necessarily change the function of the protein. Therefore, we developed a new disease cohort simulator combining 1167 WES datasets from various real cohorts, in which we introduce known breast cancer predisposition variants found in ClinVar. We simulated a breast cancer risk cohort by introducing risk variants in six genes: *BRCA1, BRCA2, PALB2, BRIP1, CHEK2* and *BARD1* (see Methods 3.4). Cohorts included in this simulated cohort, as well as the target enrichment kits used for each study are shown in Supplementary Table S3.1.

In total we had more than 2,194 WES samples available to form the simulation cohort, which were sequenced at CNAG–CRG Barcelona during 2011 and 2016. However, many of these samples were not from individuals of Spanish descent, some libraries were prepared with target enrichment kits other than Agilent SureSelect 51 or 70, or Nimblegen SeqEz v3 or had low quality. To form a high quality cohort containing individuals with highly similar genetic background we therefore used the quality control (QC) modules of REWAS (see Methods 3.4) to select 1167 samples meeting the QC criteria. For benchmarking purposes we only considered variants in regions that are targeted by all three enrichment kits. We further observed that a small subset of regions that supposed to be targeted consistently showed low coverage in a kit-specific manor, leading to strong biases as shown by PCA (Supplementary Fig. S3.2a). We solved this issue by excluding regions with less than 10x average coverage in at least one kit (Supplementary Fig. S3.2b).

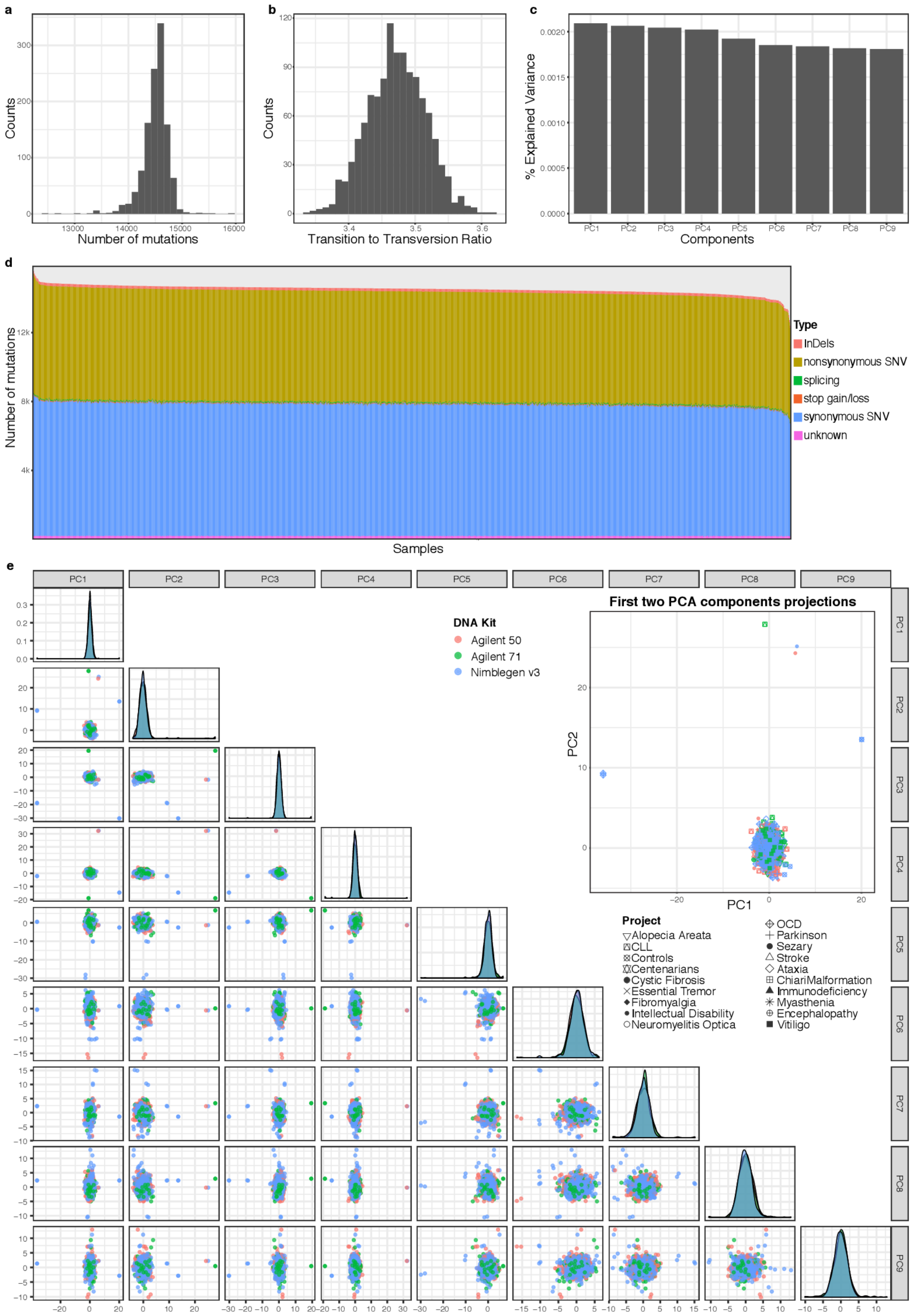

Figure 3.4: QC plots for 1167 Samples used for simulation

All figures in the panel are after removing outliers. **a** Histogram for number of mutations per sample. **b** Histogram for Ti/Tv ratio per sample. **c** Percentage of explained variance on first 9 PCA components. **d** Bar-plot for number of mutations per sample, colored by mutation's classification. **e** Projection on first nine PCA components, with zoom-in of first two PCA components projection. Samples are colored by DNA analysis kit, and in zoom-in plot samples shape correspond to the project.

Samples included in the final simulation cohort show no biases in any of the first nine components of the PCA (Figure 3.4e), and the explained variance per PCA component is low (Figure 3.4c). Furthermore, samples in the cohort show a normal distribution of the number of mutations (Figure 3.4a) and Ti/Tv ratio (Figure 3.4b), and show no bias in the fractions of InDels and synonymous, nonsynonymous or LoF SNVs (Figure 3.4d).

ClinVar variants labeled as breast cancer (BRCA) risk variants have been introduced such that three realistic complex disease architectures are obtained in which for each gene we aimed for 0.5%, 1% or 2% of the phenotypic variance explained (VE), respectively (see Methods 3.4). For some of the genes desired levels of VE were not reached in fraction of simulations. The reason is that some genes had low number of candidate ClinVar variants and possible 'unlucky' samplings for variants relative risk. Exact levels of reached VE for each architectures and gene are shown in Figure 3.5.

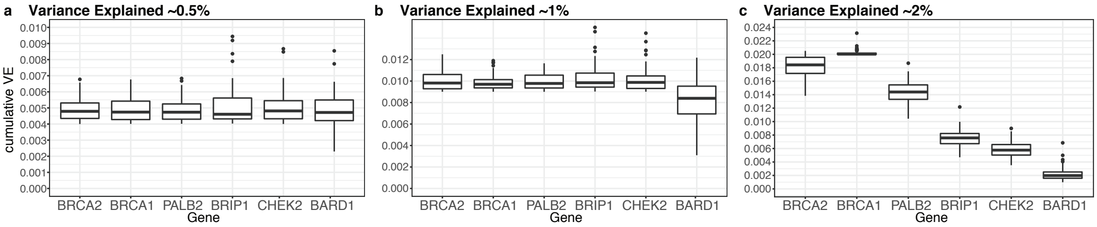

Figure 3.5: Phenotypic Variance Explained, for the six gene after introducing risk variants.

Every boxplot in the Figure is created from cumulative VE for 100 simulations (done for each gene). For some genes desired levels of VE (**a** 0.5%, **b** 1% and **c** 2%) are not reached in all 100 simulations because of low number of candidate risk variants.

ClinVar variants with a MAF between 0 and 0.01 in public databases (EVS, ExAC, and 1000 Genome project) were used as the pool for sampling risk variants. An overview of the total number of ClinVar risk variants available for the simulations per gene is shown in Table 3.2.

Gene	# of candidate variants	# of indels	# of splicing variants	# of stopgain SNVs	# of nonsynonymous SNVs
BRCA2	1213	825	44	283	61
BRCA1	1037	656	65	262	54
PALB2	49	31	0	15	3
BRIP1	25	11	1	9	4
CHEK2	17	8	1	7	1
BARD1	7	1	0	5	1

Table 3.2: ClinVar BRCA risk variants used for simulatation

Relative risk (RR) of introduced variants was sampled from a distribution that reflects the behavior such that higher RR is more common for variants with lower MAF (Supplementary Fig. S3.3, see Methods 3.4). The simulation procedure was repeated 100 times for each gene to allow for power calculation. As we are using real WES data, primarily from Spanish population cohorts, the simulation cohort already contained variants in the six genes prior to the introduction of ClinVar risk variants, which are listed in Table 3.3. These pre-existing variants are not expected to be associated to cancer risk, however we cannot completely rule out this possibility.

Gene	total # of mutations	# of possible cases	# of possible controls	# of mutations in cases	# of mutations in controls	# of affected cases	# of affected controls
BRCA2	54	387	776	21	42	30	60
BRCA1	30	389	777	16	20	24	37
PALB2	14	386	773	6	12	10	16
BRIP1	17	386	776	8	12	15	17
CHEK2	14	389	778	6	11	7	14
BARD1	9	389	778	3	7	4	15

Table 3.3: Number of mutations in six BRCA risk genes in the cohort before introducing any causal mutations.

Only mutations participating in testing (the one after filtering by multiple criteria, like MAF<0.01, no synonymous, etc.) are counted. In few samples some positions could not be called, therefore number of possible cases and controls that could have a variant(s) can be lower then total number of cases (389) and controls (778).

In order to simulate a case-control study for benchmarking of RVAS methods we randomly split the 1,167 samples in one third cases (389 samples) and two third controls (778 samples). Interestingly the random split let to a relatively high fraction of controls with pre-existing *BRCA2* variants (found in the WES data prior to introducing ClinVar variants), coinciding with a generally high number of rare variants in *BRCA2* (Table 3.3), which made RVAS for this gene specifically difficult for all benchmarked methods. Otherwise case and control cohorts showed no biases in number of variants per sample (Supplementary Fig. S3.4a) or number of variants per gene (Supplementary Figure S3.4b and c) (Note: this QC plot was performed using a different case-control split with chronic lymphocytic leukemia samples as cases. I will redo this plot for the final paper using the correct cohort split.)

3.2.4 Benchmarking RVAS Tests Using WES cohorts with ClinVar BRCA risk variants

Similar to the benchmark test using simulated disease architectures with randomly introduced variants described in 3.2.2 we used the REWAS framework to benchmark the five RVAS tests Burden, SKAT-O, KBAC, MiST and BATI on the new WES simulation cohort, described in the previous paragraph.

False Positive Rate estimates

To properly estimate the significance thresholds for the WES cohort we randomly chose one third of the samples as cases and two third as controls using the baseline WES dataset (without introduced ClinVar risk variants, see Methods 3.4), and performed RVAS with all 5 tests. By doing so we are considering that significant associations should only be found by random chance and therefore we can adjust proper thresholds for desired false positive rates (FPRs).

Prior to RVAS we filtered out all variants that had European AF>0.01 in any of the databases: 1) ExAC, 2) EVS, and 3) 1000 Genome Project. Additionally, all variants that in controls had AF>0.01 were removed. Furthermore, all variants that were annotated as synonymous or had CADD score bellow 10 (likely benign) were removed. For BATI and MiST we used CADD scores as numeric characteristic for variants and exonic function (missense, loss-of-function, frameshift) as categorical characteristic for variants. We repeated the whole process (from case-control sub-sampling to RVAS) 10 times.

We noticed that MiST has inflated zero p-values, i.e. many genes had a p-value of exactly zero, even in these randomized case-control cohorts where no gene should be significantly associated to any trait (Figure 3.6c). We found that these unexpected zero p-values occur exclusively for genes with few variants ($\lessapprox$10) across the cohort. Hence, we removed all genes with p-value 0 from MiST results (Figure 3.6d). All other methods behaved similar as in the first benchmark (six simulated disease architectures) and did not show the zero p-value artefact or unexpectedly high DIC values (Figure 3.6 and Supplementary Fig. S3.1).

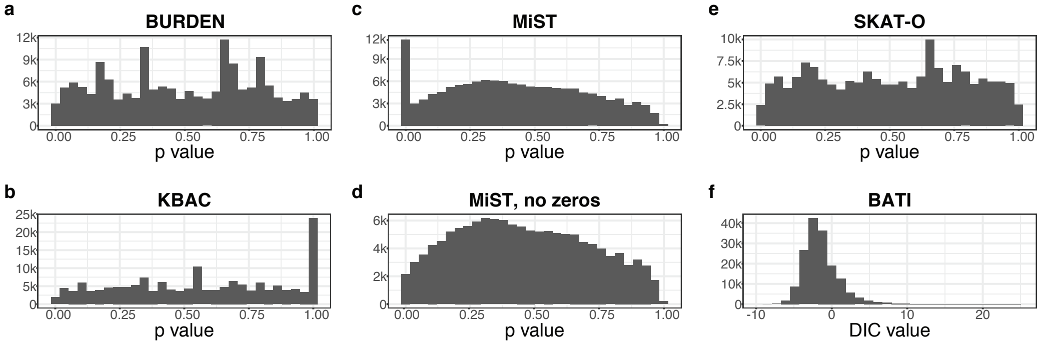

Figure 3.6: Distribution of p-values and DIC values on baseline WES data with randomized cases and controls.

Histograms are done on aggregation of p-values or DIC values from tests on 10 randomized datasets (each consisting of ~17000 genes), where no gene is expected to be significant.

Next, we calculated on each of these 10 RVAS results the p-value significance threshold for BURDEN, KBAC, SKAT-O and MiST methods to achieve 5%, 0.1% and 0.01% FPRs. Similarly, for BATI we calculated DIC significance thresholds for obtaining the same FPR levels. Estimated thresholds across the 10 randomized case-control datasets are highly similar (Figure 3.7). At 0.01% FPR estimated thresholds we have only 2 genes above the threshold, and therefore observed small fluctuation of estimated significance thresholds across 10 datasets is not surprising (note that p-values axis is in log scale in Figure 3.7). Therefore, we used the median as thresholds for subsequent power analyses (see Table 3.4).

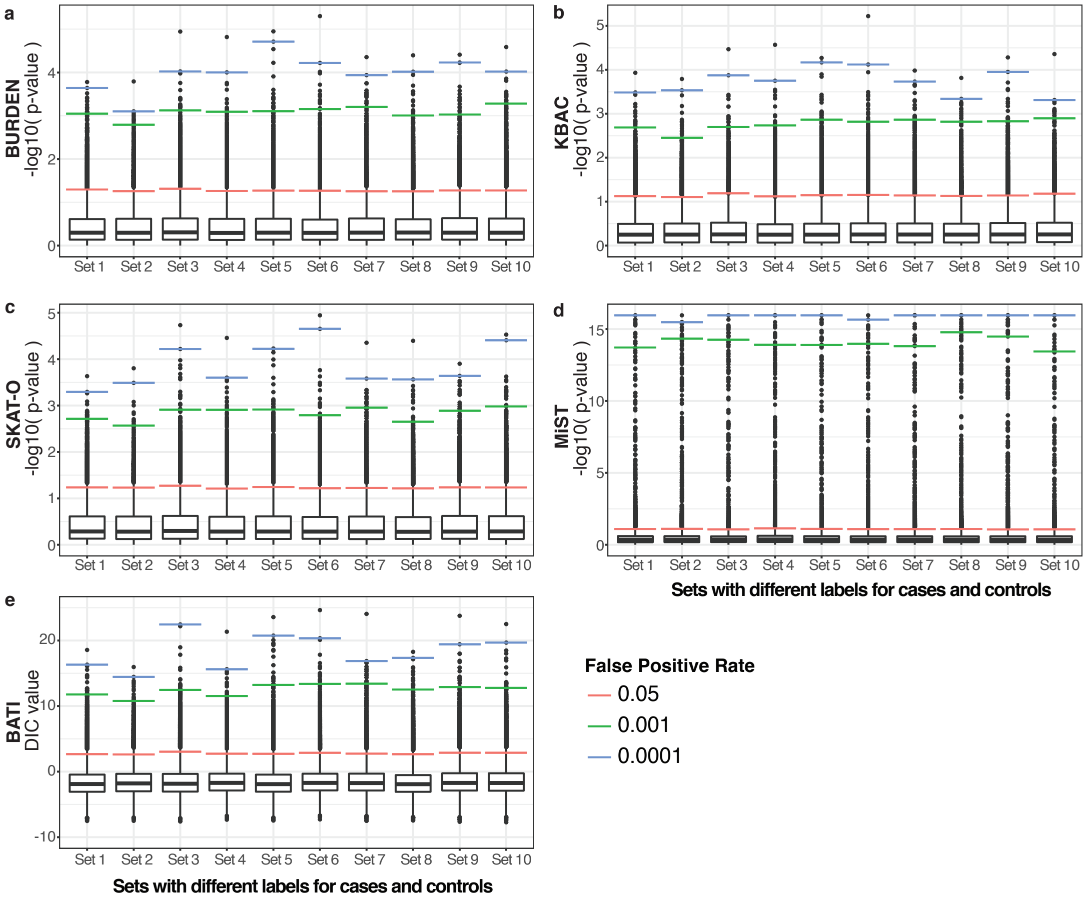

Figure 3.7: Boxplots for p-values and DICs for 10 datasets with random case-control label.

Boxplots of p-values for **a)** BURDEN, **b)** KBAC, **c)** SKAT-O, **d)** MiST methods and of DIC values for **e)** BATI. Samples in all 10 datasets had randomly assign case or control labels, with one third of samples as cases. In each boxplot lines mark 5% (red), 0.1% (green) and 0.01% (blue) significance levels.

Method	0.05 FPR	0.001 FPR	1e-04 FPR
BURDEN	5.392e-02	7.984e-04	9.602e-05
KBAC	7.229e-02	1.520e-03	1.813e-04
SKAT-O	5.834e-02	1.265e-03	2.400e-04
MiST	8.077e-02	1.160e-14	1.110e-16
BATI	2.738e+00	12.64	18.38e

Table 3.4: P-value and DIC threshold estimates for 3 FDR levels on Exome Sequencing dataset. FPR stands for False Positive Rate

Power analysis and Type-1 Error

We applied each of the benchmarked methods to each of the three simulated WES datasets (including ClinVar BRCA risk variants with cumulative variance explained (VE) of ~0.5%, ~1% and ~2% per risk gene, respectively), using the REWAS framework as described for the first benchmark above. Significance thresholds for power calculation at 5%, 0.1% and 0.01% FPR were chosen for each method as described above. We observed that all methods perform well (power close to 100% for all six simulated risk genes) if VE is high (2%) and expected FPR is 5% (Figure 3.8c, left). However, all methods except for BATI and MiST show reduced power at the same significance level, but with lower VE (see 3.8a and b for VE of 1% and 0.5%, respectively).

We note that all five methods failed to detect *BRCA2* at low VE (Figure 3.8a, left) due to the high number of variants in this gene found in our original WES cohort used for the simulation (90 samples affected, see Table 3.3). For comparison, a VE of 0.5%, 1% and 2% translate to approximately 10, 20 and 40 cases affected by a damaging variant, respectively. We conclude that genes harboring large numbers of rare variants in the population pose a substantial problem for RVAS tests, which could potentially be overcome by better annotation of functional impact of variants in the future.

We observed greater differences in performance between the five RVAS tests for lower FPR thresholds (0.1% and 0.01%) and at lower VE (1% and 0.5%). Specifically, the MiST method failed to identify any gene at lower FPR, which is especially surprising considering the good performance of MiST in the randomly generated disease architectures used in the first benchmark. Importantly, at the lowest FPR level of 0.01% only BATI is able to identify some of the risk genes in a large fraction of simulations (except for *BRCA2*, see Figure 3.8a-c, right). Similarly, at medium FPR level only BATI can identify risk genes reliably if VE is 0.5% and power is ~100% when VE is increased to 1% (again, excluding *BRCA2*, which cannot be identified by any method at these thresholds).

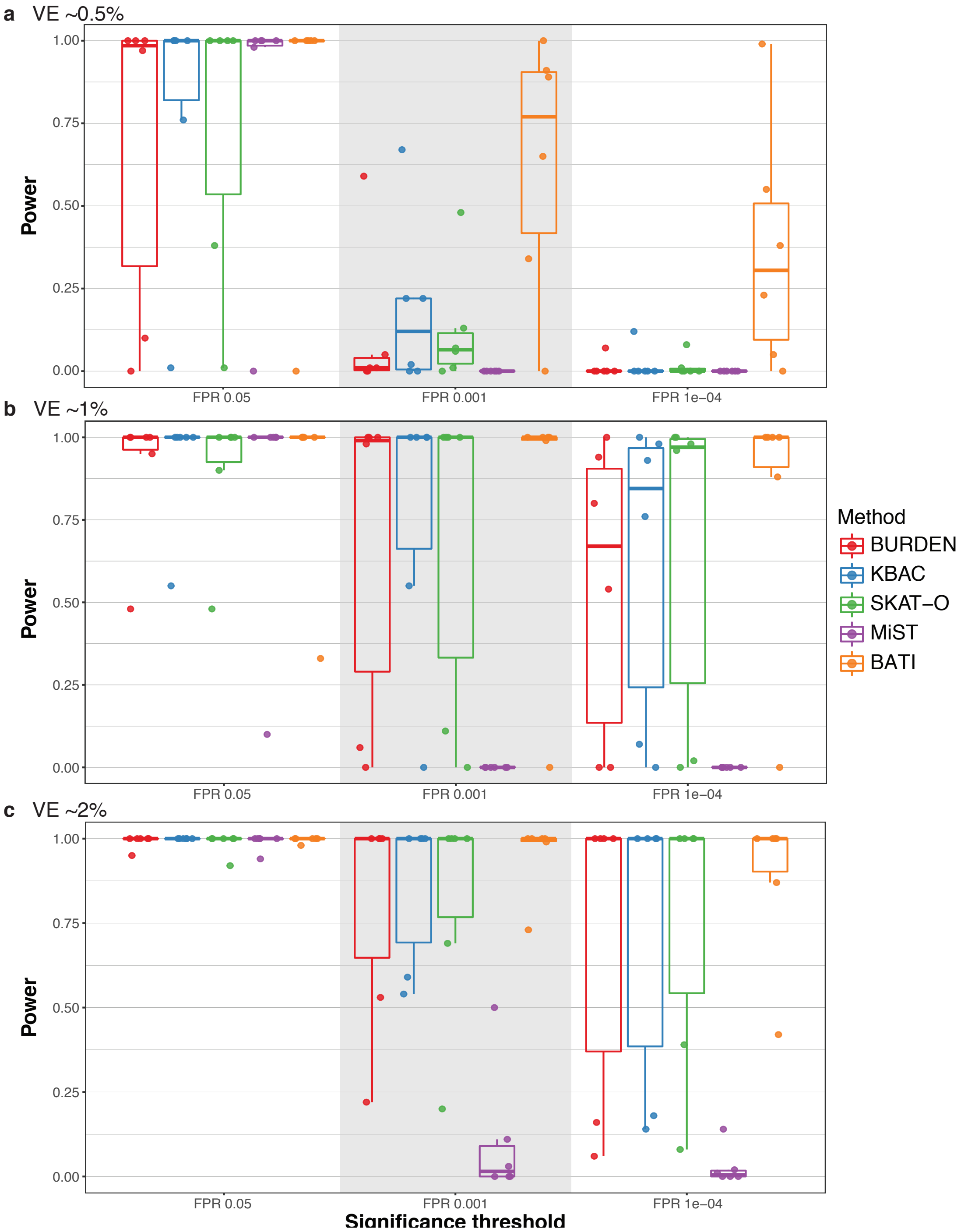

Figure 3.8: Boxplot for power of methods for 3 FPR levels.

Each dot in the plots represents a gene, and y axis value shows the fraction of the simulations in which the gene was called as significant. In **a** are shown results with simulations where variance explained is ~0.5%, in **b** with simulations where variance explained is ~1%, and in **c** with simulations where variance explained is ~2%.

Figure 3.9 shows the distribution of p-values (or DICs for BATI) across the 100 simulations for each of the six risk genes. We observe that BATI reaches significance levels for each of the tested FPR levels and each of the VE levels for 5 out of 6 risk genes (close to 100% of tests when VE is 1% and 2%, but only ~50% of tests when VE is 0.5%). BURDEN, KBAC and SKAT-O performed well for 4 out of 6 genes, *BRCA1, PALB2, BRIP1* and *CHEK2*, but have problems with *BARD1*. This is not surprising because for most of the simulations *BARD1* did not reach the desired level of VE, due to a low number of candidate risk variants found in ClinVar (see Methods 3.4). BATI did not perform worse for BARD1, indicating that the method can handle even lower levels of VE than benchmarked here. MiST fails to call significant genes in all tests using 0.1% or 0.01% as FPR thresholds. We note that at FPR of 5% we expect around 1000 false positive calls. Hence, the performance at that FPR threshold is not relevant for applications of the tests to genome-wide studies, but can indicate how tests perform in targeted studies of up to 100 genes.

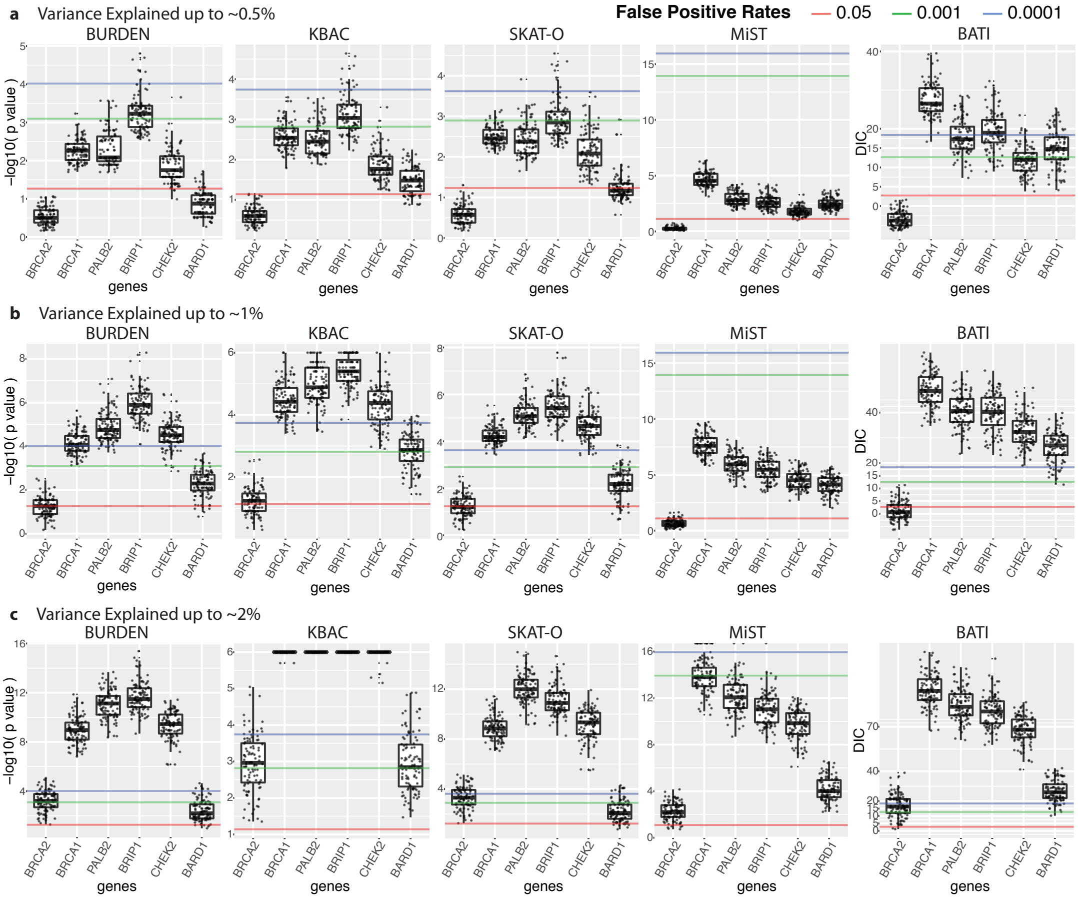

Figure 3.9: Boxplot of p-values and DICs for all simulated BRCA risk genes and methods. In **a** are shown results with simulations where variance explained (VE) is ~0.5%, in **b** with simulations where VE is ~1%, and in **c** with simulations where VE is ~2%. Each gene had 100 simulations for each VE dataset. Lines in boxplots mark: 5% (red), 1% (green) and 0.1% (blue) FPR thresholds.

In addition to benchmarking how RVAS tests behave on 'real' WES data with real sources of noise, another specific goal was to ascertain the impact of using biological features such as functional impact scores for variants (here CADD score) and categorical classification of variant function (i.e. missense vs. loss of function). For one gene, *BRCA1*, we only incorporated LoF variants from ClinVar. Indeed, the two tests that account for functional classification, MiST and BATI, always show the best performance for *BRCA1* (Figure 3.9), while Burden, KBAC and SKAT-O perform best on *BRIP1*. Specifically at the lowest FPR threshold and VE (FPR = 0.01% and VE = 0.5%) BATI is still able to identify *BRCA1* at close to 100% of tests (Figure 3.9a, left). We conclude that the novel BATI test is able to leverage categorical biological features of variants resulting in an improved performance compared to existing methods at low VE.

In *BRCA2* we only incorporated missense variants from ClinVar (but no LoF or splicing). This might contribute to the bad performance of MiST and BATI in that gene, although this cannot be distinguished from the issue of high background variant rates explained above, as all methods preformed poorly on *BRCA2*.

Finally, we tested if the observed type I error is close to the FPRs we were aiming for, which we found to be the case for most of the methods (see 3.5). Again, we observed a problem with the MiST, as 0 p-values are inflated for all genes with low number of mutations. If genes with 0-values are ignored, then the type 1 error of MiST is in an acceptable range similar to other methods.

Method	0.05 FPR	0.001 FPR	1e-04 FPR
BURDEN	4.854e-02	9.402e-04	1.763e-04
KBAC	5.001e-02	6.464e-04	0
SKAT-O	4.883e-02	1.175e-03	1.763e-04
MiST	1.025e-01	5.500e-02	5.412e-02
MiST without zero p values	4.842e-02	8.814e-04	0
BATI	4.936e-02	9.989e-04	5.876e-05

Table 3.5: T1 error with estimated P-value and DIC threshold from Table 3.4.

FPR stands at which False Positive Rate estimated threshold we used to calculate type I error

3.3 Discussion

In this work we presented the REWAS framework for rigorous rare-variant association testing using case-control cohorts sequenced with WES or WGS. Besides including five RVAS tests, it offers extensive quality control and filtering steps, which can help to prepare the data for association analysis such that a minimal rate of false positive findings is achieved. We presented a newly developed Bayesian inference based association test using Integrated Nested Laplace Approximation (BATI). We demonstrated that BATI outperforms other methods when simulated dataset is based on a real cohort analyzed by WES combined with real cancer risk variants obtained from ClinVar, providing meaningful biological features to the variants. Notably, BATI showed better power when phenotypic VE is low, a realistic scenario for many genes involved in complex disease etiology, such as cancer predisposition. Finally, we demonstrated that variant characteristics are an important source of information that RVAS tests based on hierarchical mixed-effect can leverage to improve performance. We expect that improved variant annotation methods will further increase the potential of this feature for studying various disease architectures.

Bayesian approaches have been less popular for biologists possibly due to the fact that inference is not based on classical hypothesis testing about parameters of interest, but on posterior distributions of parameters, which is a concept less familiar to those who are not specialists. Also, using Bayesian inference for estimating parameters of the generalized linear mixed models was for years requiring usage of MCMC method that are not feasible for genome-wide studies. The novel Integrated Nested Laplace Approximation approach can directly compute highly accurate approximations of the posterior marginals, that enables usage of Bayesian inference methods for genome-wide studies. Still, evaluation of the parameters of the model is not straightforward, as there are no p-values for coefficients as in classical frequentist approaches. Here we suggested using DIC difference between the full model and the null model, but the distribution of this value is not following a uniform distribution (as a distribution of p-values does). Therefore, we empirically calculated quantiles for DIC-difference, by repeating analysis on datasets with permuted case and control labels. We did not investigate how this distribution would change (if at all) with different sample sizes, different populations, or when changing any of the test parameters. We will approach this question in a future study using the larger 1000 Genomes Project cohort, currently providing close to 3000 WES samples.

We also observed that benchmark results can vary greatly depending on simulated datasets used for assessing the power of methods. Simulated datasets are often used for testing and evaluating methods mostly because they enable cheap creation of very large cohorts. Unfortunately, they do typically not capture all noise that is expected in real life studies and lead to overly optimistic benchmark results. For instance, we found that MiST outperformed all other methods when evaluated with the purely (and fairly unrealistic) simulated dataset, but under-performed when applied to a simulation based on a real WES cohort. We observed that the MiST test inflated 0 p-values when genes with low number of variants are present in the dataset (a common characteristic of highly conserved genes). Therefore, MiST became ineffective in our WES benchmark, as the type 1 error would be above expectation, unless some post-filtering of genes with low number of variants is implemented.

Population stratification is a big issue of association tests. RVAS tests specifically require the identification of rare variants in the cohort (and the filtering of common variants). This is problematic if the studied cohort is from a population that is not sufficiently represented in AF databases such as ExAC. REWAS provides the capability of splitting control cohorts to gain one group of samples used for AF estimation for the population and a second group as actual control. This approach benefits from large control groups and is not always feasible. Nonetheless, it can avoid AF estimation biases as an independent set of samples is used for the estimation of allele frequencies. The control-split functionality of REWAS has not been benchmarked in this study due to the limited size of our cohort used for simulations.

To conclude, REWAS and BATI facilitate discovery of new cancer predisposition genes using medium to large sized case-control studies. Large cancer cohorts, as the one expected to be released by PCAWG `http://docs.icgc.org/pcawg/`, are usually conglomeration of efforts from several research centers all over the world, and hence are prone to have population stratification issues and many possible biases. Therefore, quality control and filtering is mandatory functionality provided by REWAS. With large proportions of missing heritability in cancer predisposition, there is hope that we can still find new risk genes (or regulatory elements) with moderate to high effect size on phenotype. This would have great value in predictive medicine and early diagnostics. Still, genes with rare variants and low effect size on phenotype are might also play an under-appreciated role. A majority of rare variants does show low effect size [Auer and Lettre, 2015], but in sum they could fill the gap called missing heritability.

3.4 Methods

3.4.1 Data

Simulated datasets

Six simulated disease architectures (termed AR1–6) used for benchmarking of RVAS methods were obtained from [Moutsianas et al., 2015]. These datasets were created to model complex disease type 2 diabetes, and were generated using HAPGEN2 [Su et al., 2011]. For each architecture genotypes for 1500 case and 1500 controls have been simulated on 24-25 human genes of average coding length located on on chromosome 10. For every gene 100 simulations have been performed to facilitate statistical power analysis. In brief, AR1 and AR4 mimic strong selection AR2 and AR5 simulate moderate selection and AR3 weak selection of risk variants. For the first three architectures risk variants were simulated across the full site frequency spectrum (SFS) while in AR4 and AR5 risk variants had minor allele frequency (MAF) below 1%. Finally, AR6 simulates moderate selection of risk variants, using variants selected from the full SFS similar to AR2, however 50% of simulated variants were deleterious and 50% protective. In Table 3.6 all 6 architectures are summarized.

Simulated architecture	Direction of effects	Causal variant frequencies	Selection on causal alleles
AR1	All deleterious	Across full SFS	Strong
AR2	All deleterious	Across full SFS	Moderate
AR3	All deleterious	Across full SFS	Weak
AR4	All deleterious	$MAF < 1\%$	Strong
AR5	All deleterious	$MAF < 1\%$	Moderate
AR6	50% deleterious, 50% protective	Across full SFS	Moderate

Table 3.6: AR1–6 Locus architectures modeled at simulated loci.
doi:10.1371/journal.pgen.1005165.t001

Simulating disease variants in a real cohort

To allow for benchmark in a highly realistic dataset, which correctly represents all expected noise typically observed in cohorts analyzed by whole exome sequencing (WES) we simulated known cancer predisposition variants into a background of a real WES cohort from patients diagnosed with various conditions and healthy individuals. We used an in-house dataset combining 2,194 samples, which were subjected to whole exome sequencing during 2012 to 2017. The samples were collected within more than 30 different projects. The complete cohort includes individuals from eight populations: Spanish (1350), British (487), Italian (141), French (122), South African (16), Japanese (4), Moroccan (3) German (2), as well as 69 samples with unknown origin. Computational analysis and variant calling was performed according to GATK best practice guidelines (`https://software.broadinstitute.org/gatk/best-practices/`), including alignment with bwa-mem, GATK indel realignment and base quality recalibration and finally variant calling by GATK HaplotypeCaller.

To simulate a realistic case-control study we sub-select samples in order to minimize biases coming from different populations and different DNA analysis kits. To this end we only included unrelated samples (only one family member in case of e.g. trios) that belong to the Spanish population. WES libraries were prepared using seven different DNA analysis kits. We only included samples that were prepared for sequencing using one of the three exome enrichment kits: (1) Agilent SureSelect 50, (2) Agilent SureSelect 71, and (3) Nimblegen SeqEz V3. Furthermore, only SNPs and InDels located in targeted genomic regions, which were covered with an average of at least 10 reads in each kit group were included in the case-control study, as inclusion of low-coverage regions led to strong biases in QC (see Supplementary Fig. S3.2).

Samples that were identified as outliers with regard to number of called variants, transition to transversion (Ti/Tv) ratio, or on the projection of the first two PCA components were removed from further analysis. Moreover, genomic loci for which no genotype call was possible in more than 15% of samples (call rate ¡ 85%) were removed. The remaining cohort consisted of 1,167 unrelated Spanish samples, which harbored 285,658 unique loci with a non-reference genotype called in at least one of the samples. Description of the samples used for further analysis and benchmarking is s shown in Supplementary Table S3.1 and Figure 3.4.

Next, we randomly chose one third of the samples (389) to form the case group. The remaining 778 samples are treated as controls. To introduce realistic disease variants we queried the ClinVar (`www.ncbi.nlm.nih.gov/clinvar/`) database for breast cancer risk variants. We removed variants that had European MAF higher than 0.01 in any of the three databases: EVS, 1000 genomes project or ExAC. Also, variants that were not annotated as exonic or splicing were removed. We found that six genes had more than five annotated disease variants in ClinVar satisfying our criteria: *BRCA2, BRCA1, PALB2, BRIP1, CHEK2* and *BARD1*. Description of all ClinVar variants for these six genes is shown in Table 3.2.The selected variants form the pool of disease variants that were introduced into the WES cohort to simulate a cancer risk cohort, and we refer them as risk variants.

As expected, all six genes already had variants in our original cohort that are very likely not related to the trait breast cancer predisposition we are trying to simulate by introducing causal variants from ClinVar. This type of noise is expected in any case-control study using WES data, and hence increases the realistic level of the simulation. Existing exonic or splicing variants in the genes *BRCA2, BRCA1, PALB2, BRIP1, CHEK2* and *BARD1* with MAF < 0.01 are shown in Table 3.3.

Moutsianas et al. generated the six architectures mentioned above by simulating loci for which the phenotypic variance explained (VE) by genetic variants is ~1% [Moutsianas et al., 2015]. For the WES case-control simulation, we chose to test statistical power of methods in more detail by generating datasets with three different levels of VE, i.e. where the phenotypic VE by introduced ClinVar genetic variants is 1) ~0.5%, 2) ~1%, and 3) ~2%a. For BARD1 it was not possible to reach the desired VE due to an insufficient number of breast cancer risk variants found in ClinVar. Figure 3.5 shows the exact levels of VE simulated for each gene and for each simulation (100 simulations per gene).

Similar to the simulation of AR1-6 by Moutsianas et al. we used the method of So et al., which is available on-line as an R script [So et al., 2011], for calculation of variance explained (VE) at each locus. Calculate of VE requires three parameters per each variant as input: the prevalence of the trait, the population frequency of the risk allele, and the genotype relative risk. As we were simulating breast cancer risk in a Spanish population, prevalence was assumed to be 0.00085 [Ferlay et al., 2013b]. We aimed at generating realistic genotype relative risk (RR) distributions where more common risk variants (e.g. MAF$\approx$0.01) were more likely to have lower RR, while very rare risk variants (e.g. MAF$\approx$0) have a higher chance to have high RR. To this end, RR for each risk variant has been determined using the following algorithm:

1. MAF of introduced variants is estimated using the ExAC database. We define minimum MAF as 1e-10 to avoid zeros.
2. To take into account the expected exponential decay in RR variability based on the MAF range of values, we used the probability density function beta, that is a family of continuous probability distributions defined on the interval $[0, 1]$, the range of values that MAF can take, and parameterized by two positive shape parameters. The beta distribution with shape parameters $\alpha = 0.005$ and $\beta = 1$ best emulates the expected behavior, therefore we generated quantiles from this beta distribution to obtain reasonable values of RR variability across the range of MAF values. As the resolution for very rare variants (MAF$<$5e-04) is limited, a minimum between the computed value and 10 will be taken as standard deviation in next step.
3. For variants with different MAF values we sample RR from a normal distribution with mean equal to zero and standard deviation calculated as described in step 2. Then we take the absolute value and add 1.5 (as we have twice more controls, and minimum RR should in general be above 1). As final relative risk (RR) we take the minimum between the computed value and 16, as we want to avoid very extreme scenarios (e.g. a single risk variant affecting 10% or more cases).

Using the RR generation algorithm results in an RR distribution in which risk variants with MAF$\approx$0.01 have RR around 2, while for very rare risk variants (MAF$\approx$0.001) RRs between 1.5 and 16 are generated. All parameters were chosen such that the RR distribution resembles the RR distribution presented in Moutsianas et al. in Supplementary Fig. S3 [Moutsianas et al., 2015]. The RR distribution is shown in Supplementary S3.3. Given the case-control nature of our design, we have used the odds ratio (OR) instead of RR, but since the prevalence of the simulated disease is very small, the OR is a good approximation to RR. The population frequency of the risk allele can simply be estimated from the cohort.

We assumed independence between risk variants at a given locus, and thus estimated the total percentage of VE as the sum of the VE by each individual variant. The following procedure was used for introducing risk variants for each of the six simulated breast cancer risk genes:

1. Pick randomly a variant from the pool of the ClinVar candidate variants for a gene
2. Estimate MAF for this variant with ExAC database
3. Introduce an effect by sampling from the generated RR distribution
4. With MAF from (2) estimate number of affected controls. With a number of affected controls and sampled RR from (3) estimate number of affected cases
5. If number of affected cases and controls is 0 go back to step (1)
6. Introduce variant into the estimated number of cases and controls by sampling randomly cases and controls
7. Remove the risk variant from the pool of candidate variants for this simulation
8. Stop if there are no more variants in the pool of candidate variants
9. If the cumulative variance explained by variants introduced in a gene is below the specified threshold (0.5%, 1% or 2%), go to step (1) and repeat
10. If the variance is above the specified threshold stop with introducing risk variants

Additionally, for *BRCA2* we only introduced missense SNVs (but not loss of function–LoF variants) as risk variants, while for *BRCA1* we only introduced stop gain and splicing SNVs (i.e. only LoF variants). This was done so we could test if methods like MiST and BATI (see below) can benefit from categorical features that capture biological function and context of variants. The simulation procedure is repeated 100 times for each of the six simulated risk genes in order to generate 100 datasets for statistical power and false positive rates analysis.

3.4.2 Framework for Rare Variant Association Studies (RVAS)

We have implemented the 'Rare variants Exome Wide Association Study' (REWAS) framework that integrates various gene-based rare variant association study methods, quality control procedures, covariates and variants characteristics construction, approach to estimate local population AF, as well as visualization methods. REWAS is implemented as an sequence of R scripts, and is available on-line at `https://github.com/hanasusak/REWAS`.

Quality Control (QC)

The first module of REWAS implements various quality control procedures and methods. REWAS–QC checks if cases and controls are homogeneous and that there are no biases in data that could cause high false positive rates. To do so we developed an interactive R script that in user friendly way computes essential quality control (QC) measures and prepares data for association testing. The script has two mandatory input parameters: 1) sample information file path and 2) variant information file path. The sample information file has only one mandatory column specifying the names of the samples. Additional columns can be provided giving useful information about each sample, e.g. library preparation and target enrichment kits, gender, population, etc. The variant information file is similar to multi-sample VCF (variant call format) files with a few changes facilitating functional annotation of variants. A header line is used for specifying the content of columns, where column names that are associated with functional annotation start with # and column names that are imported from the original VCF do not start with #. Another difference to VCF is that sample genotypes are annotated as: NA (no call available for the sample), 0 (both alleles called as reference), 1 (heterozygous SNV) and 2 (homozygous SNV). All columns are tab separated and the following annotation columns are mandatory: #Chr, #Position, #Reference, and #Alteration. Sample names in the header have to match the names given in the sample information file. (If this is not the case there is an optional parameter allowing to specify a translation file with two columns. The first column represents the names of the samples from the variant information file and the second column is the corresponding sample name in the sample information file.)

Example of file with samples information

Samples_ID	DNA_kit	Gender	Population	Project
Sample1	Agilent71	F	Spanish	CLL
Sample2	Agilent71	F	Spanish	CLL
Sample2	Agilent50	F	Spanish	OCD

Header of example file with variants information

#Chr	#Position	#Reference	#Alteration	#Gene	Sample1	Sample2	Sample3
1	808861	G	A	FAM41C	0	NA	1
1	808922	G	A	FAM41C	1	0	0
1	808928	C	T	FAM41C	1	1	0
1	808984	A	G	FAM41C	NA	0	2
1	808991	C	T	FAM41C	0	1	0

If InDels are stored in a separate file the user can provide the additional file. The InDel file should have the same format as the genetic variants file. SNV and InDel files will be joined prior to association analysis, but an additional column will be added indicating if a variant is a SNV or an InDel.

The QC script is interactive and offers the user to perform the following steps:

1. Users will be offered to remove any samples based on any column from the sample

information file. For example, the user can choose to remove all samples that in the ‘Population’ have an entry different from ‘Spanish’

2. . Next, users could remove variants based on any annotation column. If an annotation column is categorical the user can specify specific entries to keep or remove, e.g. remove anything other than ‘exonic’ or ‘splicing’. If an annotation column is numerical users will be briefly informed on quantiles for the column and can then specify thresholds for minimum or maximum values. Such example of a column could be ‘#Segmental_duplication’, where the user would filter out all variants that have 0.9 or higher value in this column.

3. If for some variants genotypes were not called (NAs) users will be offered to remove variants based on percentage of NAs in the cohort (i.e. variant call rate).

4. A histogram of number of variants per sample is presented (Figure 3.4a). Samples that have a number of variants outside of the limits specified by the user will be removed.

5. A histogram of transition to transversion (Ti/Tv) ratio per sample is presented (Figure 3.4b). Samples that have a Ti/Tv ratio outside of the limits specified by the user will be removed.

6. Users can choose the type of variants for calculating a PCA. Given large enough variant sets, synonymous SNVs are recommended for PCA calculation as they are mostly neutral (no selection, no effect on disease). In addition very rare variants (MAF<0.005) can be removed from PCA, as they are not informative. Furthermore, all variants in linkage disequilibrium (LD ≥ 0.2) are removed from PCA analysis automatically. Optionally, user can choose color and shape for visualization of samples in the PCA projection. For PCA any genotype labeled NA will be replaced with 0, in order to significantly gain on speed with a cost of possible inaccurate projection of samples on PCA components. Still, variance explained by each component is not expected to change drastically. Then the first 20 PCA components will be calculated for selected variants and the percentage of variance explained by each PCA component will be shown to the user (Figure 3.4c). Additionally, pairwise projection of first 10 components will be shown (Figure 3.4e). Based on these plots users need to choose the number of PCA components (n) to be included as covariates in association testing.

7. With information from previous step n PCA components are recalculated, but this time not replacing NA values by 0 (therefore projection on PCA components will be accurate). Numerically this is solved with the Non-linear Iterative Partial Least Squares (NIPALS) [Wold, 1966] function. Samples are projected on the first two PCA components where each sample is colored and given shapes by previously chosen attributes (Figure 3.4e upper right corner). Users can choose minimum

and maximum values for the first two components where everything outside of chosen limits will be removed.

8. Informative (not interactive) barplot of number of variants per sample, colored by variants type/class is also produced. Users can choose any annotation column (starting with #) from the variants information file that is categorical to be used as color class. Additionally, users can aggregate and/or rename values from this column (e.g. in column '#Exonic Function' stop gains and stop losses could be renamed to stop gain/loss). An example of such a plot is shown in Figure 3.4d.

9. Finally, if the sample information file contains a column specifying cases and controls, additional informative plots will be generated:

    - Barplot of number of mutations, colored by cases and controls. An example of such plots is shown in Supplementary Fig. S3.4a. In case one color is biased towards the upper end of the plot users should check why the respective group has inflated number of variants (or the other group deflated numbers).
    - Number of variants per gene in cases versus controls. An example of such plots is shown in Supplementary Fig. S3.4a and b. Here, user can assess if the number of variants per gene is balanced (equally distributed) between cases and controls.

    Plots in this step are performed on filtered data from previous steps and users can again compare cases and controls on specific types/classes of variants.

The output of the QC scripts will be all mentioned plots, before and after filtering if applicable. Additionally, two text files will be generated, one with sample information for all samples passing all filtering steps plus added n number of PCA component projections. The second text file is a variant information file with variants passing the all filters. These files are used as input for RVAS methods. All user choices and any other relevant information from QC is saved in the log file.

Rare Variant Association Study (RVAS) Methods

The current REWAS version implements five RVAS test methods, including the four published methods" BURDEN [Lee et al., 2012b], KBAC [Liu and Leal, 2010], SKAT-O [Lee et al., 2012b][Lee et al., 2012a] and MiST [Sun et al., 2013]. These four RVAS tests have been selected based on their superior performance (power) compare to other tests in previous benchmark papers [Moutsianas et al., 2015]. Furthermore, each of these methods is available as implementation in R. Moreover, the four methods are complementary, as they address different difficulties with rare-variant association testing.

BUDREN Test

BURDEN is a unidirectional test that aggregates the rare variants within a focal gene (or any other defined region or functional unit) as a single burden variable, which can then be tested for association with any trait of interest. It is more powerful when a large fraction of rare variants in the set is causal and the effects are mainly deleterious with similar magnitude. If some of these assumptions do not hold we expect that power with BURDEN tests will be significantly reduced.

In REWAS the BURDEN test was implemented using the package SKAT (`cran.r-project.org/web/packages/SKAT/index.html`) version 1.3.0. The Null model, which only contains covariates, was generated using the `SKAT_Null_Model` function with output set to dichotomous outcome (`out_type ="D"`) and no sample adjustment (`Adjustment=FALSE`). All other parameters were set to default. For the actual Burden test with genetic information and covariates we used the function `SKATBinary` with all parameters at default values except for method, which was set to `"Burden"` and for weights, for which we used MAF of variants transformed with the `Get_Logistic_Weights` function with default parameters.

KBAC

KBAC is a unidirectional test that combines variant classification and association testing in a coherent framework. Compared to BURDEN test it is expected to have better power if there is a mixture of non-causal and causal variants in a gene.

KBAC is available as R implementation (`tigerwang.org/software/kbac`), but with some restrictions in parameter options as compared to the KBAC standalone software implementation. For example, with the KBAC R package version 0.1 there was no possibility to include covariates. For association testing we used the function `KbacTest` with parameters `alpa=2.5e-06`, `num.permutation=1000000`, and with all other parameters set to default values.

SKAT-O

SKAT-O approach is a weighted linear combination of a unidirectional burden test and the SKAT variance component bidirectional test. SKAT uses a multiple regression model to directly regress the phenotype on genetic variants in a defined region and on covariates, and therefore allows for variants to have different directions and/or effects size [Wu et al., 2011]. Therefore, SKAT-O automatically infers from the data what test, burden or SKAT, is more appropriate and optimize for the best of both.

SKAT-0 was implemented in REWAS using the package SKAT, the same package as used for BURDEN test. The Null model, which only contains covariates, was generated using the `SKAT_Null_Model` function with output set to dichotomous outcome (`out_type ="D"`) and no sample adjustment (`Adjustment=FALSE`). All other parameters were set to default. For SKAT-O association testing with genetic information and covariates we used the function `SKATBinary` with all default parameters except for method, which was set to `"optimal.adj"` corresponding to SKAT-O test, and for weights, for which we used MAF of variants transformed with the `Get_Logistic_Weights` function with default parameters.

MiST

The MiST test is based on a type of mixed-effect models implemented using hierarchical modeling, which can utilize known characteristics of variants as e.g. functional impact. Apart from continuous damage score this could also benefit from inclusion of categorical variables as e.g. missense vs. LoF SNV. It tests for fixed and random effects in two steps, and similar to SKAT-O can handle causal variants with different directions and magnitude of effects. For additional details about MiST implementation we refer to Sun et al. [Sun et al., 2013].

MiST was developed as standalone R package and is available at CRAN repository: `cran.r-project.org/web/packages/MiST/index.html`. Here we used MiST R package Version 1.0 and we used the function `logit.weight.test` with all default parameters to perform the MiST association test.

BATI

We have developed a novel genetic association test (BATI) based on Integrated Nested Laplace Approximation (INLA; [Rue et al., 2009]). INLA allows implementation of Bayesian inference on the generalized linear mixed model framework. The method is conceptually similar to the method MiST [Sun et al., 2013] in the sense that it considers heterogeneous effects by specifying individual effects of variants as random effects, allowing the inclusion of prior knowledge about the variants (such as functionality or damaging scores) and incorporation of confounders at patient level (such as gender or population stratification).

Integrated Nested Laplace Approximation

Integrated Nested Laplace Approximation (INLA) is a new approach to implement Bayesian inference on latent Gaussian models, which are a very wide and flexible class of models ranging from (generalized) linear mixed models (GLMMs) to spatial and spatio-temporal models. Thanks to this, INLA can be used in a great variety of applications [Li et al., 2012b, Ruiz-Cárdenas et al., 2012, Martino et al., 2011, Roos and Held, 2011, Schrödle et al., 2011, Schrödle and Held, 2011, Paul et al., 2010]. Unlike MiST, BATI thanks to INLA can make inference based on full model estimation, and therefore allows us to obtain more information about estimates of model parameters.

INLA provides approximations to the posterior marginals of the latent variables, which are both very accurate and extremely fast to compute [Rue et al., 2009]. In particular, it has been developed as a computationally efficient alternative to MCMC and presents two main advantages over MCMC techniques. On the one hand, it is worth noting that INLA's fast speed allows to work on models with huge dimensional latent fields. On the other hand, INLA treats latent Gaussian models in a unified way thus allowing greater automation of the inference process. A detailed description of the INLA method and a thorough comparison with MCMC results can be found in [Rue et al., 2009].

Model specification

Assume we have N individuals, and let y_i $(i = 1, \cdots, N)$ be the observed trait of the i_{th} individual that belongs to an exponential family

$$y_i \sim \pi(y_i;\, \mu_i, \theta) \tag{3.1}$$

where the expected value $\mu = E(Y_i)$ is linked to a linear predictor η_i through a known link function $g(\cdot)$, so that $g(\cdot) = \eta_i$. In our case Y_i is a binary variable which is assumed to follow a Bernouilli probability distribution, and the common specification for $g(\cdot)$ is the logit function. Instead, we propose to construct the likelihood of the data based on a logistic distribution and use the identity function for $g(\cdot)$. The linear predictor η_i is defined to account for potential confounding covariates at individual-level and RVs effects:

$$\eta_i = X_i^t \alpha + G_i^t \beta \tag{3.2}$$

where X_i is a $m \times 1$ vector of confounding covariates and G_i denotes a $p \times 1$ vector of genotypes for p RVs and each genotype is coded as 0, 1, or 2, representing the number of minor alleles. α and β are the regression vectors of coefficients.

As proposed by [Sun et al., 2013] we can allow to account as well for individual variant characteristics under the assumption that similar variant-specific characteristics have similar effect on the trait, while still allowing for potential individual heterogeneity effect. Therefore β can be modeled in a hierarchical way as:

$$\beta_j = Z_j^t \omega + \delta_j \tag{3.3}$$

where ω is a vector of $q \times 1$ $(j = 1, \cdots, q)$ variant-specific regression coefficients, Z^t is a $p \times q$ matrix, and δ is a $p \times 1$ random effects vector which is assumed to follow a multivariate Gaussian distribution with mean 0 and covariance matrix τQ . If no dependency structure is defined across variants as in [Sun et al., 2013] Q is a $p \times p$ identity matrix. However, if one thinks that the variants have a certain correlation structure such as physical distance dependency across the variants, then Q is constructed so that it reflects this structure. The advantage of INLA is that the Laplace approximation of the posterior distributions allows the estimation of the full model for complex structures of random effects.

Plugging expression 3.3 into expression 3.2 we have the expression of a generalized linear mixed effects model (GLMM):

$$\eta_i = X_i^t \alpha + (G_i^t Z)\omega + G_i^t \delta \tag{3.4}$$

with α and ω as fixed effects coefficients and δ as random effects coefficients.

Assuming that the vector of parameters is represented by $\theta = \{\alpha,\, \omega,\, \delta\}$, the objectives of the Bayesian computation are the marginal posterior distributions for each of the elements of the parameters vector $p(\theta_s|y)$ and for the hyper-parameter $p(\tau|y)$.

Thus, firstly we need to compute $p(\tau|y)$ and $p(\theta_s|\tau, y)$, which is needed to compute the marginal posterior for the parameters. The INLA approach exploits the assumptions of the model to produce a numerical approximation to the posteriors of interest, based on the Laplace approximation [Tierney and Kadane, 1986].

More details on these methods can be found in [Rue et al., 2009, Martins et al., 2013, Blangiardo et al., 2013].

Model selection

For association test we use a model selection criteria comparing the likelihood of a null model with no genetic effects and the likelihood of an alternative model with some genetic effects:

$$H_0 :\, \eta_i = X_i^t \alpha \tag{3.5}$$

$$H_1 :\, \eta_i = X_i^t \alpha + (G_i^t Z)\omega + G_i^t \delta \tag{3.6}$$

Therefore we use the difference in deviance information criteria (DIC) between models [Spiegelhalter et al., 2002].

Computational implementation

INLA is implemented as an R-package within the R freeware statistical program, which offers a friendly framework and a very good tool for inference on latent Gaussian models. In our REWAS pipeline we used R package INLA version 17.6.20. To run the null model as independent variables we used only covariates, and for full model we added genetic information. In full model as the latent Gaussian field we used classical random effect model (by specification of INLA it is equivalent to setting `model="z"` in formula for random effects). In both models variable `family` was set `"logistic"`, variable `control.compute` to list of values `dic=TRUE, cpo=TRUE, config=TRUE`, number of threads (`"num.threads"`) to number of available cores specified by user (default 1), and the step-length for the gradient calculations for the hyperparameters (`h`) to value specified by user (default 1e-3). All other parameters are left to default values.

Rare Variant Association Study – Running the REWAS pipeline

All tests are implemented in an R package called REWAS requiring as input two mandatory parameters, path to the samples information file and path to the variants information file, that are produced in QC step of analysis. Optional parameters are allele frequency threshold (default is 0.01 if not set), path to the file with list of genes (or another analogous list) to be tested, and path to the output folder. If output folder is not defined a new folder will be automatically created in the current directory starting with 'Risk_analysis' and ending with the current date (format Y-m-d-HMS) where all results will be saved. The QC part of the analysis can be performed in two ways, running the R script interactively or running the R script as bash job (which can be distributed on a compute cluster or cloud if necessary). The Interactive way is more intuitive and easier for users with no Linux experience but requires a stable connection with the server where the analysis is performed. It is important to mention that analysis can take ~9 days if genes of the whole exome are tested (~20000 genes) on relatively small cohort (~1200 cases and controls together) if one core is only available and all 5 tests are performed. For this reason we developed an alternative way to perform analysis where users need to setup a config file and submit the script as a batch job to a server or compute cluster. BATI is the slowest association test from the five, but it can benefit from multiple cores if they are available. From our experience with REWAS, times drop linearly with increase of numbers of available cores. For example, for the same dataset, on the same computer architecture and the same type of analysis, with 1 core it was necessary ~9 days to finish, while with 5 cores available it took a bit less then 2 days.
All the variables that user can set in config file with some examples are:

```
READ_FROM="bash" # mode for the script, default 'stdin' - interactive mode
R_MAX_MC_CORES=1 # number of cores available, default 1
R_SEED=160185 # random seed, default 160185
R_PERMUTATIONS=1 # number of random splits of controls in two datasets

R_INLA="T"  # logical indicator (T or F) if BATI test should be performed, default T
R_MIST="T" # logical indicator (T or F) if MiST test should be performed, default T
R_KBAC="T" # logical indicator (T or F) if KBAC test should be performed, default T
R_SKATO="T" # logical indicator (T or F) if SKAT-O test should be performed, default T
R_BURDEN="T" # logical indicator (T or F) if BURDEN test should be performed, default T
MAF_LOGICAL="F" # logical indicator (T or F) if MAF of the variants should be uses as
    weights for BATI test, default F
H_STEP= 0.001 # step for gradient decent in INLA, default 1e-3. In case INLA is not
    converging step should be reduced

R_COV_MAT_COLS="PC1,␣PC2,␣PC3" #  column names in samples info file that should be
    used as covariates, default "". If not set it will not use covariates for test
    where possible
R_AGG_COL="#Gene" # column name in variants info file indicate what should be used to
    aggregate variants. This is usually Genes name column.
```

```
R_PROJECT_COL="case_ind" # column name in samples info file that contain information
    what are cases and what are controls samples
R_CASES_NAME="case" # in R_PROJECT_COL what are the cases. If multiple values, they
    should be comma separated
R_CONTROLS_NAME="control" # in R_PROJECT_COL what are the controls. If multiple
    values, they should be comma separated. If not provided all values different form
    R_CASES_NAME in R_PROJECT_COL will be used as controls.

R_CHAR_FILT_COL="#ExonicFunction" # column names in variants info file that are
    categorical and we want to filter from. Separated with |
R_CHAR_FILT_VAL="synonymous␣SNV" # Values for each of the columns in R_CHAR_FILT_COL
    that we want to filter out. When starting with values from different
    R_CHAR_FILT_COL column separate with |, but in-between one  R_CHAR_FILT_COL
    column separate with comma

R_NUM_FILT_COL="#EurEVSFrequency|#Eur1000GenomesFrequency|#ExAC_NFE|#Cadd2" # column
    names from variants info file that are numerical and we want to filter on.
    Separated with |
R_NUM_FILT_VAL="h|h|h|l" # If we want to filter out higher (h) or lower (l) then
    specific threshold. separated with |
R_NUM_FILT_TRH="0.01|0.01|0.01|10" # Threshold for each column in R_NUM_FILT_COL,
    separated with |

R_DUMMY_VAR="#ExonicFunction" # column name from variants info file that you want to
    use as variants categorical description
R_DUMMY_MERGE="frameshift␣deletion,␣frameshift␣insertion,␣nonframeshift␣deletion,␣
    nonframeshift␣insertion␣|␣stopgain␣SNV,␣stoploss␣SNV,␣splicing" # values from
    column R_DUMMY_VAR that you want to merge and/or rename. Groups for
    merging/renaming are separated with |, and values from same group with comma.
R_DUMMY_RENAME="indels|stop_gain_loss_splicing" # new names for each group in
    R_DUMMY_MERGE, separated with |
R_NUMERIC_VAR="#Cadd2" # column name from variants info file that you want to use as
    variants numerical description

R_CONTROLS_NUM=778 # number of samples to be used as controls from available amount.
    If not set all available controls will be used as controls.
```

If user wants to run RVAS analysis as batch job than mandatory variables to set are: READ_FROM, R_AGG_COL, R_PROJECT_COL and R_CASES_NAME. Others or have default value or are not obligatory. For example, if user wants to run only one specific test, e.g. BATI, few lines of bash script would be sufficient:

```
#!/bin/sh
export READ_FROM="bash"
export R_INLA="T" # not nessesary to put here as it is default value
export R_MIST="F"
export R_KBAC="F"
export R_SKATO="F"
export R_BURDEN="F"
export R_AGG_COL="#Gene"
export R_PROJECT_COL="case_ind"
export R_CASES_NAME="case"

Rscript RVAS_tests.R -m "path␣to␣variant␣info␣file" -d "path␣to␣
    samples␣info␣file" -f 0.01
```

Here, all the variants that have AF estimated from controls higher then a given threshold (0.01) will be removed from analysis. If user wanted to filter out all variants that have AF higher then given threshold in public databases he would need to use R_NUM_FILT_COL variable and specify the columns or follow instruction when done in interactive mode. It is reasonable to argue that using controls to estimate AF might cause biases in analysis. Therefore, there is an option of splitting controls in two datasets, where user choose on which fractions he wants to split controls. One fraction of controls will be used to estimate allele frequency (AF), while the other fraction will be used as regular controls dataset. All the variants that have AF in the first dataset higher then AF threshold that is passed as parameter, will be removed from analysis. To avoid unlucky random splits this process can be repeated N times. Default is 100, but user can controls this with R_PERMUTATIONS variable or follow instruction when done in interactive mode. If multiple cores are available and set with variable R_MAX_MC_CORES every spit-controls analysis will take first available core. If there is more then one core available and controls are not divided in two datasets, then BATI analysis will take all available cores as it is possible to parallelize process for estimating parameters and it speeds up time linearly.

Each of the five tests will produce one output file. All output files will have following columns

genes Aggregation column values

total.mut Total number of unique variants participating in analysis for this gene (or other specified aggregation unit)

cases.mut Number of unique variants in cases.

controls.mut Number of unique variants in controls

num.cases Number of affected cases

num.controls Number of affected controls

no.na.cases Number of cases that participated in analysis

no.na.controls Number of controls that participated in analysis

p.val.pi Only MiST results have this column. P-value for test $\pi = 0$

p.val.tau Only MiST results have this column. P-value for test $\tau = 0$

p.val.overall P-value for the gene. BATI results do not have this column

adjust P-value corrected for multiple testing (BH). BATI results do not have this column

dic.diff Only BATI results have this column. Difference between DIC values for models when genetic information is used and when only covariates are used. The higher the DIC difference, the more likely the genotypes in the gene can explain the trait.

cpo.diff Only BATI results have this column.. Difference between CPO values for models when genetic information is used and when only covariates are used. The higher the CPO difference, the more likely the genotypes in the gene can explain the trait.

All columns except **genes** have suffix *_permX*, where X goes from 1 to the N (number of random splits of controls in two sets). In addition, a log file is saved in the results folder. A flowchart of the pipeline is shown in Figure 3.1.

3.5 Supplementary Information

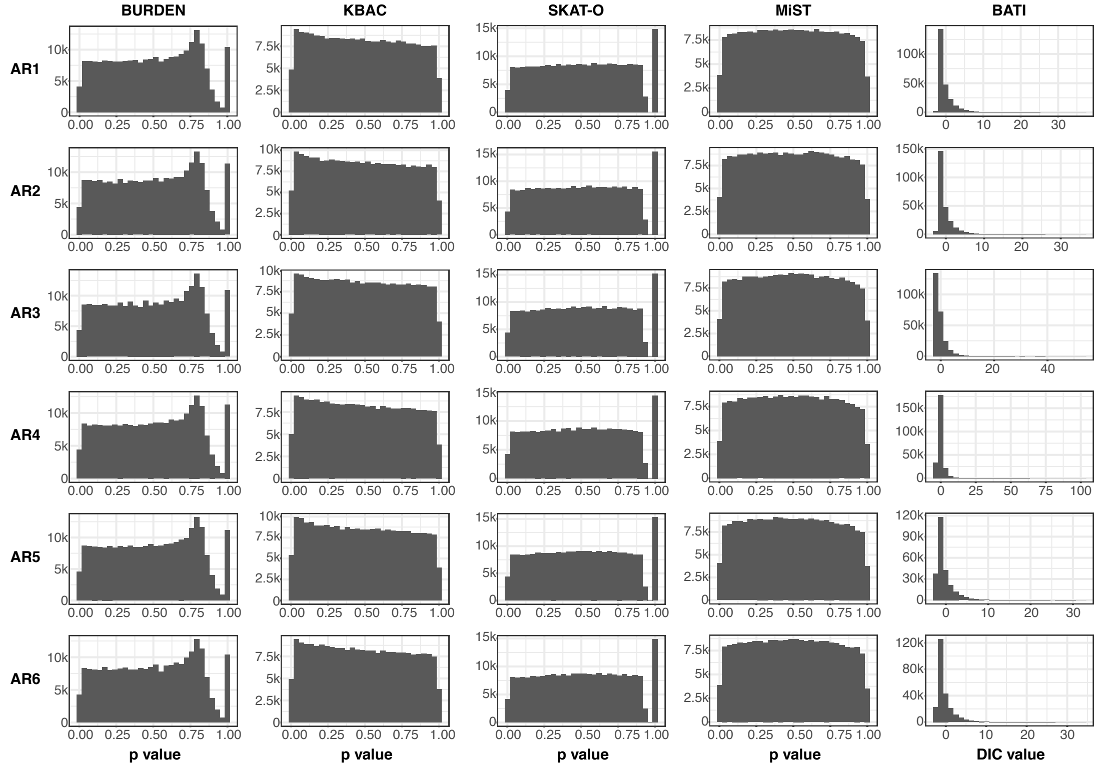

Figure S3.1: Distribution of p-values and DIC for each method and architecture

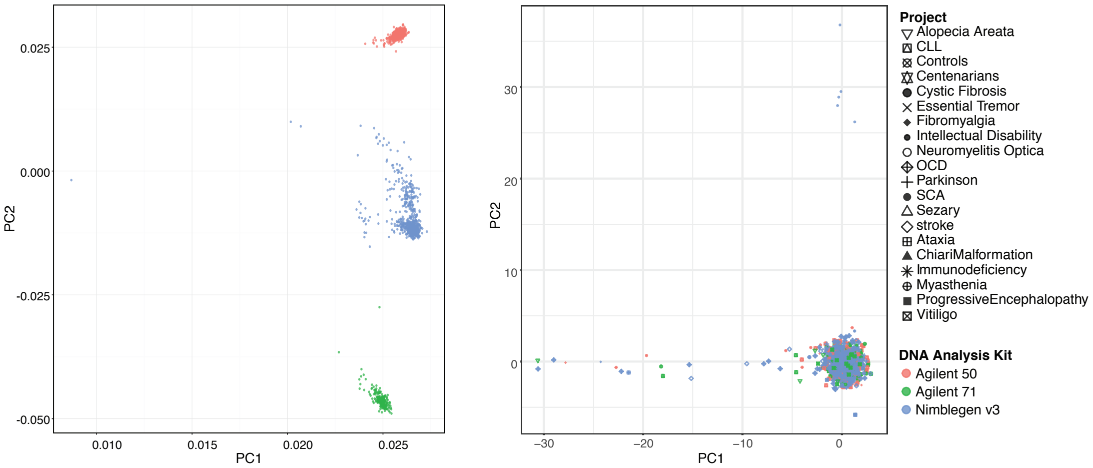

Figure S3.2: First two PCA components, colored by DNA analysis kits.

a We show first two PCA components projections of samples when used only SNPs and InDels that are in intersect of defined genomic regions by used DNA analysis kits. **b** SNPs and InDels used for second PCA projection were found in intersect of regions that were covered at least 10x by all three DNA analysis kits. In both cases PCA projection was done on synonymous variants and pruned for LD.

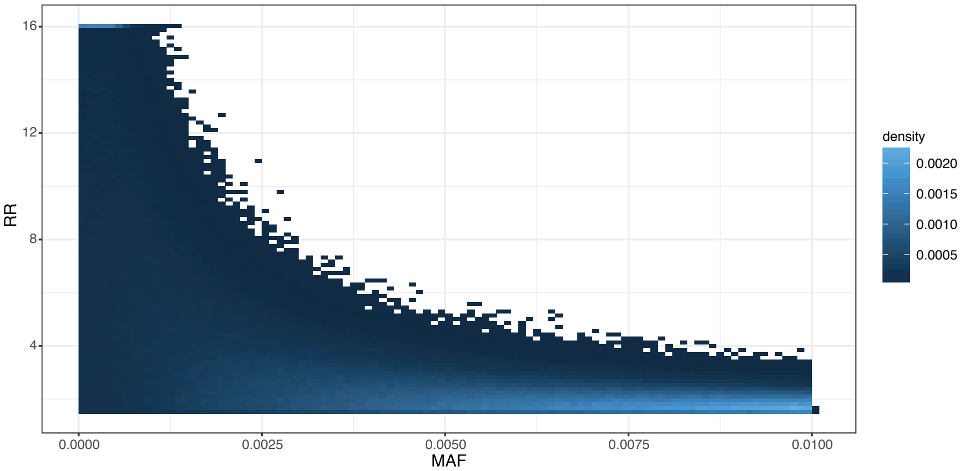

Figure S3.3: Distribution we sample relative risk (RR) across MAF spectrum [0, 0.01]

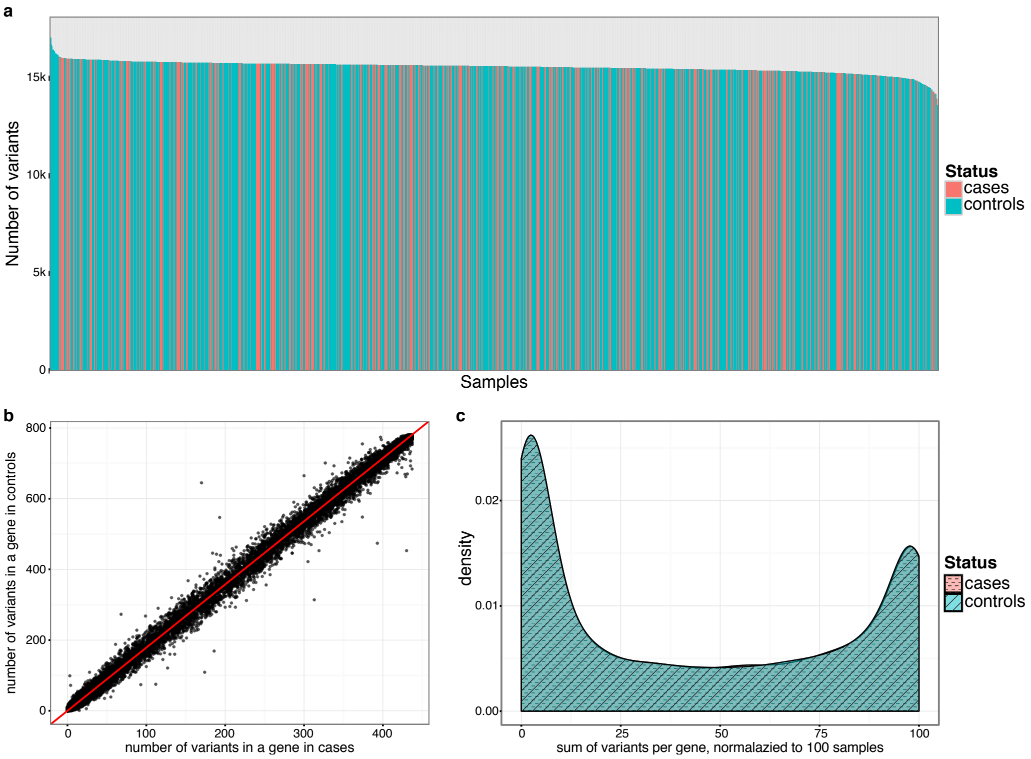

Figure S3.4: Example of exploitative plots for comparing cases and controls

In all three plots as cases are taken CLL samples and all other samples are considered as controls. **a** Barplot for number of variants per sample. **b** Each dot is a gene, and their x and y axis values are number of variants for the gene in cases and controls. Red line represents ratio of number of controls and cases. **c** For each gene is calculated in how many cases and controls is mutated. These numbers are normalized to 100 samples in each group, and then density for such numbers is plotted, colored by case/control groups.

DNA kit Project	Agilent 50	Agilent 71	Nimblegen v3
Alopecia Areata	0	25	0
Chronic Lymphocytic Leukemia	276	160	0
Controls	0	0	63
Centenarians	1	0	0
Cystic Fibrosis	10	16	0
Essential Tremor	0	0	4
Fibromyalgia	39	0	49
Intellectual Disability	79	0	1
Neuromyelitis Optica	14	0	0
Obsessive Compulsive Disorder	0	0	260
Parkinson	0	0	38
Sezary	0	6	0
Stroke	0	0	79
Ataxia	0	0	12
ChiariMalformation	0	0	2
Immunodeficiency	0	0	9
Myasthenia	0	0	6
Progressive Encephalopathy	0	0	4
Vitiligo	0	14	0
Total per DNA kit	419	221	527

Table S3.1: Description of WES samples used for simulation .

Chapter 4

DISCUSSION

In this thesis I have presented two projects: (*i*) A Bayesian statistical model for inferring cancer driver genes, and (*ii*) A comprehensive statistical framework for rare-variant association studies. Looking back in history, most of the ideas used in this study to improve statistical methods for precision medicine and large cohort analysis have their roots in research performed decades ago. For instance, the hypothesis that clonal mutations are enriched for driver genes in most cancer types has been proposed by Peter Nowell in 70's, and later observed by other groups [McGranahan et al., 2015].
Still, late driver mutations that had no time to expand (until the time that specimen was sampled from the patient in a biopsy), driver mutations that are observed in subclones that are physical limited to expand, and driver mutations that are in subclones that optimally 'co-operate' in specific fractions, will not show strong clonal signatures, if at all. Thus, we believe that there is no one 'magical' signature of positive selection that can identify all driver events in cancer, as cancers are a complex and highly heterogeneous disease on multiple levels. Considering many signatures of selection at the same time is the way that promises to identify the complete landscape of cancer driver genes. In this work we have made a first step by integration of three signatures of selection. However, other signatures and variant types could be added, e.g. differential expression, copy number variants or mutation clustering. Bayesian inference as well as other machine learning based methods are powerful tools able to integrate large number of features (signatures) and are furthermore able to learn from prior knowledge. We expect to see more of these approaches in the future used to tackle the issue of identifying all causes of cancer in a comprehensive manor.

Rare-variant association studies (RVAS) gained huge interest once the GWAS method started to show limitations and the issue of the missing heritability became obvious [Lee et al., 2014, Manolio et al., 2009]. Especially the advent of NGS allowing for whole-genome or whole-exome sequencing made new approaches necessary that can utilize rare variants. Most loci identified by GWAS have a modest effects on disease risk, which in practice often means a long journey until clinical diagnostics (or treatment) can benefit from this knowledge. That rare variants play important role in complex diseases has been shown in many studies [Roth et al., 2012, Stein et al., 2012, Rivas et al., 2011, Jonsson et al., 2012], and several rare risk variants influencing cancer risk are known [Gudmundsson et al., 2012]. With the emergence of sequencing data for large cancer cohorts generated by TCGA and ICGC consortia and the release of predicted somatic and germline variants, we saw the opportunity to analyze the germline genome of cancer patients in order to identify new cancer risk genes. Although, most of the samples in this cohorts are not familial, we know that many sporadic cancers have a strong genetic component [Lichtenstein et al., 2000]. Lack of good matching controls motivated us to develop a comprehensive framework for rare-variant association testing of case-control cohorts not only incorporating state of the art RVAS methods, but also modules for quality control and population stratification. RVAS approaches used on cancer cohorts have the potential to reveal a number of moderate to strong effect variants/genes, which will prove useful in predictive personalized medicine, and could represent new drug targets.

As result from the first project (Chapter 2), we presented the standalone R package cDriver that is focusing on analysis of somatic mutations from tumor tissue. It utilize multiple signatures of positive selection, from which cancer cell fraction (CCF) is for the first time used as part of any software to infer cancer driver genes. We first demonstrate that CCF is a potent signature of positive selection, and subsequently we used CCF with other signatures as evidence (or not) for mutations to be true drivers, all integrated in Bayesian statistical inference model. Further, we have performed extensive benchmarking of cDriver and four other commonly used driver prediction tools. From benchmark analyses we came to conclusion that cDriver performs as good or better then the best of the four competing tools. Nonetheless, we found that ensembles of two or more methods still outperformed any single method. Importantly, we could show that cDriver contributes to an improved result of the ensemble of methods.

In second part of Chapter 2 we focused on driver genes that are mutated in a very small fraction of cancer patients (lowly recurrent driver mutations). Methods that exploit recurrence as sign of positive selection where not successful in identifying driver genes that are found in <5% of cancer patients [Vogelstein et al., 2013]. We exploited the observation that many cancer driver genes are shared as driver events between different tumor types. We identified ‘tumor type-driver gene’ (TTGD) connections for genes that are not frequently mutated in that particular tumor type, but frequent in another tumor type. These novel cancer driver genes were enriched for chromatin modifiers. Interestingly we found that mutations in chromatin modifiers are frequent in almost all tumor types, with more than 40% of patients presenting with at least one mutated chromatin modifier.

In Chapter 3 we contributed to the fields of cancer heritability and genetic association study in two ways, first by implementing a novel powerful RVAS method, and second, by developing a comprehensive framework (REWAS) facilitating user-friendly and intuitive analysis of case-control cohort studies. We demonstrated that new RVAS method have better power then other methods when tested on datasets that resemble real case-control studies. More importantly, the new method shows moderate to high power for identification of genes that have low percentage of phenotypic variance explained (0.5%) when other methods did not reach exome-wide significance (0.01% of false positives expected). We observed different results in simulated datasets that do not reflect real case-control studies. Using these simulated datasets that are depleted of any kind of noise and utilize randomized disease variants the performance order of tested tools inverted as compared to the more realistic simulation. This might question the validity of many benchmark tests based on purely simulated data for evaluation RVAS methods and conclusions drown from such evaluations are likely not applicable to real WES or WGS datasets. In the end, the best method is the one that can handle noise and biases in the data that are regularly present in WES case-control datasets and hence the simulations need to represent these biases and noise sources.

We did not benchmark REWAS in real case-control study (yet) as we would not be able to measure power without costly replication studies. Nevertheless, we did use REWAS on various disease studies, including chronic lymphocytic leukemia, obsessive compulsive disorder, breast cancer and the PanCancer Analysis of Whole Genomes cohort. For OCD a replication study is in process, for breast cancer results were highly overlapping between methods but of unknown functional significance (the cohort was also of very limited size), and for CLL and PCAWG the results are still being analyzed (although unfortunately no ‘low hanging fruit’ was found which immediately jumped into the eye).

Here we neglected any analysis of common variants which also play an important role in cancer predisposition [Varghese and Easton, 2010, Chang et al., 2014]. However, there is strong evidence that mechanisms for cancer predisposition genes with rare risk variants and high effects differ from cancer predispositions genes that have common risk variants and low effects [Rahman, 2014]. Thus, their analysis and identification can be done separately without loosing power or information about disease mechanisms (e.g. pathways analysis). Nonetheless our REWAS framework is readily able to include common variants in the association test, or to integrate other biological features which in the future could help to improve results.

4.1 Future perspective

In this work somatic and germline events in cancer genomes have been considered and analyzed separately. It is clear that a cancer patient is a carrier of both types of variants, a fact that in this work has been neglected. It has been shown that there is a great overlap between cancer predisposition genes and cancer driver genes [Rahman, 2014]. Moreover, the "Two-Hit Hypothesis" [Knudson, 1971] proposes that tumor suppressor genes are often affected by a Heterozygous germline variant affecting one allele, while the function of the second allele is lost later in life by a somatic event. Finally, it has been demonstrated that germline variants in some genes can affect somatic mutation profiles (e.g. variants in the APOBEC3 gene cluster [Middlebrooks et al., 2016]). Some attempts to integrate these two types of cancer mutations within a single statistical analysis of an ovarian cancer cohort gave promising results [Kanchi et al., 2014]. Recently, germline-somatic interactions was used in almost 6,000 patients to infer germline variants that affect cancer evolution (without any controls dataset) and classifying tumors based on germline profiles revealed new driver genes [Carter et al., 2017]. I have contributed myself to a study (manuscript in preparation) by the PCAWG consortium in which the analysis of germline and somatic mutations in 2834 WGS datasets revealed new connections between germline variants and somatic mutation profiles. This is a promising sign that cancer genomics will benefit from aggregating information from multiple sources (like different types of alternations, somatic and germline mutations, multiple signatures of selection, other OMICs types etc.). More complex models and more complex 'Big Data' approaches will be necessary to tackle this type of multi-dimensional data analysis. Another way to approach complex datasets would be 'divide and conquer', where smaller problems are first solved (e.g. every method applied to only one signature) and then results from multiple sources are aggregated. Tamborero et al. showed that this approach, sometimes termed 'ensemble method' can yield novel results, by showing that combining results of driver prediction methods was superior to individual results of any method alone [Tamborero et al., 2013b]. However, using this approach any interaction between signatures is lost and hence we propose to continue with the development of integrative statistical analysis methods using Bayesian inference and other machine learning methods.

Chapter 5

CONCLUSIONS

- Signatures of tumor evolution are highly informative for prediction of cancer driver genes, as mutations ‘beneficial’ for cancer cells are showing properties of positive selection.

- Cancer cell fraction of somatic mutations is one of the signatures of positive selection, and it can be used to approximate fitness gains by different genotypes detected in heterogeneous cancer tissues.

- Exploitation of multiple signatures of positive selection helps to generate a comprehensive landscape of cancer driver mutations and genes, especially for drivers that have low recurrence (‘long tail of driver genes’).

- Genes identified as drivers in a particular tumor type have a higher probability to be drivers for other tumor types. Therefore, exploiting this prior knowledge allowed us to identify new connections between tumor types and lowly recurrent mutated genes.

- The R package (cDriver) was implemented, using Bayesian statistical inference and three signatures of positive selection to predict cancer driver genes. The package additionally provides a sophisticated background mutations rate model, a CCF calculation method, various visualizations of results and input data.

- Quality Control and filtering methods are of great importance for statistical analysis of case-control studies, as natural and/or artificial noise are evident in almost all datasets. Failing to address this issue will lead to high false positive rates in genotype-phenotype association studies.

- Biological characteristics of genetic variants (e.g. functional impact) improve the identification of disease risk genes in association tests.

- The BATI approach developed in this thesis, using Bayesian Inference combined with parameter estimation by Integrated Nested Laplace Approximations (INLA) approach, demonstrated to have the best power to detect risk genes in a benchmark using a real WES cohort and ClinVar risk variants.
- We developed the REWAS framework, an all-in-one solution for association studies, offering modules for: quality control and filtration of case-control datasets, preparation of covariates and variant characteristics for downstream analysis, and five different rare-variant association tests.

APPENDIX

List of publications during PhD studies

Published

Willmann, M., Bezdan, D., Zapata, L., **Susak, H.**, Vogel, W., Schröppel, K., Liese, J., Weidenmaier, C., Autenrieth, I.B. & Ossowski, S. (2015). Analysis of a long-term outbreak of XDR Pseudomonas aeruginosa: a molecular epidemiological study, *Journal of Antimicrobial Chemotherapy*. 70(5):1322–1330

Accepted

Zapata, L.*, **Susak, H.***, Drechsel, O., Friedländer, M., Estivill, X., Ossowski, S. (2016). Signatures of positive selection reveal a universal role of chromatin modifiers as cancer driver genes. *Scientific Reports*

List of manuscripts in preparation

To be submitted

As participant at PCAWG (`dcc.icgc.org/pcawg`) work package 8 paper:
At *Nature*: Germline determinants of the somatic mutation landscape in 2,642 cancer genomes.

Sepahi, I., Faust, U., Sturm, M., Bosse, K., Kehrer, M, Heilig, M., Grundman-Hauser, K., Bauer, P., Ossowski, S., **Susak, H.**, Bick, U., Schröck, E., Niederacher, D., Auber, B., Sutter, C., Arnold, N., Hahnen, E., Dworniczak, B., Wang-Gorke, S., Gehrig, A., Weber, B. H.F., Engel, C., Lemke, J., Nguyen, H. H. P., Riess, O., Schroeder, C. Search for rare genetic variants in DNA repair genes in *BRCA1* mutation carriers with early and late age at onset of Breast cancer.

In preparation

Susak, H.*, Escaramis-Babiano, G.*, Serra-Saurina L., Bosio, M., Rabionet, K., Domenech-Salgado, L., Ozkan, S., Estivill, X., Ossowski S., Bayesian Rare Variant Association Test using Integrated Nested Laplace Approximation.

Bibliography

[Alexandrov et al., 2013] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Borresen-Dale, A.-L. L., Boyault, S., Burkhardt, B., Butler, A. P., Caldas, C., Davies, H. R., Desmedt, C., Eils, R., Eyfjörd, J. E., Foekens, J. A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D. T. W., Jones, D., Knappskog, S., Kool, M., Lakhani, S. R., López-Otín, C., Martin, S., Munshi, N. C., Nakamura, H., Northcott, P. A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J. V., Puente, X. S., Raine, K., Ramakrishna, M., Richardson, A. L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T. N., Span, P. N., Teague, J. W., Totoki, Y., Tutt, A. N. J., Valdés-Mas, R., van Buuren, M. M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L. R., Zucman-Rossi, J., Futreal, P. A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S. M., Siebert, R., Campo, E., Shibata, T., Pfister, S. M., Campbell, P. J., and Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–21.

[Alioto et al., 2014] Alioto, T. S., Derdak, S., Beck, T. A., Boutros, P. C., Bower, L., Buchhalter, I., Eldridge, M. D., Harding, N. J., Heisler, L. E., and Hovig, E. (2014). A comprehensive assessment of somatic mutation calling in cancer genomes. *bioRxiv*, page 012997.

[Álvarez Silva et al., 2015] Álvarez Silva, M. C., Yepes, S., Torres, M. M., and Barrios, A. F. G. (2015). Proteins interaction network and modeling of igvh mutational status in chronic lymphocytic leukemia. *Theor Biol Med Model*, 12:12.

[Anders and Huber, 2010] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.*, 11(10):R106.

[Aran et al., 2015] Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat Commun*, 6:8971.

[Arsenic et al., 2015] Arsenic, R., Treue, D., Lehmann, A., Hummel, M., Dietel, M., Denkert, C., and Budczies, J. (2015). Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC Clin Pathol*, 15:20.

[Ascierto et al., 2012] Ascierto, P. A., Kirkwood, J. M., Grob, J. J., Simeone, E., Grimaldi, A. M., Maio, M., Palmieri, G., Testori, A., Marincola, F. M., and Mozzillo, N. (2012). The role of BRAF V600 mutation in melanoma. *J Transl Med*, 10:85.

[Auer and Lettre, 2015] Auer, P. L. and Lettre, G. (2015). Rare variant association studies: considerations, challenges and opportunities. *Genome Med*, 7(1):16.

[Babenko et al., 2006] Babenko, V. N., Basu, M. K., Kondrashov, F. A., Rogozin, I. B., and Koonin, E. V. (2006). Signs of positive selection of somatic mutations in human cancers detected by est sequence analysis. *BMC Cancer*, 6:36.

[Balmain, 2001] Balmain, A. (2001). Cancer genetics: from boveri and mendel to microarrays. *Nature Reviews Cancer*, 1(1):77–82.

[Balmain and Nagase, 1998] Balmain, A. and Nagase, H. (1998). Cancer resistance genes in mice: models for the study of tumour modifiers. *Trends Genet.*, 14(4):139–144.

[Bassaganyas et al., 2013] Bassaganyas, L., Beà, S., Escaramís, G., Tornador, C., Salaverria, I., Zapata, L., Drechsel, O., Ferreira, P. G., Rodriguez-Santiago, B., Tubio, J. M. C., Navarro, A., Martín-García, D., López, C., Martínez-Trillos, A., López-Guillermo, A., Gut, M., Ossowski, S., López-Otín, C., Campo, E., and Estivill, X. (2013). Sporadic and reversible chromothripsis in chronic lymphocytic leukemia revealed by longitudinal genomic analysis. *Leukemia*.

[Beerenwinkel et al., 2015] Beerenwinkel, N., Schwarz, R. F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: mathematical models and computational inference. *Systematic biology*, 64(1):e1–e25.

[Biegel et al., 2014] Biegel, J. A., Busse, T. M., and Weissman, B. E. (2014). Swi/snf chromatin remodeling complexes and cancer. *Am J Med Genet C Semin Med Genet*, 166C(3):350–66.

[Bishop, 1981] Bishop, J. M. (1981). Enemies within: the genesis of retrovirus oncogenes. *Cell*, 23(1):5–6.

[Bissell and Hines, 2011] Bissell, M. J. and Hines, W. C. (2011). Why don't we get more cancer? a proposed role of the microenvironment in restraining cancer progression. *Nature medicine*, 17(3):320–329.

[Blangiardo et al., 2013] Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spat Spatiotemporal Epidemiol*, 4:33–49.

[Bodmer and Tomlinson, 2010] Bodmer, W. and Tomlinson, I. (2010). Rare genetic variants and the risk of cancer. *Curr. Opin. Genet. Dev.*, 20(3):262–267.

[Bodnar et al., 1998] Bodnar, A. G., Ouellette, M., Frolkis, M., Holt, S. E., Chiu, C. P., Morin, G. B., Harley, C. B., Shay, J. W., Lichtsteiner, S., and Wright, W. E. (1998). Extension of life-span by introduction of telomerase into normal human cells. *Science*, 279(5349):349–352.

[Bolli et al., 2014] Bolli, N., Avet-Loiseau, H., Wedge, D. C., Van Loo, P., Alexandrov, L. B., Martincorena, I., Dawson, K. J., Iorio, F., Nik-Zainal, S., Bignell, G. R., Hinton, J. W., Li, Y., Tubio, J. M. C., McLaren, S., O' Meara, S., Butler, A. P., Teague, J. W., Mudie, L., Anderson, E., Rashid, N., Tai, Y.-T. T., Shammas, M. A., Sperling, A. S., Fulciniti, M., Richardson, P. G., Parmigiani, G., Magrangeas, F., Minvielle, S., Moreau, P., Attal, M., Facon, T., Futreal, P. A., Anderson, K. C., Campbell, P. J., and Munshi, N. C. (2014). Heterogeneity of genomic evolution and mutational profiles in multiple myeloma. *Nat Commun*, 5:2997.

[Boveri, 1914] Boveri, T. (1914). *Zur frage der entstehung maligner tumoren*. Gustav Fischer.

[Broca, 1866] Broca, P. (1866). *Traité des tumeurs*. Number v. 1 in Traité des tumeurs. P. Asselin.

[Cai et al., 2014] Cai, Y., Geutjes, E. J., De Lint, K., Roepman, P., Bruurs, L., Yu, L. R., Wang, W., van Blijswijk, J., Mohammad, H., and de Rink, I. (2014). The nurd complex cooperates with dnmts to maintain silencing of key colorectal tumor suppressor genes. *Oncogene*, 33(17):2157–2168.

[Campbell et al., 2010] Campbell, P. J., Yachida, S., Mudie, L. J., Stephens, P. J., Pleasance, E. D., Stebbings, L. A., Morsberger, L. A., Latimer, C., McLaren, S., Lin, M.-L. L., McBride, D. J., Varela, I., Nik-Zainal, S. A., Leroy, C., Jia, M., Menzies, A., Butler, A. P., Teague, J. W., Griffin, C. A., Burton, J., Swerdlow, H., Quail, M. A., Stratton, M. R., Iacobuzio-Donahue, C., and Futreal, P. A. (2010). The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature*, 467(7319):1109–13.

[Carter et al., 2017] Carter, H., Marty, R., Hofree, M., Gross, A. M., Jensen, J., Fisch, K. M., Wu, X., DeBoever, C., Van Nostrand, E. L., Song, Y., Wheeler, E., Kreisberg,

J. F., Lippman, S. M., Yeo, G. W., Gutkind, J. S., and Ideker, T. (2017). Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov*, 7(4):410–423.

[Carter et al., 2012] Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., and Getz, G. (2012). Absolute quantification of somatic dna alterations in human cancer. *Nat Biotechnol*, 30(5):413–21.

[Chang et al., 2014] Chang, C. Q., Yesupriya, A., Rowell, J. L., Pimentel, C. B., Clyne, M., Gwinn, M., Khoury, M. J., Wulf, A., and Schully, S. D. (2014). A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur. J. Hum. Genet.*, 22(3):402–408.

[Chang et al., 2016] Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Gao, J., Socci, N. D., Solit, D. B., Olshen, A. B., Schultz, N., and Taylor, B. S. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.*, 34(2):155–163.

[Chassaing et al., 2016] Chassaing, N., Davis, E. E., McKnight, K. L., Niederriter, A. R., Causse, A., David, V., Desmaison, A., Lamarre, S., Vincent-Delorme, C., Pasquier, L., Coubes, C., Lacombe, D., Rossi, M., Dufier, J. L., Dollfus, H., Kaplan, J., Katsanis, N., Etchevers, H. C., Faguer, S., and Calvas, P. (2016). Targeted resequencing identifies PTCH1 as a major contributor to ocular developmental anomalies and extends the SOX2 regulatory network. *Genome Res.*, 26(4):474–485.

[Chudnovsky et al., 2014] Chudnovsky, Y., Kim, D., Zheng, S., Whyte, W. A., Bansal, M., Bray, M.-A. A., Gopal, S., Theisen, M. A., Bilodeau, S., Thiru, P., Muffat, J., Yilmaz, O. H., Mitalipova, M., Woolard, K., Lee, J., Nishimura, R., Sakata, N., Fine, H. A., Carpenter, A. E., Silver, S. J., Verhaak, R. G. W., Califano, A., Young, R. A., Ligon, K. L., Mellinghoff, I. K., Root, D. E., Sabatini, D. M., Hahn, W. C., and Chheda, M. G. (2014). Zfhx4 interacts with the nurd core member chd4 and regulates the glioblastoma tumor-initiating cell state. *Cell Rep*, 6(2):313–24.

[Cibulskis et al., 2013] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31(3):213–219.

[Cohen et al., 2004] Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R., and Hobbs, H. H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–872.

[Comino-Mendez et al., 2011] Comino-Mendez, I., Gracia-Aznarez, F. J., Schiavi, F., Landa, I., Leandro-Garcia, L. J., Leton, R., Honrado, E., Ramos-Medina, R., Caronia, D., Pita, G., Gomez-Grana, A., de Cubas, A. A., Inglada-Perez, L., Maliszewska, A., Taschin, E., Bobisse, S., Pica, G., Loli, P., Hernandez-Lavado, R., Diaz, J. A., Gomez-Morales, M., Gonzalez-Neira, A., Roncador, G., Rodriguez-Antona, C., Benitez, J., Mannelli, M., Opocher, G., Robledo, M., and Cascon, A. (2011). Exome sequencing identifies MAX mutations as a cause of hereditary pheochromocytoma. *Nat. Genet.*, 43(7):663–667.

[Cordova-Alarcon et al., 2005] Cordova-Alarcon, E., Centeno, F., Reyes-Esparza, J., Garcia-Carranca, A., and Garrido, E. (2005). Effects of HRAS oncogene on cell cycle progression in a cervical cancer-derived cell line. *Arch. Med. Res.*, 36(4):311–316.

[Couch et al., 2014] Couch, F. J., Nathanson, K. L., and Offit, K. (2014). Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science*, 343(6178):1466–1470.

[Cowles and Carlin, 1996] Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

[Cox et al., 2005] Cox, C., Bignell, G., Greenman, C., Stabenau, A., Warren, W., Stephens, P., Davies, H., Watt, S., Teague, J., Edkins, S., Birney, E., Easton, D. F., Wooster, R., Futreal, P. A., and Stratton, M. R. (2005). A survey of homozygous deletions in human cancer genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 102(12):4542–4547.

[Cruchaga et al., 2014] Cruchaga, C., Karch, C. M., Jin, S. C., Benitez, B. A., Cai, Y., Guerreiro, R., Harari, O., Norton, J., Budde, J., Bertelsen, S., Jeng, A. T., Cooper, B., Skorupa, T., Carrell, D., Levitch, D., Hsu, S., Choi, J., Ryten, M., Sassi, C., Bras, J., Gibbs, R. J., Hernandez, D. G., Lupton, M. K., Powell, J., Forabosco, P., Ridge, P. G., Corcoran, C. D., Tschanz, J. T., Norton, M. C., Munger, R. G., Schmutz, C., Leary, M., Demirci, F. Y., Bamne, M. N., Wang, X., Lopez, O. L., Ganguli, M., Medway, C., Turton, J., Lord, J., Braae, A., Barber, I., Brown, K., Pastor, P., Lorenzo-Betancor, O., Brkanac, Z., Scott, E., Topol, E., Morgan, K., Rogaeva, E., Singleton, A., Hardy, J., Kamboh, M. I., George-Hyslop, P. S., Cairns, N., Morris, J. C., Kauwe, J. S. K., and Goate, A. M. (2014). Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer's disease. *Nature*, 505(7484):550–554.

[Cybulski et al., 2015] Cybulski, C., Carrot-Zhang, J., Klu?niak, W., Rivera, B., Kashyap, A., Woko?orczyk, D., Giroux, S., Nadaf, J., Hamel, N., Zhang, S., Huzarski, T., Gronwald, J., Byrski, T., Szwiec, M., Jakubowska, A., Rudnicka, H., Lener, M., Masoj?, B., Tonin, P. N., Rousseau, F., Gorski, B., D?bniak, T., Majewski, J., Lubi?ski, J., Foulkes, W. D., Narod, S. A., and Akbari, M. R. (2015). Germline RECQL mutations are associated with breast cancer susceptibility. *Nat. Genet.*, 47(6):643–646.

[Davies et al., 2002] Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W., Davis, N., Dicks, E., Ewing, R., Floyd, Y., Gray, K., Hall, S., Hawes, R., Hughes, J., Kosmidou, V., Menzies, A., Mould, C., Parker, A., Stevens, C., Watt, S., Hooper, S., Wilson, R., Jayatilake, H., Gusterson, B. A., Cooper, C., Shipley, J., Hargrave, D., Pritchard-Jones, K., Maitland, N., Chenevix-Trench, G., Riggins, G. J., Bigner, D. D., Palmieri, G., Cossu, A., Flanagan, A., Nicholson, A., Ho, J. W., Leung, S. Y., Yuen, S. T., Weber, B. L., Seigler, H. F., Darrow, T. L., Paterson, H., Marais, R., Marshall, C. J., Wooster, R., Stratton, M. R., and Futreal, P. A. (2002). Mutations of the BRAF gene in human cancer. *Nature*, 417(6892):949–954.

[de Miranda et al., 2014] de Miranda, N. F., Georgiou, K., Chen, L., Wu, C., Gao, Z., Zaravinos, A., Lisboa, S., Enblad, G., Teixeira, M. R., and Zeng, Y. (2014). Exome sequencing reveals novel mutation targets in diffuse large b-cell lymphomas derived from chinese patients. *Blood*, 124(16):2544–2553.

[Decker et al., 2017] Decker, B., Allen, J., Luccarini, C., Pooley, K. A., Shah, M., Bolla, M. K., Wang, Q., Ahmed, S., Baynes, C., Conroy, D. M., Brown, J., Luben, R., Ostrander, E. A., Pharoah, P. D., Dunning, A. M., and Easton, D. F. (2017). Rare, protein-truncating variants in ATM, CHEK2 and PALB2, but not XRCC2, are associated with increased breast cancer risks. *J. Med. Genet.*

[Dees et al., 2012] Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., and Ding, L. (2012). Music: identifying mutational significance in cancer genomes. *Genome Res*, 22(8):1589–98.

[DePristo et al., 2011] DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, 43(5):491–498.

[Dieci et al., 2016] Dieci, M. V., Smutná, V., Scott, V., Yin, G., Xu, R., Vielh, P., Mathieu, M.-C. C., Vicier, C., Laporte, M., Drusch, F., Guarneri, V., Conte, P., Delaloge, S., Lacroix, L., Fromigué, O., André, F., and Lefebvre, C. (2016). Whole exome sequencing of rare aggressive breast cancer histologies. *Breast Cancer Res Treat*, 156(1):21–32.

[Dohm et al., 2008] Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, 36(16):e105.

[Easton et al., 1993] Easton, D. F., Bishop, D. T., Ford, D., and Crockford, G. P. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.*, 52(4):678–701.

[Economopoulou et al., 2015] Economopoulou, P., Dimitriadis, G., and Psyrri, A. (2015). Beyond BRCA: new hereditary breast cancer susceptibility genes. *Cancer Treat. Rev.*, 41(1):1–8.

[Eicher et al., 2015] Eicher, J. D., Landowski, C., Stackhouse, B., Sloan, A., Chen, W., Jensen, N., Lien, J. P., Leslie, R., and Johnson, A. D. (2015). GRASP v2.0: an update on the Genome-Wide Repository of Associations between SNPs and phenotypes. *Nucleic Acids Res.*, 43(Database issue):799–804.

[Eilbeck et al., 2017] Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nat. Rev. Genet.*, 18(10):599–612.

[Fearon and Vogelstein, 1990] Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767.

[Ferlay et al., 2013a] Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., and Bray, F. (2013a). Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. available from: http://globocan.iarc.fr, accessed on 7/9/2017. *Lyon, France: International Agency for Research on Cancer*.

[Ferlay et al., 2013b] Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J., Comber, H., Forman, D., and Bray, F. (2013b). Cancer incidence and mortality patterns in europe: Estimates for 40 countries in 2012. *European Journal of Cancer*, 49(6):1374 – 1403.

[Fernandez et al., 2016] Fernandez, L. C., Torres, M., and Real, F. X. (2016). Somatic mosaicism: on the road to cancer. *Nat. Rev. Cancer*, 16(1):43–55.

[Fialkow, 1979] Fialkow, P. J. (1979). Clonal origin of human tumors. *Annu. Rev. Med.*, 30:135–143.

[Fischer et al., 2014] Fischer, A., Vázquez-García, I., Illingworth, C. J. R., and Mustonen, V. (2014). High-definition reconstruction of clonal composition in cancer. *Cell Rep*.

[Forment et al., 2012] Forment, J. V., Kaidi, A., and Jackson, S. P. (2012). Chromothripsis and cancer: causes and consequences of chromosome shattering. *Nat. Rev. Cancer*, 12(10):663–670.

[Friend et al., 1986] Friend, S. H., Bernards, R. A., Rogelj, S., Weinberg, R. A., Rapaport, J. M., Albert, D. M., and Dryja, T. P. (1986). A human dna segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature*, 323(6089):643–646.

[Futreal et al., 2004] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–83.

[Gerlinger et al., 2012] Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., and Tarpey, P. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, 366(10):883–892.

[Gerlinger and Swanton, 2010] Gerlinger, M. and Swanton, C. (2010). How darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine. *Br J Cancer*, 103(8):1139–43.

[Gerstung et al., 2012] Gerstung, M., Beisel, C., Rechsteiner, M., Wild, P., Schraml, P., Moch, H., and Beerenwinkel, N. (2012). Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nat Commun*, 3:811.

[Gonzaga-Jauregui et al., 2012] Gonzaga-Jauregui, C., Lupski, J. R., and Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annu. Rev. Med.*, 63:35–61.

[Gonzalez-Perez and Lopez-Bigas, 2012] Gonzalez-Perez, A. and Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Res*, 40(21):e169.

[Gonzalez-Perez et al., 2013] Gonzalez-Perez, A., Mustonen, V., Reva, B., Ritchie, G. R. S., Creixell, P., Karchin, R., Vazquez, M., Fink, J. L., Kassahn, K. S., Pearson, J. V., Bader, G. D., Boutros, P. C., Muthuswamy, L., Ouellette, B. F. F., Reimand, J., Linding, R., Shibata, T., Valencia, A., Butler, A., Dronov, S., Flicek, P., Shannon, N. B., Carter, H., Ding, L., Sander, C., Stuart, J. M., Stein, L. D., and Lopez-Bigas, N.

(2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nat Methods*, 10(8):723–9.

[Greaves and Maley, 2012] Greaves, M. and Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381):306–313.

[Gudmundsson et al., 2012] Gudmundsson, J., Sulem, P., Gudbjartsson, D. F., Masson, G., Agnarsson, B. A., Benediktsdottir, K. R., Sigurdsson, A., Magnusson, O. T., Gudjonsson, S. A., Magnusdottir, D. N., Johannsdottir, H., Helgadottir, H. T., Stacey, S. N., Jonasdottir, A., Olafsdottir, S. B., Thorleifsson, G., Jonasson, J. G., Tryggvadottir, L., Navarrete, S., Fuertes, F., Helfand, B. T., Hu, Q., Csiki, I. E., Mates, I. N., Jinga, V., Aben, K. K., van Oort, I. M., Vermeulen, S. H., Donovan, J. L., Hamdy, F. C., Ng, C. F., Chiu, P. K., Lau, K. M., Ng, M. C., Gulcher, J. R., Kong, A., Catalona, W. J., Mayordomo, J. I., Einarsson, G. V., Barkardottir, R. B., Jonsson, E., Mates, D., Neal, D. E., Kiemeney, L. A., Thorsteinsdottir, U., Rafnar, T., and Stefansson, K. (2012). A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat. Genet.*, 44(12):1326–1329.

[Gunby et al., 2007] Gunby, R. H., Sala, E., Tartari, C. J., Puttini, M., Gambacorti-Passerini, C., and Mologni, L. (2007). Oncogenic fusion tyrosine kinases as molecular targets for anti-cancer therapy. *Anticancer Agents Med Chem*, 7(6):594–611.

[Hanahan and Weinberg, 2011] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74.

[Hanks et al., 2004] Hanks, S., Coleman, K., Reid, S., Plaja, A., Firth, H., Fitzpatrick, D., Kidd, A., Mehes, K., Nash, R., Robin, N., Shannon, N., Tolmie, J., Swansbury, J., Irrthum, A., Douglas, J., and Rahman, N. (2004). Constitutional aneuploidy and cancer predisposition caused by biallelic mutations in BUB1B. *Nat. Genet.*, 36(11):1159–1161.

[Heisterkamp et al., 1983] Heisterkamp, N., Stephenson, J. R., Groffen, J., Hansen, P. F., de Klein, A., Bartram, C. R., and Grosveld, G. (1983). Localization of the c-abl oncogene adjacent to a translocation break point in chronic myelocytic leukaemia. *Nature*, 306(5940):239–242.

[Hodgson, 2008] Hodgson, S. (2008). Mechanisms of inherited cancer susceptibility. *J Zhejiang Univ Sci B*, 9(1):1–4.

[Hogenbirk et al., 2016] Hogenbirk, M. A., Heideman, M. R., de Rink, I., Velds, A., Kerkhoven, R. M., Wessels, L. F., and Jacobs, H. (2016). Defining chromosomal translocation risks in cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 113(26):E3649–3656.

[Holtzman and Marteau, 2000] Holtzman, N. A. and Marteau, T. M. (2000). Will genetics revolutionize medicine? *N. Engl. J. Med.*, 343(2):141–144.

[Iyer et al., 2006] Iyer, R. R., Pluciennik, A., Burdett, V., and Modrich, P. L. (2006). DNA mismatch repair: functions and mechanisms. *Chem. Rev.*, 106(2):302–323.

[Jonsson et al., 2012] Jonsson, T., Atwal, J. K., Steinberg, S., Snaedal, J., Jonsson, P. V., Bjornsson, S., Stefansson, H., Sulem, P., Gudbjartsson, D., Maloney, J., Hoyte, K., Gustafson, A., Liu, Y., Lu, Y., Bhangale, T., Graham, R. R., Huttenlocher, J., Bjornsdottir, G., Andreassen, O. A., Jonsson, E. G., Palotie, A., Behrens, T. W., Magnusson, O. T., Kong, A., Thorsteinsdottir, U., Watts, R. J., and Stefansson, K. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature*, 488(7409):96–99.

[Kanchi et al., 2014] Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D., Wendl, M. C., Zhang, Q., Koboldt, D. C., Xie, M., Kandoth, C., McMichael, J. F., Wyczalkowski, M. A., Larson, D. E., Schmidt, H. K., Miller, C. A., Fulton, R. S., Spellman, P. T., Mardis, E. R., Druley, T. E., Graubert, T. A., Goodfellow, P. J., Raphael, B. J., Wilson, R. K., and Ding, L. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun*, 5:3156.

[Kandoth et al., 2013] Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., Shen, H., Robertson, A. G., Pashtan, I., Shen, R., Benz, C. C., Yau, C., Laird, P. W., Ding, L., Zhang, W., Mills, G. B., Kucherlapati, R., Mardis, E. R., and Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, 497(7447):67–73.

[Kim et al., 2013] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36.

[Kircher et al., 2014] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46(3):310–5.

[Knudson, 1971] Knudson, A. G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823.

[Koboldt et al., 2012a] Koboldt, D. C., Larson, D. E., Chen, K., Ding, L., and Wilson, R. K. (2012a). Massively parallel sequencing approaches for characterization of structural variation. *Methods Mol. Biol.*, 838:369–384.

[Koboldt et al., 2012b] Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012b).

VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, 22(3):568–576.

[Kontham et al., 2013] Kontham, V., von Holst, S., and Lindblom, A. (2013). Linkage analysis in familial non-Lynch syndrome colorectal cancer families from Sweden. *PLoS ONE*, 8(12):e83936.

[Korthauer and Kendziorski, 2015] Korthauer, K. D. and Kendziorski, C. (2015). MADGiC: a model-based approach for identifying driver genes in cancer. *Bioinformatics*, 31(10):1526–1535.

[Kulis and Esteller, 2010] Kulis, M. and Esteller, M. (2010). DNA methylation and cancer. *Adv. Genet.*, 70:27–56.

[Lai and Wade, 2011] Lai, A. Y. and Wade, P. A. (2011). Cancer biology and nurd: a multifaceted chromatin remodelling complex. *Nat Rev Cancer*, 11(8):588–96.

[Landau et al., 2015] Landau, D. A., Tausch, E., Taylor-Weiner, A. N., Stewart, C., Reiter, J. G., Bahlo, J., Kluth, S., Bozic, I., Lawrence, M., Böttcher, S., Carter, S. L., Cibulskis, K., Mertens, D., Sougnez, C. L., Rosenberg, M., Hess, J. M., Edelmann, J., Kless, S., Kneba, M., Ritgen, M., Fink, A., Fischer, K., Gabriel, S., Lander, E. S., Nowak, M. A., Döhner, H., Hallek, M., Neuberg, D., Getz, G., Stilgenbauer, S., and Wu, C. J. (2015). Mutations driving cll and their evolution in progression and relapse. *Nature*, 526(7574):525–30.

[Landau and Wu, 2013] Landau, D. A. and Wu, C. J. (2013). Chronic lymphocytic leukemia: molecular heterogeneity revealed by high-throughput genomics. *Genome Med*, 5(5):47.

[Lander et al., 2001] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty,

K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.-F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H.-C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G. R., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F. A., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S.-P., Yeh, R.-F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Patrinos, A., and Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359.

[Langmead et al., 2009] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25.

[Larrea et al., 2010] Larrea, A. A., Lujan, S. A., and Kunkel, T. A. (2010). SnapShot: DNA mismatch repair. *Cell*, 141(4):730.e1.

[Larson et al., 2012] Larson, D. E., Harris, C. C., Chen, K., Koboldt, D. C., Abbott, T. E., Dooling, D. J., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2012).

SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*, 28(3):311–317.

[Lawrence et al., 2014] Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501.

[Lawrence et al., 2013] Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., Kim, J., Sougnez, C., Ambrogio, L., Nickerson, E., Shefler, E., Cortés, M. L., Auclair, D., Saksena, G., Voet, D., Noble, M., Dicara, D., Lin, P., Lichtenstein, L., Heiman, D. I., Fennell, T., Imielinski, M., Hernandez, B., Hodis, E., Baca, S., Dulak, A. M., Lohr, J., Landau, D.-A. A., Wu, C. J., Melendez-Zajgla, J., Hidalgo-Miranda, A., Koren, A., McCarroll, S. A., Mora, J., Lee, R. S., Crompton, B., Onofrio, R., Parkin, M., Winckler, W., Ardlie, K., Gabriel, S. B., Roberts, C. W. M., Biegel, J. A., Stegmaier, K., Bass, A. J., Garraway, L. A., Meyerson, M., Golub, T. R., Gordenin, D. A., Sunyaev, S., Lander, E. S., and Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218.

[Le Gallo et al., 2012] Le Gallo, M., O'Hara, A. J., Rudd, M. L., Urick, M. E., Hansen, N. F., O'Neil, N. J., Price, J. C., Zhang, S., England, B. M., Godwin, A. K., Sgroi, D. C., Hieter, P., Mullikin, J. C., Merino, M. J., and Bell, D. W. (2012). Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat Genet*, 44(12):1310–5.

[Lee et al., 2015] Lee, J.-Y. . Y., Yoon, J.-K. . K., Kim, B., Kim, S., Kim, M. A., Lim, H., Bang, D., and Song, Y.-S. . S. (2015). Tumor evolution and intratumor heterogeneity of an epithelial ovarian cancer investigated using next-generation sequencing. *BMC Cancer*, 15(1).

[Lee et al., 2014] Lee, S., Abecasis, G. R., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.*, 95(1):5–23.

[Lee et al., 2012a] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., and Lin, X. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.*, 91(2):224–237.

[Lee et al., 2012b] Lee, S., Wu, M. C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

[Leiserson et al., 2015] Leiserson, M. D. M., Vandin, F., Wu, H.-T. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., Papoutsaki, A., Kim, Y., Niu, B., McLellan, M., Lawrence, M. S., Gonzalez-Perez, A., Tamborero, D., Cheng, Y., Ryslik, G. A., Lopez-Bigas, N., Getz, G., Ding, L., and Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*, 47(2):106–14.

[Li and Leal, 2008] Li, B. and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*, 83(3):311–21.

[Li and Li, 2014] Li, B. and Li, J. Z. (2014). A general framework for analyzing tumor subclonality using snp array and dna sequencing data. *Genome Biol*, 15(9):473.

[Li and Durbin, 2010] Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li et al., 2012a] Li, X., Zhang, J., Su, C., Zhao, X., Tang, L., and Zhou, C. (2012a). The association between polymorphisms in the DNA nucleotide excision repair genes and RRM1 gene and lung cancer risk. *Thorac Cancer*, 3(3):239–248.

[Li et al., 2012b] Li, Y., Brown, P., Rue, H., al Maini, M., and Fortin, P. (2012b). Spatial modelling of lupus incidence over 40 years with changes in census areas. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(1):99–115.

[Lichtenstein et al., 2000] Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.*, 343(2):78–85.

[Lin et al., 2017] Lin, X., Chen, Z., Gao, P., Gao, Z., Chen, H., Qi, J., Liu, F., Ye, D., Jiang, H., Na, R., Yu, H., Shi, R., Lu, D., Zheng, S. L., Mo, Z., Sun, Y., Ding, Q., and Xu, J. (2017). TEX15: A DNA repair gene associated with prostate cancer risk in Han Chinese. *Prostate*, 77(12):1271–1278.

[Liu and Leal, 2010] Liu, D. J. and Leal, S. M. (2010). A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet.*, 6(10):e1001156.

[Loveday et al., 2011] Loveday, C., Turnbull, C., Ramsay, E., Hughes, D., Ruark, E., Frankum, J. R., Bowden, G., Kalmyrzaev, B., Warren-Perry, M., Snape, K., Adlard, J. W., Barwell, J., Berg, J., Brady, A. F., Brewer, C., Brice, G., Chapman, C., Cook, J., Davidson, R., Donaldson, A., Douglas, F., Greenhalgh, L., Henderson, A., Izatt, L., Kumar, A., Lalloo, F., Miedzybrodzka, Z., Morrison, P. J., Paterson, J., Porteous, M., Rogers, M. T., Shanley, S., Walker, L., Eccles, D., Evans, D. G., Renwick, A., Seal, S., Lord, C. J., Ashworth, A., Reis-Filho, J. S., Antoniou, A. C., and Rahman, N. (2011). Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat. Genet.*, 43(9):879–882.

[Madsen and Browning, 2009] Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, 5(2):e1000384.

[Makalowski and Boguski, 1998] Makalowski, W. and Boguski, M. S. (1998). Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 95(16):9407–9412.

[Manolio et al., 2009] Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

[Martincorena et al., 2017] Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., and Campbell, P. J. (2017). Universal patterns of selection in cancer and somatic tissues. *bioRxiv*.

[Martincorena and Campbell, 2015] Martincorena, I. n. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *scienceScience*, 349(6255):1483–1489.

[Martino et al., 2011] Martino, S., Aas, K., Lindqvist, O., Neef, L. R., and Rue, H. (2011). Estimating stochastic volatility models using integrated nested laplace approximations. *The European Journal of Finance*, 17(7):487–503.

[Martins et al., 2013] Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with inla: New features. *Computational Statistics & Data Analysis*, 67:68 – 83.

[Marusyk et al., 2012] Marusyk, A., Almendro, V., and Polyak, K. (2012). Intratumour heterogeneity: a looking glass for cancer? *Nat Rev Cancer*, 12(5):323–34.

[Maruvka et al., 2017] Maruvka, Y. E., Mouw, K. W., Karlic, R., Parasuraman, P., Kamburov, A., Polak, P., Haradhvala, N. J., Hess, J. M., Rheinbay, E., Brody, Y., Koren, A., Braunstein, L. Z., D'Andrea, A., Lawrence, M. S., Bass, A., Bernards, A., Michor, F., and Getz, G. (2017). Analysis of somatic microsatellite indels identifies driver events in human tumors. *Nat. Biotechnol.*

[Massague, 2008] Massague, J. (2008). TGFbeta in Cancer. *Cell*, 134(2):215–230.

[McGranahan et al., 2015] McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., and Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med*, 7(283):283ra54.

[McKenna et al., 2010] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9):1297–1303.

[McLaren et al., 2010] McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, 26(16):2069–2070.

[Meijers-Heijboer et al., 2002] Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., Elstrodt, F., van Duijn, C., Bartels, C., Meijers, C., Schutte, M., McGuffog, L., Thompson, D., Easton, D., Sodha, N., Seal, S., Barfoot, R., Mangion, J., Chang-Claude, J., Eccles, D., Eeles, R., Evans, D. G., Houlston, R., Murday, V., Narod, S., Peretz, T., Peto, J., Phelan, C., Zhang, H. X., Szabo, C., Devilee, P., Goldgar, D., Futreal, P. A., Nathanson, K. L., Weber, B., Rahman, N., and Stratton, M. R. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.*, 31(1):55–59.

[Merlo et al., 2006] Merlo, L. M. F., Pepper, J. W., Reid, B. J., and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6(12):924–35.

[Meyerson et al., 1997] Meyerson, M., Counter, C. M., Eaton, E. N., Ellisen, L. W., Steiner, P., Caddle, S. D., Ziaugra, L., Beijersbergen, R. L., Davidoff, M. J., Liu, Q., Bacchetti, S., Haber, D. A., and Weinberg, R. A. (1997). hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. *Cell*, 90(4):785–795.

[Middlebrooks et al., 2016] Middlebrooks, C. D., Banday, A. R., Matsuda, K., Udquim, K. I., Onabajo, O. O., Paquin, A., Figueroa, J. D., Zhu, B., Koutros, S., Kubo, M., Shuin, T., Freedman, N. D., Kogevinas, M., Malats, N., Chanock, S. J., Garcia-Closas, M., Silverman, D. T., Rothman, N., and Prokunina-Olsson, L. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.*, 48(11):1330–1338.

[Miki et al., 1994] Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L. M., and Ding, W. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science*, 266(5182):66–71.

[Miller et al., 2014] Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. (2014). Sciclone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol*, 10(8):e1003665.

[Morgenthaler and Thilly, 2007] Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.*, 615(1-2):28–56.

[Morozova and Marra, 2008] Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264.

[Moutsianas et al., 2015] Moutsianas, L., Agarwala, V., Fuchsberger, C., Flannick, J., Rivas, M. A., Gaulton, K. J., Albers, P. K., Consortium, G., McVean, G., Boehnke, M., Altshuler, D., and McCarthy, M. I. (2015). The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLOS Genetics*, 11(4):1–24.

[Nakagawa et al., 2015] Nakagawa, H., Wardell, C. P., Furuta, M., Taniguchi, H., and Fujimoto, A. (2015). Cancer whole-genome sequencing: present and future. *Oncogene*, 34(49):5943–5950.

[Neale et al., 2011] Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.*, 7(3):e1001322.

[Nelson et al., 2012] Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S. A., Fraser, D., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., Topp, S., Hall, M. D., Nangle, K., Wang, J., Abecasis, G., Cardon, L. R., Zollner, S., Whittaker, J. C., Chissoe, S. L., Novembre, J., and Mooser, V. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337(6090):100–104.

[News, 2010] News, N. (2010). Human genome: Genomes by the thousand. *Nature*, 467(7319):1026–1027.

[Nho et al., 2017] Nho, K., Kim, S., Horgusluoglu, E., Risacher, S. L., Shen, L., Kim, D., Lee, S., Foroud, T., Shaw, L. M., Trojanowski, J. Q., Aisen, P. S., Petersen, R. C., Jack, C. R., Weiner, M. W., Green, R. C., Toga, A. W., and Saykin, A. J. (2017). Association analysis of rare variants near the APOE region with CSF and neuroimaging biomarkers of Alzheimer's disease. *BMC Med Genomics*, 10(Suppl 1):29.

[Nielsen et al., 2016] Nielsen, F. C., van Overeem Hansen, T., and Sørensen, C. S. (2016). Hereditary breast and ovarian cancer: new genes in confined pathways. *Nat. Rev. Cancer*, 16(9):599–612.

[Nielsen, 2005] Nielsen, R. (2005). Molecular signatures of natural selection. *Annu Rev Genet*, 39:197–218.

[Nik-Zainal et al., 2016] Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L. B., Martin, S., Wedge, D. C., Van Loo, P., Ju, Y. S., Smid, M., Brinkman, A. B., Morganella, S., Aure, M. R., Lingjaerde, O. C., Langerod, A., Ringnér, M., Ahn, S.-M. M., Boyault, S., Brock, J. E., Broeks, A., Butler, A., Desmedt, C., Dirix, L., Dronov, S., Fatima, A., Foekens, J. A., Gerstung, M., Hooijer, G. K. J., Jang, S. J., Jones, D. R., Kim, H.-Y. Y., King, T. A., Krishnamurthy, S., Lee, H. J., Lee, J.-Y. Y., Li, Y., McLaren, S., Menzies, A., Mustonen, V., O'Meara, S., Pauporté, I., Pivot, X., Purdie, C. A., Raine, K., Ramakrishnan, K., Rodríguez-González, F. G., Romieu, G., Sieuwerts, A. M., Simpson, P. T., Shepherd, R., Stebbings, L., Stefansson, O. A., Teague, J., Tommasi, S., Treilleux, I., Van den Eynden, G. G., Vermeulen, P., Vincent-Salomon, A., Yates, L., Caldas, C., Veer, L. V., Tutt, A., Knappskog, S., Tan, B. K. T., Jonkers, J., Borg, A., Ueno, N. T., Sotiriou, C., Viari, A., Futreal, P. A., Campbell, P. J., Span, P. N., Van Laere, S., Lakhani, S. R., Eyfjord, J. E., Thompson, A. M., Birney, E., Stunnenberg, H. G., van de Vijver, M. J., Martens, J. W. M., Borresen-Dale, A.-L. L., Richardson, A. L., Kong, G., Thomas, G., and Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54.

[Nowell, 1976] Nowell, P. C. (1976). The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28.

[Nowell and Hungerford, 1960] Nowell, P. C. and Hungerford, D. A. (1960). Chromosome studies on normal and leukemic human leukocytes. *Journal of the National Cancer Institute*, 25(1):85–109.

[Oesper et al., 2013] Oesper, L., Mahmoody, A., and Raphael, B. J. (2013). Theta: Inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome Biology*, 14(7):R80.

[Orvis et al., 2014] Orvis, T., Hepperla, A., Walter, V., Song, S., Simon, J., Parker, J., Wilkerson, M. D., Desai, N., Major, M. B., Hayes, D. N., Davis, I. J., and Weissman, B. (2014). Brg1/smarca4 inactivation promotes non-small cell lung cancer aggressiveness by altering chromatin organization. *Cancer Res*, 74(22):6486–98.

[O'Shaughnessy and Hendrich, 2013] O'Shaughnessy, A. and Hendrich, B. (2013). Chd4 in the dna-damage response and cell cycle progression: not so nurdy now. *Biochem Soc Trans*, 41(3):777–82.

[Ostrow et al., 2014] Ostrow, S. L., Barshir, R., DeGregori, J., Yeger-Lotem, E., and Hershberg, R. (2014). Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet*, 10(3):e1004239.

[Pabinger et al., 2013] Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2013). A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform*.

[Paul et al., 2010] Paul, M., Riebler, A., Bachmann, L. M., Rue, H., and Held, L. (2010). Bayesian bivariate meta-analysis of diagnostic test studies using integrated nested laplace approximations. *Statistics in Medicine*, 29(12):1325–1339.

[Pepper et al., 2009] Pepper, J. W., Scott Findlay, C., Kassen, R., Spencer, S. L., and Maley, C. C. (2009). Cancer research meets evolutionary biology. *Evol Appl*, 2(1):62–70.

[Ponder, 2001] Ponder, B. A. (2001). Cancer genetics. *Nature*, 411(6835):336–341.

[Price et al., 2010] Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.*, 86(6):832–838.

[Priest et al., 2016] Priest, J. R., Osoegawa, K., Mohammed, N., Nanda, V., Kundu, R., Schultz, K., Lammer, E. J., Girirajan, S., Scheetz, T., Waggott, D., Haddad, F., Reddy, S., Bernstein, D., Burns, T., Steimle, J. D., Yang, X. H., Moskowitz, I. P., Hurles, M., Lifton, R. P., Nickerson, D., Bamshad, M., Eichler, E. E., Mital, S., Sheffield, V., Quertermous, T., Gelb, B. D., Portman, M., and Ashley, E. A. (2016). De Novo and Rare Variants at Multiple Loci Support the Oligogenic Origins of Atrioventricular Septal Heart Defects. *PLoS Genet.*, 12(4):e1005963.

[Puente et al., 2015] Puente, X. S., Beà, S., Valdés-Mas, R., Villamor, N., Gutiérrez-Abril, J., Martín-Subero, J. I., Munar, M., Rubio-Pérez, C., Jares, P., Aymerich, M., Baumann, T., Beekman, R., Belver, L., Carrio, A., Castellano, G., Clot, G., Colado, E., Colomer, D., Costa, D., Delgado, J., Enjuanes, A., Estivill, X., Ferrando, A. A., Gelpí, J. L., González, B., González, S., González, M., Gut, M., Hernández-Rivas, J. M., López-Guerra, M., Martín-García, D., Navarro, A., Nicolás, P., Orozco, M., Payer, A. R., Pinyol, M., Pisano, D. G., Puente, D. A., Queirós, A. C., Quesada, V., Romeo-Casabona, C. M., Royo, C., Royo, R., Rozman, M., Russiñol, N., Salaverría, I., Stamatopoulos, K., Stunnenberg, H. G., Tamborero, D., Terol, M. J., Valencia, A., López-Bigas, N., Torrents, D., Gut, I., López-Guillermo, A., López-Otín, C., and Campo, E. (2015). Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 526(7574):519–24.

[Pyatnitskiy et al., 2015] Pyatnitskiy, M., Karpov, D., Poverennaya, E., Lisitsa, A., and Moshkovskii, S. (2015). Bringing down cancer aircraft: Searching for essential hypomutated proteins in skin melanoma. *PLoS One*, 10(11):e0142819.

[Rahman, 2014] Rahman, N. (2014). Realizing the promise of cancer predisposition genes. *Nature*, 505(7483):302–8.

[Rahman et al., 2007] Rahman, N., Seal, S., Thompson, D., Kelly, P., Renwick, A., Elliott, A., Reid, S., Spanova, K., Barfoot, R., Chagtai, T., Jayatilake, H., McGuffog, L., Hanks, S., Evans, D. G., Eccles, D., Easton, D. F., and Stratton, M. R. (2007). PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.*, 39(2):165–167.

[Reimand and Bader, 2013] Reimand, J. and Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol*, 9:637.

[Rivas et al., 2011] Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J. M., Kuruvilla, F., Lagace, C., Neale, B., Lo, K. S., Schumm, P., Torkvist, L., Dubinsky, M. C., Brant, S. R., Silverberg, M. S., Duerr, R. H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M., McGovern, D. P., Cho, J. H., Rioux, J. D., Xavier, R. J., Daly, M. J., Brant, S. R., Cho, J. H., Duerr, R. H., McGovern, D. P., Rioux, J. D., Silverberg, M. S., Parkes, M., Lee, J., Zhang, H., Bredin, F., Ahmad, T., Satsangi, J., Nimmo, E., Drummond, H., Lees, C., Mansfield, J., Mathew, C. G., Prescott, N., Harrison, K., Sanderson, J., Newman, W., Phillips, A., Mowat, C., Edwards, C., Wilson, D. C., Barrett, J., Anderson, C., Gray, E., Edkins, S., Russell, R. K., Henderson, P., Ahmad, T., Anderson, C. A., Annese, V., Baldassano, R. N., Balschun, T., Barclay, M., Barrett, J. C., Bayless, T. M., Bis, J. C., Brand, S., Brant, S. R., Bumpstead, S., Buning, C., Cho, J. H., Cohen, A., Colombel, J. F., Cottone, M., D'Amato, M., D'Inca, R., Daly, M. J., Denson, T., Dubinsky, M., Duerr, R. H., Edwards, C., Ellinghaus, D., Florin, T., Franchimont, D., Franke, A., Gearry, R., Georges, M., Glas, J., Van Gossum, A., Griffiths, A. M., Guthery, S. L., Hakonarson, H., Haritunians, T., Hugot, J. P., de Jong, D. J., Jostins, L., Kugathasan, S., Kullack-Ublick, G., Latiano, A., Laukens, D., Lawrance, I., Lee, J., Lees, C. W., Lemann, M., Levine, A., Libioulle, C., Louis, E., Mansfield, J. C., Mathew, C. G., McGovern, D. P., Mitrovic, M., Montgomery, G. W., Mowat, C., Newman, W., Palmieri, O., Panes, J., Parkes, M., Phillips, A., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Proctor, D. D., Radford-Smith, G. L., Regueiro, M., Rioux, J. D., Roberts, R., Rotter, J. I., Rutgeerts, P., Sanderson, J., Sans, M., Satsangi, J., Schreiber, S., Schumm, P., Seibold, F., Sharma, Y., Silverberg, M. S., Simms, L. A., Steinhart, A. H., Targan, S. R., Taylor, K. D., Torkvist, L., Vermeire, S., Halfvarson, J., Verspaget, H. W., De Vos, M., Walters, T., Wang, K., Weersma, R. K., Whiteman, D., and Wijmenga, C. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.*, 43(11):1066–1073.

[Roberts and Orkin, 2004] Roberts, C. W. M. and Orkin, S. H. (2004). The swi/snf complex–chromatin and cancer. *Nat Rev Cancer*, 4(2):133–42.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Roos and Held, 2011] Roos, M. and Held, L. (2011). Sensitivity analysis in bayesian generalized linear mixed models for binary data. *Bayesian Anal.*, 6(2):259–278.

[Roth et al., 2014] Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., and Shah, S. P. (2014). Pyclone: statistical inference of clonal population structure in cancer. *Nature methods*, 11(4):396–398.

[Roth et al., 2012] Roth, E. M., McKenney, J. M., Hanotin, C., Asset, G., and Stein, E. A. (2012). Atorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. *N. Engl. J. Med.*, 367(20):1891–1900.

[Rowley, 1973] Rowley, J. D. (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. *Nature*, 243(5405):290–293.

[Rue and Martino, 2007] Rue, H. and Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of Statistical*

Planning and Inference, 137(10):3177 – 3192. Special Issue: Bayesian Inference for Stochastic Processes.

[Rue et al., 2009] Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

[Ruiz-Cárdenas et al., 2012] Ruiz-Cárdenas, R., Krainski, E. T., and Rue, H. (2012). Direct fitting of dynamic models using integrated nested laplace approximations - inla. *Comput. Stat. Data Anal.*, 56(6):1808–1828.

[Ruiz-Pinto et al., 2017] Ruiz-Pinto, S., Pita, G., Martin, M., Alonso-Gordoa, T., Barnes, D. R., Alonso, M. R., Herraez, B., Garcia-Miguel, P., Alonso, J., Perez-Martinez, A., Carton, A. J., Gutierrez-Larraya, F., Garcia-Saenz, J. A., Benitez, J., Easton, D. F., Patino-Garcia, A., and Gonzalez-Neira, A. (2017). Exome array analysis identifies ETFB as a novel susceptibility gene for anthracycline-induced cardiotoxicity in cancer patients. *Breast Cancer Res. Treat.*

[Sabarinathan et al., 2016] Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., and Lopez-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, 532(7598):264–267.

[Sakoparnig et al., 2015] Sakoparnig, T., Fried, P., and Beerenwinkel, N. (2015). Identification of constrained cancer driver genes based on mutation timing. *PLoS Comput Biol*, 11(1):e1004027.

[Sanger et al., 1977] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467.

[Saunders et al., 2012] Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*, 28(14):1811–1817.

[Schrödle and Held, 2011] Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using inla. *Environmetrics*, 22(6):725–734.

[Schrödle et al., 2011] Schrödle, B., Held, L., Riebler, A., and Danuser, J. (2011). Using integrated nested laplace approximations for the evaluation of veterinary surveillance data from switzerland: a case-study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(2):261–279.

[Schuh et al., 2012] Schuh, A., Becq, J., Humphray, S., Alexa, A., Burns, A., Clifford, R., Feller, S. M., Grocock, R., Henderson, S., Khrebtukova, I., Kingsbury, Z., Luo, S., McBride, D., Murray, L., Menju, T., Timbs, A., Ross, M., Taylor, J., and Bentley, D. (2012). Monitoring chronic lymphocytic leukemia progression by whole genome sequencing reveals heterogeneous clonal evolution patterns. *Blood*, 120(20):4191–6.

[Seal et al., 2006] Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., North, B., McGuffog, L., Evans, D. G., Eccles, D., Easton, D. F., Stratton, M. R., and Rahman, N. (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.*, 38(11):1239–1241.

[Shtir et al., 2016] Shtir, C., Aldahmesh, M. A., Al-Dahmash, S., Abboud, E., Alkuraya, H., Abouammoh, M. A., Nowailaty, S. R., Al-Thubaiti, G., Naim, E. A., ALYounes, B., Binhumaid, F. S., ALOtaibi, A. B., Altamimi, A. S., Alamer, F. H., Hashem, M., Abouelhoda, M., Monies, D., and Alkuraya, F. S. (2016). Exome-based case-control association study using extreme phenotype design reveals novel candidates with protective effect in diabetic retinopathy. *Hum. Genet.*, 135(2):193–200.

[Siegel et al., 2013] Siegel, R., Naishadham, D., and Jemal, A. (2013). Cancer statistics, 2013. *CA Cancer J Clin*, 63(1):11–30.

[Sinville and Soper, 2007] Sinville, R. and Soper, S. A. (2007). High resolution DNA separations using microchip electrophoresis. *J Sep Sci*, 30(11):1714–1728.

[Smith and Haigh, 1974] Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical research*, 23(01):23–35.

[Smith et al., 2013] Smith, M. J., O'Sullivan, J., Bhaskar, S. S., Hadfield, K. D., Poke, G., Caird, J., Sharif, S., Eccles, D., Fitzpatrick, D., Rawluk, D., du Plessis, D., Newman, W. G., and Evans, D. G. (2013). Loss-of-function mutations in SMARCE1 cause an inherited disorder of multiple spinal meningiomas. *Nat. Genet.*, 45(3):295–298.

[So et al., 2011] So, H. C., Gui, A. H., Cherny, S. S., and Sham, P. C. (2011). Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases. *Genet. Epidemiol.*, 35(5):310–317.

[Sorensen and Gianola, 2007] Sorensen, D. and Gianola, D. (2007). *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Statistics for Biology and Health. Springer New York.

[Sottoriva et al., 2013] Sottoriva, A., Spiteri, I., Piccirillo, S. G., Touloumis, A., Collins, V. P., Marioni, J. C., Curtis, C., Watts, C., and Tavaré, S. (2013). Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proceedings of the National Academy of Sciences*, 110(10):4009–4014.

[Spiegelhalter et al., 2002] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

[Stadler et al., 2010] Stadler, Z. K., Gallagher, D. J., Thom, P., and Offit, K. (2010). Genome-wide association studies of cancer: principles and potential utility. *Oncology (Williston Park, N.Y.)*, 24(7):629–637.

[Stafford et al., 2017] Stafford, J. L., Dyson, G., Levin, N. K., Chaudhry, S., Rosati, R., Kalpage, H., Wernette, C., Petrucelli, N., Simon, M. S., and Tainsky, M. A. (2017). Reanalysis of BRCA1/2 negative high risk ovarian cancer patients reveals novel germline risk loci and insights into missing heritability. *PLoS ONE*, 12(6):e0178450.

[Stamatoyannopoulos et al., 2009] Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nat. Genet.*, 41(4):393–395.

[Stehelin et al., 1976] Stehelin, D., Varmus, H. E., Bishop, J. M., and Vogt, P. K. (1976). DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature*, 260(5547):170–173.

[Stein et al., 2012] Stein, E. A., Mellis, S., Yancopoulos, G. D., Stahl, N., Logan, D., Smith, W. B., Lisbon, E., Gutierrez, M., Webb, C., Wu, R., Du, Y., Kranz, T., Gasparino, E., and Swergold, G. D. (2012). Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N. Engl. J. Med.*, 366(12):1108–1118.

[Stratton et al., 2009] Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

[Su et al., 2011] Su, Z., Marchini, J., and Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics*, 27(16):2304–2305.

[Sukumar et al., 1983] Sukumar, S., Notario, V., Martin-Zanca, D., and Barbacid, M. (1983). Induction of mammary carcinomas in rats by nitroso-methylurea involves malignant activation of H-ras-1 locus by single point mutations. *Nature*, 306(5944):658–661.

[Sun et al., 2013] Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genet. Epidemiol.*, 37(4):334–344.

[Supek et al., 2011] Supek, F., BoÅ¡njak, M., Å kunca, N., and Å muc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLOS ONE*, 6(7):1–9.

[Supek and Lehner, 2015] Supek, F. and Lehner, B. (2015). Differential dna mismatch repair underlies mutation rate variation across the human genome. *Nature*.

[Supek et al., 2014] Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell*, 156(6):1324–35.

[Tabin et al., 1982] Tabin, C. J., Bradley, S. M., Bargmann, C. I., Weinberg, R. A., Papageorge, A. G., Scolnick, E. M., Dhar, R., Lowy, D. R., and Chang, E. H. (1982). Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149.

[Talbot and Crawford, 2004] Talbot, S. J. and Crawford, D. H. (2004). Viruses and tumours–an update. *Eur. J. Cancer*, 40(13):1998–2005.

[Tamborero et al., 2013a] Tamborero, D., Gonzalez-Perez, A., and Lopez-Bigas, N. (2013a). Oncodriveclust: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–44.

[Tamborero et al., 2013b] Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., and Lopez-Bigas, N. (2013b). Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep*, 3:2650.

[Tamborero et al., 2013c] Tamborero, D., Lopez-Bigas, N., and Gonzalez-Perez, A. (2013c). Oncodrive-cis: a method to reveal likely driver genes based on the impact of their copy number changes on expression. *PLoS One*, 8(2):e55489.

[Tan et al., 2017] Tan, P. L., Garrett, M. E., Willer, J. R., Campochiaro, P. A., Campochiaro, B., Zack, D. J., Ashley-Koch, A. E., and Katsanis, N. (2017). Systematic Functional Testing of Rare Variants: Contributions of CFI to Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.*, 58(3):1570–1576.

[Taparowsky et al., 1982] Taparowsky, E., Suard, Y., Fasano, O., Shimizu, K., Goldfarb, M., and Wigler, M. (1982). Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change. *Nature*, 300(5894):762–765.

[TCGANetwork, 2008] TCGANetwork (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8.

[TCGANetwork, 2011] TCGANetwork (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615.

[TCGANetwork, 2012a] TCGANetwork (2012a). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–25.

[TCGANetwork, 2012b] TCGANetwork (2012b). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–7.

[TCGANetwork, 2012c] TCGANetwork (2012c). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

[TCGANetwork, 2013] TCGANetwork (2013). Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med*, 368(22):2059–74.

[TCGANetwork, 2015] TCGANetwork (2015). Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature*, 517(7536):576–82.

[Tennessen et al., 2012] Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., Kang, H. M., Jordan, D., Leal, S. M., Gabriel, S., Rieder, M. J., Abecasis, G., Altshuler, D., Nickerson, D. A., Boerwinkle, E., Sunyaev, S., Bustamante, C. D., Bamshad, M. J., and Akey, J. M. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69.

[Thiel and Ristimaki, 2013] Thiel, A. and Ristimaki, A. (2013). Toward a Molecular Classification of Colorectal Cancer: The Role of BRAF. *Front Oncol*, 3:281.

[Tierney and Kadane, 1986] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

[Torre et al., 2016] Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). Global Cancer Incidence and Mortality Rates and Trends–An Update. *Cancer Epidemiol. Biomarkers Prev.*, 25(1):16–27.

[Trapnell et al., 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–5.

[Van Loo et al., 2010] Van Loo, P., Nordgard, S. H., Lingj?rde, O. C., Russnes, H. G., Rye, I. H., Sun, W., Weigman, V. J., Marynen, P., Zetterberg, A., Naume, B., Perou, C. M., B?rresen-Dale, A. L., and Kristensen, V. N. (2010). Allele-specific copy number analysis of tumors. *Proc. Natl. Acad. Sci. U.S.A.*, 107(39):16910–16915.

[Varghese and Easton, 2010] Varghese, J. S. and Easton, D. F. (2010). Genome-wide association studies in common cancers–what have we learnt? *Curr. Opin. Genet. Dev.*, 20(3):201–209.

[Vogelstein et al., 2013] Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., and Kinzler, K. W. (2013). Cancer genome landscapes. *science*, 339(6127):1546–1558.

[Vohra and Biggin, 2013] Vohra, S. and Biggin, P. C. (2013). Mutationmapper: a tool to aid the mapping of protein mutation data. *PLoS One*, 8(8):e71711.

[Wang et al., 2010] Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, 38(16):e164.

[Wang et al., 2011] Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., Werner, L., Sivachenko, A., DeLuca, D. S., and Zhang, L. (2011). Sf3b1 and other novel cancer genes in chronic lymphocytic leukemia. *New England Journal of Medicine*, 365(26):2497–2506.

[Wei et al., 2011] Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.*, 39(19):e132.

[Weinberg, 1995] Weinberg, R. A. (1995). The retinoblastoma protein and cell cycle control. *Cell*, 81(3):323–330.

[Weinstein et al., 2013] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., Wheeler, D., Ally, A., Balasundaram, M., Birol, I., Butterfield, S. N., Chu, A., Chuah, E., Chun, H. J., Dhalla, N., Guin, R., Hirst, M., Hirst, C., Holt, R. A., Jones, S. J., Lee, D., Li, H. I., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Robertson, A. G., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Varhol, R. J., Beroukhim, R., Bhatt, A. S., Brooks, A. N., Cherniack, A. D., Freeman, S. S., Gabriel, S. B., Helman, E., Jung, J., Meyerson, M., Ojesina, A. I., Pedamallu, C. S., Saksena, G., Schumacher, S. E., Tabak, B., Zack, T., Lander, E. S., Bristow, C. A., Hadjipanayis, A., Haseley, P., Kucherlapati, R., Lee, S., Lee, E., Luquette, L. J., Mahadeshwar, H. S., Pantazi, A., Parfenov, M., Park, P. J., Protopopov, A., Ren, X., Santoso, N., Seidman, J., Seth, S., Song, X., Tang, J., Xi, R., Xu, A. W., Yang, L., Zeng, D., Auman, J. T., Balu, S., Buda, E., Fan, C., Hoadley, K. A., Jones, C. D., Meng, S., Mieczkowski, P. A., Parker, J. S., Perou, C. M., Roach, J., Shi, Y., Silva, G. O., Tan, D., Veluvolu, U., Waring, S., Wilkerson, M. D., Wu, J., Zhao, W., Bodenheimer, T., Hayes, D. N., Hoyle, A. P., Jeffreys, S. R., Mose, L. E., Simons, J. V., Soloway, M. G., Baylin, S. B., Berman, B. P., Bootwalla, M. S., Danilova, L., Herman, J. G., Hinoue, T., Laird, P. W., Rhie, S. K., Shen, H., Triche, T., Weisenberger, D. J., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Sougnez, C., Wang, M., Saksena, G., Carter, S. L., Cibulskis, K., Chin, L., Zhang, J., Getz, G., Dinh, H., Doddapaneni, H. V., Gibbs, R., Gunaratne, P., Han, Y., Kalra, D., Kovar, C., Lewis, L., Morgan, M., Morton, D., Muzny, D., Reid, J., Xi, L., Cho, J., DiCara, D., Frazer, S., Gehlenborg, N., Heiman, D. I., Kim, J., Lawrence, M. S., Lin, P., Liu, Y., Noble, M. S., Stojanov, P., Voet, D., Zhang, H., Zou, L., Stewart, C., Bernard, B., Bressler, R., Eakin, A., Iype, L., Knijnenburg, T., Kramer, R., Kreisberg, R., Leinonen, K., Lin, J., Liu, Y., Miller, M., Reynolds, S. M., Rovira, H., Shmulevich, I., Thorsson, V., Yang, D., Zhang, W., Amin, S., Wu, C. J., Wu, C. C., Akbani, R., Aldape, K., Baggerly, K. A., Broom, B., Casasent, T. D., Cleland, J., Creighton, C., Dodda, D., Edgerton, M., Han, L., Herbrich, S. M., Ju, Z., Kim, H., Lerner, S., Li, J., Liang, H., Liu, W., Lorenzi, P. L., Lu, Y., Melott, J., Mills, G. B., Nguyen, L., Su, X., Verhaak, R., Wang, W., Weinstein, J. N., Wong, A., Yang, Y., Yao, J., Yao, R., Yoshihara, K., Yuan, Y., Yung, A. K., Zhang, N., Zheng, S., Ryan, M., Kane, D. W., Aksoy, B. A., Ciriello, G., Dresdner, G., Gao, J., Gross, B., Jacobsen, A., Kahles, A., Ladanyi, M., Lee, W., Lehmann, K. V., Miller, M. L., Ramirez, R., Ratsch, G., Reva, B., Sander, C., Schultz, N., Senbabaoglu, Y., Shen, R., Sinha, R., Sumer, S. O., Sun, Y., Taylor, B. S., Weinhold, N., Fei, S., Spellman, P., Benz, C., Carlin, D., Cline, M., Craft, B., Ellrott, K., Goldman, M., Haussler, D., Ma, S., Ng, S., Paull, E., Radenbaugh, A., Salama, S., Sokolov, A., Stuart, J. M., Swatloski, T., Uzunangelov, V., Waltman, P., Yau, C., Zhu, J., Hamilton, S. R., Getz, G., Sougnez, C., Abbott, S., Abbott, R., Dees, N. D., Delehaunty, K., Ding, L., Dooling, D. J., Eldred, J. M., Fronick, C. C., Fulton, R., Fulton, L. L., Kalicki-Veizer, J., Kanchi, K. L., Kandoth, C., Koboldt, D. C., Larson, D. E., Ley, T. J., Lin, L., Lu, C., Magrini, V. J., Mardis, E. R., McLellan, M. D., McMichael, J. F., Miller, C. A., O'Laughlin, M., Pohl, C., Schmidt, H., Smith, S. M., Walker, J., Wallis, J. W., Wendl, M. C., Wilson, R. K., Wylie, T., Zhang, Q., Burton, R., Jensen, M. A., Kahn, A., Pihl, T., Pot, D., Wan, Y., Levine, D. A., Black, A. D., Bowen, J., Frick, J., Gastier-Foster, J. M., Harper, H. A., Helsel, C., Leraas, K. M., Lichtenberg, T. M., McAllister, C., Ramirez, N. C., Sharpe, S., Wise, L., Zmuda, E., Chanock, S. J., Davidsen, T., Demchok, J. A., Eley, G., Felau, I., Ozenberger, B. A., Sheth, M., Sofia, H., Staudt, L., Tarnuzzer, R., Wang, Z., Yang, L., Zhang, J., Omberg, L., Margolin, A., Raphael, B. J., Vandin, F., Wu, H. T., Leiserson, M. D., Benz, S. C., Vaske, C. J., Noushmehr, H., Knijnenburg, T., Wolf, D., Van 't Veer, L., Collisson, E. A., Anastassiou, D., Ou Yang, T. H., Lopez-Bigas, N., Gonzalez-Perez, A., Tamborero, D., Xia, Z., Li, W., Cho, D. Y., Przytycka, T., Hamilton, M., McGuire, S., Nelander, S., Johansson, P., Jornsten, R., Kling, T., and Sanchez, J. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat.*

Genet., 45(10):1113–1120.

[Williams et al., 2016] Williams, M. J., Werner, B., Barnes, C. P., Graham, T. A., and Sottoriva, A. (2016). Identification of neutral tumor evolution across cancer types. *Nat. Genet.*, 48(3):238–244.

[Willis et al., 2004] Willis, A., Jung, E. J., Wakefield, T., and Chen, X. (2004). Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene*, 23(13):2330–2338.

[Wold, 1966] Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. *Research Papers in Statistics*, 630.

[Wooster et al., 1995] Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature*, 378(6559):789–792.

[Wu, 2012] Wu, C. J. (2012). Cll clonal heterogeneity: an ecology of competing subpopulations. *Blood*, 120(20):4117–8.

[Wu et al., 2011] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89(1):82–93.

[Xie et al., 2014] Xie, M., Lu, C., Wang, J., McLellan, M. D., Johnson, K. J., Wendl, M. C., McMichael, J. F., Schmidt, H. K., Yellapantula, V., Miller, C. A., Ozenberger, B. A., Welch, J. S., Link, D. C., Walter, M. J., Mardis, E. R., Dipersio, J. F., Chen, F., Wilson, R. K., Ley, T. J., and Ding, L. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med*, 20(12):1472–8.

[Yang et al., 2010] Yang, H., Zhong, Y., Peng, C., Chen, J. Q., and Tian, D. (2010). Important role of indels in somatic mutations of human cancer genes. *BMC Med. Genet.*, 11:128.

[Zack et al., 2013] Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhsng, C. Z., Wala, J., Mermel, C. H., Sougnez, C., Gabriel, S. B., Hernandez, B., Shen, H., Laird, P. W., Getz, G., Meyerson, M., and Beroukhim, R. (2013). Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45(10):1134–1140.

[Zhang et al., 2010] Zhang, Y., Gostissa, M., Hildebrand, D. G., Becker, M. S., Boboila, C., Chiarle, R., Lewis, S., and Alt, F. W. (2010). The role of mechanistic factors in promoting chromosomal translocations found in lymphoid and other cancers. *Adv. Immunol.*, 106:93–133.

[Zhao et al., 2014] Zhao, B., Hemann, M. T., and Lauffenburger, D. A. (2014). Intratumor heterogeneity alters most effective drugs in designed combinations. *Proc Natl Acad Sci U S A*, 111(29):10773–8.