Leveraging genomic and molecular variations to understand the regulatory landscape in human cancers and differentiating stem cells

Lara Hanne Urban

European Bioinformatics Institute (EMBL-EBI)

Wolfson College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy.

May 2019

Declaration of Originality

This dissertation is the result of my own work and includes nothing that is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

Abstract

Leveraging genomic and molecular variations to understand the regulatory landscape in human cancers and differentiating stem cells

Lara Hanne Urban

Genetic and molecular variations are closely intertwined; while genetic factors drive phenotypic differences ranging from gene expression to organismal traits, phenotypic variations are the target of evolutionary selection, what eventually results in genetic changes. As technological advances have resulted in high-throughput assays for different molecular dimensions, it has become challenging to turn these large-scale data into meaningful insights and to delineate biological cause and consequence. In this thesis, I use computational modelling to detect and understand biologically meaningful associations between genetic variation and gene expression alterations.

First, we use data across 27 human cancer types to probe associations between different genetic factors and gene expression levels. We describe the tumours' regulatory landscape that is highly heterogeneous across cancer types, and quantify the relationship between gene expression and various genetic features that characterise local and global mutational burden as well as distinct mutational processes.

Next, we study the relationship between genetic and epigenetic variation and alternative splicing. This analysis extends studies of splicing events in bulk data to variability in splicing between single cells from the same tissue: We analyse DNA methylation and alternative splicing across single cells derived from one human donor to characterise splicing variation and its determinants across genes. Thus, we identify relevant genetic determinants of splicing in induced pluripotent stem cells as well as during their differentiation, and a significant contribution of DNA methylation to splicing variation across cells.

Finally, we show how gene expression-mutagenesis screens can be applied to understand complex mutational signatures, using the cancer hallmark of DNA repair deficiency as an example. The molecular cause and consequence of homologous recombination repair deficiency are not yet fully understood. We explore genome-wide molecular aberrations caused by this repair deficiency beyond the few previously known genes. Our preliminary results point towards a genetically dominant effect of *BRCA1* mutagenesis.

Taken together, this thesis highlights novel dimensions of genotype-phenotype associations in highly heterogeneous molecular datasets. We describe the complex regulatory landscape across human cancer types, as well as molecular alterations and relevant epigenetic effects in differentiating pluripotent stem cells.

Acknowledgements

First, I would like to acknowledge Oliver Stegle and his research group who have provided an exciting and collaborative platform for me to learn about and carry out research for more than three years. Next, I would like to thank my thesis advisory committee, Jan Korbel, John Marioni and Willem Ouwehand, for valuable advice; Sebastian Waszak for supervision of my last project and for sharing his scientific insights and boundless creativity with me; Stephanie Linker and Marc-Jan Bonder for being the driving forces in the single-cell splicing project; the PCAWG consortium and specifically Claudia Calabrese, Alvis Brazma and the remaining PCAWG working group 3 for the collaborative efforts in the PCAWG project; and Dmitry Gordenin, Natalie Saini and Steven Roberts for their input on mutational signatures.

I thank the European Union's Horizon2020 research and innovation programme for funding of my PhD (grant agreement number N635290, 'PanCanRisk'); and EMBL Heidelberg, the Fuchs Fund, the Genetics Society UK, the OpenPlant initiative, University of Cambridge, the Burke's Peerage Foundation, University of Washington, and Wolfson College Cambridge for additional funding and support.

I specifically thank Wolfson College Cambridge that has provided me with the most intellectually curious and open scientific community I have ever experienced. Special thanks to communication officer Fiona Glisenan for her truly motivating interest and clever questions, to the Science Society for providing me with a platform to nourish my scientific interests, and to the Wolfson community and rowing club, for reminding me that Cambridge is about more than just obtaining your PhD.

I would like to acknowledge the PuntSeq team that has kept my scientific enthusiasm up in times of doubts. I have learned a lot about how essential excitement, collaboration and public engagement are for conducting good research.

Finally, despite not mentioning all individual names I would like to thank my family and friends who have supported me on many different levels. Special thanks go to Na Cai and Elo Madissoon, for being such good friends and the most amazing role models I could have wished for. Thanks to Alejandro de Miquel Bleier, for believing in everything I do while being the kindest person I have ever met, and for listening to genetics jabbering for hours.

Preface

The 2nd chapter involves the contribution of several PCAWG working group 3 members. I designed and realised the gene expression-mutational signature association study and all related downstream analyses. I additionally collaborated on the PCAWG analyses concerning allele-specific expression, the effect of germline genetic variants on gene expression and the effect of somatic genetic variants on gene expression. While the entire PCAWG consortium and particularly PCAWG working group 3 were involved in this project, crucial contributions to the presented work have been made by Claudia Calabrese, Kjong Lehmann, Fenglin Liu, Roland Schwarz and Nuno Fonseca. Oliver Stegle, Alvis Brazma, Gunnar Raetsch, Zemin Zhang and Angela Brooks guided and supervised the analyses.

In the 3rd chapter, my work focuses on the analysis of splicing variation across single cells and deep learning modelling. I realised all splicing variation studies, and assessed the splicing category switches during stem cell differentiation. All work has been done in close collaboration with Stephanie Linker, who pre-processed the majority of the data, and Marc-Jan Bonder, who pre-processed the methylation data and co-supervised the project. Davis McCarthy was involved in the gene expression data processing and the cell-level feature calculations. Mariya Chhatriwala, Stephen Clark, Shradha Amatya, Ludovic Vallier, and Wolf Reik generated the data.

The 4th chapter was a follow-up project of the mutational signatures analyses presented in chapter 2. I broadened the concept of signature-expression associations to more complex molecular signatures with a focus on DNA damage repair deficiency. This research has taken place under supervision of Sebastian Waszak and Oliver Stegle.

Contents

List of figures

Nomenclature

AEI	allelic expression imbalance
ASE	allele-specific expression
AUC	area under the receiver operating characteristic curve
β	effect size
BMA	Bayesian Model Averaging
bPSI	pseudo bulk alternative splicing rate
CADD	combined annotation dependent depletion
cDNA	complementary deoxyribonucleic acid
CN	copy-number
CNN	convolutional neural network
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA	deoxyribonucleic acid
eQTL	expression quantitative trait locus
ES	enrichment score
ES cell	embryonic stem cell
FDR	false discovery rate
FPKM	fragments per kilo base per million reads mapped
FWER	family-wise error rate
GTEx	genotype-tissue expression project
GWAS	genome-wide association study
HER2	human epidermal growth factor receptor type 2
HipSci	Human Induced Pluripotent Stem Cell Initiative
HR	homologous recombination repair
HRD	homologous recombination repair deficiency
ICGC	International Consortium of Cancer Genomes
indel	short insertion or deletion
iPS cell	induced pluripotent stem cell
KEGG	Kyoto Encyclopedia of Genes and Genomes
LD	linkage disequilibrium
LLR	log-likelihood ratio
MAF	minor allele frequency
MLE	maximum likelihood estimator
MMEJ	microhomology-mediated end joining
MMR	mismatch repair
mRNA	messenger ribonucleic acid

NGS	next-generation sequencing
NHEJ	non-homologous end joining
NMF	Nonnegative matrix factorization
PARP	Poly(ADP-Ribose)-Polymerase
PARPi	Poly(ADP-Ribose)-Polymerase inhibition
PC(A)	principal components; principal component analysis
PCAWG	Pan-Cancer Analysis of Whole Genomes
PCR	polymerase chain reaction
PSI	alternative splicing rate
qPCR	quantitative polymerase chain reaction
QQ plot	quantile-quantile plot
QTL	quantitative trait locus
ROC	receiver operating characteristic
ROS	reactive oxygen species
RNA	ribonucleic acid
RNA-Seq	ribonucleic acid sequencing
rRNA	ribosomal ribonucleic acid
RT	reverse transcription
SAGE	serial analysis of gene expression
scM&T-Seq	single-cell methylation and transcriptome sequencing
SNP	single-nucleotide polymorphism
SNV	single-nucleotide variant
SCNA	somatic copy-number alteration
SV	structural variant
TCGA	The Cancer Genome Atlas
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer ribonucleic acid
TSS	transcription start site
UTR	untranslated region
UV	ultraviolet
VAF	variant allele frequency
VEP	variant effect predictor
WGS	whole-genome sequencing

1 Introduction

Genetic and molecular variation are strongly dependent on each other: Genetic factors drive phenotypic differences ranging from gene expression to organismal traits while phenotypic variations are under evolutionary selection what eventually results in genetic changes. In this introduction, we first review sources of genetic variation in the human genome, followed by different forms of molecular variation that can now be quantified in unprecedented numbers due to technological advances and the availability of high-throughput assays. Specifically, we describe in detail how total gene expression and alternative splicing can be quantified and serve as molecular proxies for protein expression (Section 1.1).

The prevailing challenge in statistical genomics is the conversion of large-scale heterogeneous data into meaningful insights, including the predictability of molecular alterations and the associated delineation of biological cause and consequence. We focus on gene expression as molecular readout, and explore how the link between genetic and gene expression variation has been modelled so far, and how different sources of data have been integrated in previous analyses (Section 1.2): We introduce genome-wide association studies (GWAS) and quantitative trait locus (QTL) analyses, and explain how their statistical fundament, linear regression models, can be used to correct for known and hidden determinants of gene expression variability like batch effects, experimental confounders or population structure. We also give an overview of how complex machine learning methods can be employed to analyse non-linear associations in genomic data, using convolutional neural networks (CNNs) as exemplar.

We then introduce human cancer as the central biological study system of this thesis (Section 1.3). While the regulatory landscape of human cancers is highly heterogeneous, the relevance of the transcriptome for diagnosis and prognosis of the disease, and for the activity of mutational processes on the molecular level, has been well established.

1.1 Functional consequences of genetic variation

1.1.1 DNA and genetic variation

The genome of an organism contains all necessary information to encode a living organism, and stores information about its evolutionary past. The genome of all living organisms and some viruses consists of DNA (deoxyribonucleic acid), a molecule composed of two antiparallel strains of polynucleotides kept together by hydrogen bonds. Each polynucleotide is composed of nucleotides, with each of them containing one of the four nitrogenous bases adenine ('A'), cytosine ('C'), guanine ('G'), and thymine ('T'). On a chemical level A and G constitute purines, C

and G pyrimidines, aromatic heterocyclic organic compounds that consist of one ring in the case of pyrimidines and of two rings in the case of purines. The hydrogen bonds that keep the two polynucleotide strands together operate between one purine and one pyrimidine base: A binds to C, and G to T. This makes the sequences of the two DNA strands complementary to each other (Alberts et al., 2009).

The length and organisation of DNA differ between evolutionary taxa. Here, we focus on the nuclear DNA of eukaryotic organisms, which is organised into distinct chromosomes. Humans and most other mammals are diploid organisms, *i.e.*, they have two homologous copies of each chromosome, one inherited from the mother and one from the father. Specifically, the human genome consists of 23 pairs of chromosomes, namely 22 pairs of autosomal chromosomes and one pair of allosomes. Whereas an allosome describes any chromosome that differs from the typical autosomes in, *e.g.*, size or behaviour, the human allosomes constitute sex chromosomes that determine the genetic sex of human individuals. Human females have two copies of the X allosome, whereas males have one copy of the X and one of the Y allosome (Alberts et al., 2009). Each chromosome contains protein-coding genes (approximately 1.5% of the whole genome) and inter-genic non-coding regions. Current estimates predict the presence of 19,000 to 20,000 protein-coding genes in the human genome (Ezkurdia et al., 2014), with alternative splicing further diversifying possible function of each individual gene (Black, 2003; Section 1.1.2). While these protein-coding regions are transcribed into messenger RNA (mRNA), the human genome also contains genes that are transcribed into non-coding RNA, which can be functional on its own (*e.g.*, tRNA (transfer RNA) and rRNA (ribosomal RNA), which are key players of the translation process) (Palazzo & Lee, 2015). The human genome is approximately 3.24G base pairs long (Venter et al., 2001; Lander et al., 2001). Every pair of human individuals share on average 99.5% of these base pairs (Levy et al., 2007).

This leads us to genetic differences within a species: DNA sequence alterations that exist within an individual or between individuals within a population or between populations are called genetic variants. The alternative variants that occur at a specific genetic locus are called alleles. The genotype of an individual describes this individual's specific combination of alleles at a (normally diploid) genetic locus. In the case of diploid organisms, the genotype at a locus can hence be homozygous (*i.e.*, with two copies of the same allele) or heterozygous (*i.e.*, with two different alleles).

The pool of genetic variation is large. Genetic variants can be categorised according to their length, their origin, their allele frequency, their functional effect, the type of chromosome they lie on, and if they are located in coding or non-coding stretches of the genome.

The most common form of a genetic variant is the single-nucleotide variant (SNV), which describes the substitution of a single base pair (Frazer et al., 2009). SNVs can be of somatic origin, *i.e.*, they originate from a *de novo* genetic mutation in somatic tissue, or of germline origin, *i.e.*, they are inherited from parents to the offspring. The latter ones are referred to as single-nucleotide polymorphisms (SNPs). Variation in SNPs across individuals and populations

enables the inference of genetic population structure (Novembre et al., 2008). Most germline genetic variants in the human genome are assumed to be evolutionary neutral (Kimura, 1968), hypothesising that population differences have mainly arisen due to chance and stochasticity. Linkage disequilibrium (LD) describes the non-random association of SNPs at different genetic loci across individuals, and can serve as a readout for genetic processes that affect a population. Important evolutionary factors shape the LD of a genetic variant, including selection, recombination, mutation rate, and genetic drift (Slatkin, 2008). As the concept of LD relies on the assumption that those SNPs that are located close to each other tend to be inherited together (a phenomenon also known as 'gene linkage'), LD is affected by the distance between SNPs. In addition, LD patterns across the genome are influenced by the recombination rate: Genetic recombination during meiosis in eukaryotes can shuffle genetic material between homologous chromosomes and therefore create new genetic combinations in the offspring. Recombination enables alignment of chromosomes during meiosis to ensure proper segregation of chromosomes and avoid deleterious genetic disjunctions resulting in, *e.g.*, aneuploidy (Hassold et al., 2007). Over several generations, recombination generates genetic diversity in populations due to new combinations of alleles (Slatkin, 2008).

SNPs can be categorised according to their frequency in a population. If a genetic variant is bi-allelic, its population-level frequency is described by its minor allele frequency (MAF). This allows categorisation of genetic variants based on predefined thresholds, *e.g.*, into rare (MAF < 0.05) and common (MAF ≥ 0.05) variants (Frazer et al., 2009). The minimum MAF of genetic variants that can be accurately assessed within a particular study cohort therefore strongly depends on the sample size of genetic cohorts, which is expected to continue to increase for many years (Roundtable on Translating Genomic-Based Research for Health, 2016).

Whereas most SNPs are bi-allelic, larger genetic variants, so-called structural variants (SVs), often show more than two alleles in a population. SVs include short insertions and deletions (indels), larger copy-number (CN) changes due to insertions, deletions, and duplications, as well as inversions that range in size from kilo base pairs to whole chromosome arms (Frazer et al., 2009). The functional relevance of somatic copy-number alterations (SCNAs) has hereby attracted more and more attention, particularly in cancer research where large genetic rearrangements have been shown to contribute to tumour development and progression (Wu et al., 2014).

Genetic variants also differ in terms of their relative genomic position, *i.e.*, they can lie in coding or non-coding parts of the genome. The functional consequences of protein-coding genetic variants can be inferred by determining if the variant affects the amino acid sequence of a gene's encoded protein or not. Non-synonymous variants alter the amino acid sequence of the downstream protein; synonymous variants don't alter the amino acid sequence due to the degeneracy of the genetic code. Non-synonymous variants can hereby cause nonsense mutations due to premature stop codons that result in truncated proteins, or missense mutations that lead to the incorporation of a different amino acid into the protein. Additionally, it has

become clear that many genetic variants in non-coding areas of the genome are functionally relevant, *e.g.*, by modulating transcription factor (TF) binding efficiency, altering the DNA's spatial structure, affecting alternative splicing rates, or influencing the degradation rate of the mRNA. The functional classification of non-coding genetic variants is more difficult than the categorisation of coding variants since they might result in much more complex phenotypes than the alteration of a protein's amino acid composition. To make sense of all genome-wide genetic variation, many of the functional consequences of these non-coding variants for molecular variation and ultimately for the organismal phenotype have yet to be understood (Alberts et al., 2009).

The next section describes the extent of molecular variation that can be assessed by existing technology and that can be used to understand the functional consequences of both, coding and non-coding genetic variants.

1.1.2 Phenotypic effects of genetic variation

The very first effect of different genetic alleles on a phenotype was described by Gregor Johann Mendel when he studied the discrete inheritance of dichotomous traits in plants (Mendel, 1866). This was extended to quantitative phenotypes by Francis Galton (1871) who observed that, opposed to dichotomous traits in plants, height was inherited in an averaged manner in humans. Traits such as height that do not follow discrete Mendelian inheritance are referred to as complex traits. By today, we understand that most organismal traits have a genetically complex architecture and follow heritability patterns as observed by Galton. The biostatistician and (eu)geneticist Ronald A. Fisher used mathematical approaches to combine Mendelian genetics with Darwinist natural selection hypotheses to explain Galton's observation by an additive genetic model (Fisher, 1918). The additive genetic model assumes small effects of a large number of genetic loci that ultimately result in the normal distribution of the affected quantitative trait (Charlesworth & Charlesworth, 2010).

Levels of phenotypic variation

While Mendel, Galton and Fisher relied on organismal traits for their studies, high-throughput assays now allow assaying a much broader range of phenotypes, and in particular molecular phenotypes. Possible molecular phenotypes range from direct DNA modifications to observable traits on the organismal level (**Figure 1**). Direct DNA modifications can concern the three-dimensional structure of DNA and hence the chromatin accessibility, with regulatory impact on gene expression (Degner et al., 2012). Further, DNA can be affected by epigenetic modifications, heritable chemical or physical changes of the DNA that do not affect the actual DNA sequence. These include modifications of the DNA itself (most prominently DNA methylation, *i.e.*, the covalent addition of a methyl group to cytosine) or of the histone proteins

that the DNA is wrapped around, resulting in structural changes of the chromatin. DNA methylation in promoter regions of genes can act to repress gene expression and therefore have a strong functional impact. The first expression level of genomic information is the transcriptome level. The number of transcribed RNA (ribonucleic acid) sequences per gene can hereby serve as a proxy of the gene's activity. Most genes can exert different functions via the mechanism of alternative splicing: Alternative splicing ultimately leads to different transcripts of a single gene, another level of gene expression complexity (see below). While some RNA is functional on its own (*e.g.*, tRNA, rRNA; Section 1.1.1), most RNA (mRNA; Section 1.1.1) encodes proteins by translation of a sequence of nitrogenous bases into a sequence of amino acids. Further, the function of a protein itself can be affected by sequence or structural alterations, by modulation of its quantity, or by changes of its interaction partners. Notably, the protein group of TFs binds to regulatory genetic regions, affects the transcription levels of other genes and hence their proteomic output. Proteins are involved in a vast range of functions in living organisms, including cellular structure and transport, metabolism, response to stimuli and DNA replication. Their quantity and function translates into organismal phenotypes, including physiological traits like height and hair colour or susceptibility to disease (**Figure 1**; Alberts et al., 2009).

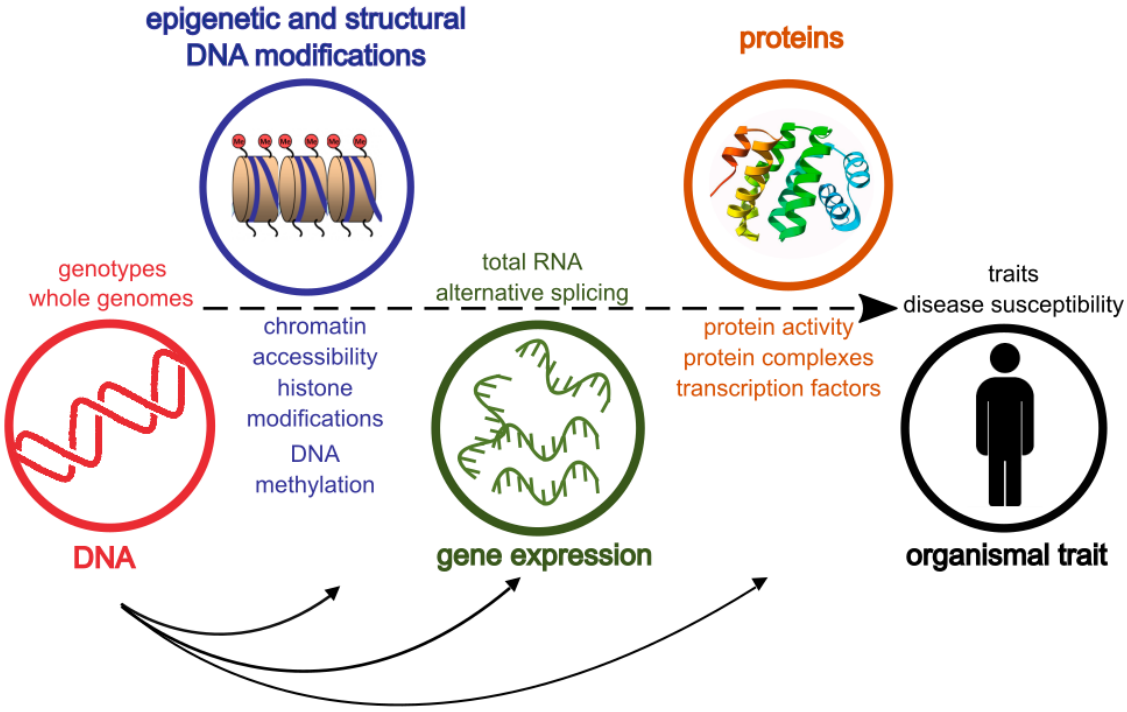

Figure 1. Different levels of molecular variation that can be modulated by genetic variation. While alterations of the epigenome and the DNA structure, gene expression, protein expression, and organismal traits can be directly caused by underlying genetic variation (solid arrows), indirect effects additionally complicate the regulatory landscape (dashed arrow).

Gene expression

In this thesis, we focus on gene expression, *i.e*., the transcriptome, as a molecular phenotype. In general, the transcriptome describes the complete set of transcripts in a tissue or cellular sample, and their respective quantity. As a precursor of protein expression, mRNA can serve as a proxy of gene expression levels. Multiple approaches have been developed to measure cellular mRNA levels, including hybridization- and sequencing-based approaches. In the case of hybridization-based methodology, reverse transcription (RT) is used to generate a complementary DNA (cDNA) template of the mRNA. When this cDNA template is being amplified with labelled hybridization probes via quantitative polymerase chain reaction (qPCR), fluorescence is emitted according to the oligonucleotides that are being incorporated. Based on the fluorescence signal, the genetic sequence of the original mRNA strand can be reconstructed. Alternatively, a hybridization microarray contains pre-defined probes for transcripts of every known gene of one or several species. Transcripts that are not known *a priori* can be detected with tag-based methods such as SAGE (Serial Analysis of Gene Expression); SAGE uses small tags that cover only fragments of a transcript as probes, and can therefore, opposed to hybridization microarray chips, also discover transcripts whose full sequence is unknown. However, a large proportion of the tags used by SAGE does not map to unique regions of a reference genome due to their short length, and can therefore not be used for transcript quantification. Further, tag-based approaches do not ensure the analysis of the entire transcriptome, and can generally not discover alternative splicing events (Wang et al., 2009).

RNA sequencing (RNA-Seq) based on next-generation sequencing (NGS) of the cDNA allows for genome-wide quantification of the transcriptome. After obtaining one (single-read RNA-Seq) or two paired (paired-end RNA-Seq) sequence reads per cDNA fragment, the sequencing reads are either aligned to a reference genome or are assembled *de novo*. From the number of RNA-Seq reads that map to a particular gene an estimation of gene expression can be deduced. In this thesis, we use FPKM (fragments per kilo base per million reads mapped) for gene expression estimations. FPKM quantify the number of reads that are assigned to a given gene, normalised by gene length and the total sequencing depth (Wang et al., 2009).

Opposed to the hybridization- and tag-based methods, NGS allows for identifying completely new genes, previously unknown genetic variants in the genes, variation in alternative splicing (see below), or post-transcriptional modifications. In addition, RNA-Seq can quantify the vast array of non-coding RNA molecules (Section 1.1.1). Altogether, sequencing-based assessments of the transcriptome deliver more detailed insights into gene expression variability than hybridization- or tag-based approaches.

In this thesis, we quantify gene expression by RNA-Seq measurements of mRNA levels. Besides analysing total gene expression and its alterations across cell types and individuals, we also investigate alternative splicing variation. Alternative splicing is a molecular process that

only takes place in eukaryotic cells. Eukaryotic genes are discontinuous, and consist of protein-coding exons and non-coding introns (Sharp, 1994). Exons are normally shorter than introns: In humans, 80% of exons are shorter than 200 base pairs whereas the mean intron length in is 2600 base pairs (Sakharkar et al., 2005). The human genome project revealed that human DNA consists of surprisingly few exons (1.1% of the genome), whereas introns cover 24% of the genome (Venter et al., 2001; Lander et al., 2001). The number of genes was also found to be smaller than expected, with around 30,000 being identified in 2001, and 19,000 genes being the latest estimate at the time that this thesis is written (Ezkurdia et al., 2014). However, this relatively small number of genes can exert various functions in a large number of complex processes, since alternative splicing diversifies the functionality of each gene by including or excluding exons in the transcript (Black, 2003). Importantly, the amount of alternatively spliced genes has been shown to be correlated with phenotypic complexity in higher eukaryotes (Kim et al., 2007): Alternative splicing occurs in more than 95% of human genes, compared with 63% of mouse genes (Barbosa-Morais et al., 2012). While alternative splicing plays an important role in many regulatory functions like cell differentiation and diversification of neuronal wiring (Kelemen et al., 2013), it has also been found to affect the development of various diseases like autism spectrum disorder, spinal muscular atrophy and different cancers (Xiong et al., 2015).

Splicing is carried out by the spliceosome, a ribonucleoprotein machinery. It consists of five small ribonucleoprotein complexes (U1, U2, U4/U6, and U5) (Wahl et al., 2009): U1 recognizes and binds to the 5' end of the intron, whereas U2 binds to a distinct sequence within the intron. Binding of these complexes ensures recognition of the introns and of the ends of the retained exons (Matlin et al., 2005). Subsequently, the complexes U4 and U5 bind to the genomic region, U6 replaces U1, U4 is removed, and finally a transesterification reaction removes the introns and connects the adjacent exons (Fica et al., 2013).

While different types of alternative splicing have been observed (Sammeth et al., 2008), the so-called exon skipping within a cassette exon is the most frequently observed type. A cassette exon describes an alternative exon that is flanked by two constitutive exons separated by introns. The alternative exon can either be spliced out or retained in the transcript (Black, 2003). In this thesis, we will only focus on alternative splicing events at cassette exons. Alternative splicing at cassette exons can be quantified using RNA-Seq data by calculating the ratio of the number of transcript reads that include the alternative exon to the total number of transcript reads that map to the cassette exon (Black, 2003; Section 3.2.1).

Genetic regulation of molecular traits

While molecular phenotypes are interesting to study in their own right, modern genetics research, frequently also termed 'systems genetics', considers these traits as mediating factors between genetic variation and variability in observable traits (**Figure 1**). The rationale behind this is that intermediate phenotypes reflect the immediate consequences of genetic variation, which can reduce noise and allows for mechanistic insights into deregulated biological pathways

(*e.g.*, Chen et al., 2016). Molecular phenotypes such as gene expression are hereby highly cell type-specific. Relevant associations between genetic variants and gene expression may therefore be missed if the most relevant tissue is not profiled (see Fairfax et al. (2012) who demonstrated substantial differences in genetic variant-gene expression associations between monocytes and B lymphocytes). In addition to cell type specificity, genetic variant-gene expression associations have also been shown to depend on the context of the tissue in question; *e.g.*, Fairfax et al. (2014) discovered certain associations in monocytes only when the cells were induced with immune stimuli. These cell type- and context-specificities have provided a first insight into the complexity of associations between genetic and molecular variation. Therefore, defining molecular phenotypes as mediating factors of organismal phenotypes remains a difficult task. Also, not all associations between genetic variants and gene expression result in variation of an observable trait, and for many of the organismal traits no molecular mediators have yet been identified. However, some studies revealed individual cases of striking colocalisation between molecular and disease phenotypes, *e.g.*, between inflammatory bowel disease and cell type-specific *CARD9* expression, a gene that has been shown to have a regulatory role in cell apoptosis processes (Chen et al., 2016).

Genetic variation can be associated with all levels of molecular variation, ranging from direct DNA modifications to organismal traits (**Figure 1**). These effects can be direct, *e.g.*, by changing the sequential composition of RNA or proteins, or indirect, *e.g.*, by altering regulatory genetic regions and therefore TF binding efficiency, or by modulating epigenetic modification, DNA structure and gene accessibility (**Figure 1**). The effect of a genetic variant can further be classified as *cis, i.e.*, the genetic variant in question regulates the molecular phenotype of a neighbouring gene, or as *trans*, *i.e.*, the genetic variant in question regulates the molecular phenotype of a different gene via intramolecular interaction, *e.g.*, via affecting proteins like TFs. The latter mode of gene regulation often implies that the genetic variant is distant from the regulated gene and frequently located on a different chromosome (Yao et al., 2017). Epistasis hereby describes the interaction between genes in regulating a phenotype: One genetic locus can alter or even mask the effect of other genetic loci, complicating or hampering the study of regulatory effects. Exploiting the phenomenon of epistasis may, however, help identify previously unknown genetic variants and their regulatory effects (Cordell, 2002).

Germline variants are known to be causal for gene expression changes. While the cause-effect relationship between these genetic variants and molecular phenotypes is therefore known, it is often difficult to identify the actually causal individual variant due to the LD structure of the genome (Section 1.1.1). In the case of somatic variants, even the cause-effect relationship between the genetic variant and molecular variation is often intractable: While somatic variants may affect gene expression, molecular mechanisms can also introduce *de novo* somatic mutations to the DNA, *e.g.*, due to erroneous DNA replication or insufficient DNA repair.

In the next section, we give an overview of the statistical models that this thesis applies to tackle the challenges presented by the analysis of genetic molecular datasets.

1.2 Integrative modelling of genomic and molecular variation

1.2.1 Expression quantitative trait locus analyses

GWAS link genetic variation to variability in organismal traits (Frazer et al., 2009); these studies perform under the assumption that a genetic variant that has been found to be statistically significantly associated with a trait is in strong LD with the actually causal variant. GWAS therefore leverage the LD structure of the genome to detect genetic loci responsible for a specific trait. QTL studies can associate genetic variants with tissue-specific molecular alterations across individuals; we here focus on expression QTL (eQTL) studies that associate genetic variants with tissue-specific gene expression (Schadt et al., 2008). QTL studies have, however, recently been extended to assess the effect of genetic variants on other molecular phenotypes, including DNA methylation (Gaunt et al., 2016), histone modifications (Grubert et al., 2015), chromatin architecture (Waszak et al., 2015), and protein expression (Stark et al., 2014) (see **Figure 1** for an overview of molecular traits that can be leveraged in association studies).

A persistent challenge in eQTL studies is the strong dependence of associations between genetic effects and gene expression on the cell or tissue type and the context of the study (Section 1.1.2). Therefore, eQTL studies have to be realised in various cellular and environmental contexts to fully understand the associations between genetic and gene expression variation. Other environmental variables, such as age or sex of the individuals, might also have an impact on these associations. Hence, to understand true biological associations, these variables have to be accounted for in eQTL studies.

The GTEx (Genotype-Tissue Expression) project started tackling these problems by obtaining gene expression profiles from a broad range of tissues of post-mortem human donors, and by collecting environmental variables of the same individuals (Lonsdale et al., 2013; The GTEx Consortium et al., 2015). In addition, projects like the Human Cell Atlas will result in an even more detailed classification of cell types by defining molecular cell types based on single-cell transcriptomic data (Regev et al., 2017) what will help identify truly cell type-specific associations.

However, association testing between many molecular measurements and many genetic variants in multiple tissues and contexts entails an even larger multiple testing burden than in conventional GWAS. For this reason, eQTL studies have mostly been restricted to *cis* eQTL studies, mapping only proximal genetic variants to the expression of the respective gene since they are more likely to affect the molecular trait. Increasing the sample size of gene expression studies has already helped boost the statistical power of association studies (The GTEx

Consortium et al., 2015), but ideally samples from multiples studies could be pooled and analysed by one approach to alleviate the sample size problem.

Data heterogeneity, particularly of large-scale or pooled studies, leads us to the problem of confounding factors across individuals that are often not well defined or unknown, leading to spurious association between genetic variants and gene expression and therefore reducing the statistical power to identify real associations (McClellan & King, 2010). Many approaches for solving this challenge in eQTL studies have been proposed; here, we will focus on linear (mixed) models, which have emerged as a robust method to control for confounding factors (Kang et al., 2008) and which have been implemented in a computationally efficient way (*e.g.*, Lippert et al., 2014).

1.2.2 Linear regression models in genomics

Linear regression models a continuous variable y_i as linearly related to C input variables, referred to as covariates, across N samples with $i \in \{1, 2, ..., N\}$:

$$y_i = \sum_{c=1}^{C} x_{ic}\,\beta_c + \psi_i \text{ , with } \psi_i \sim \mathrm{N}(0, \sigma_e^2) \qquad (i)$$

The residual term ψ_i accounts for deviations from the model due to noise or covariates that are not taken into account. In a simple linear model, these residual terms are assumed to be independent and identically distributed with variance σ_e^2. The weights β_c describe the magnitude and directionality of the effects of the model covariates x_{ic} on the phenotype y_i. These so-called effect sizes can be estimated by a maximum likelihood approach (see below). We here denote matrices (including vectors) - as opposed to scalar values - by bold scripture or capital letters. By introducing Y, X, and $\boldsymbol{\beta}$ ($Y \in \mathbb{R}^{\mathrm{N\,x\,1}}$; $X \in \mathbb{R}^{\mathrm{N\,x\,C}}$; $\boldsymbol{\beta} \in \mathbb{R}^{\mathrm{C\,x\,1}}$), and the identity matrix $I \in \mathbb{R}^{\mathrm{N\,x\,N}}$, we can specify the probability distribution p of the data as follows:

$$p\,(Y|X, \boldsymbol{\beta}, \sigma_e^2) \;=\; \mathrm{N}(Y|X\boldsymbol{\beta}, \sigma_e^2 I) \qquad (ii)$$

This probability p from equation (*ii*) is also known as the likelihood of the model $\mathcal{L}(\boldsymbol{\beta}, \sigma_e^2)$. The maximum likelihood estimator (MLE) is then the set of parameters $\boldsymbol{\beta}$ and σ_e^2 that maximises the likelihood (or typically the log likelihood, which is easier to estimate due to the mathematical properties of the logarithm). The MLE of these parameters, here denoted as $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2_e$, can be defined by the following equation (*iii*):

$$\widehat{\boldsymbol{\beta}}, \hat{\sigma}_e^2 = argmax_{\boldsymbol{\beta}, \sigma_e^2} \mathcal{L}(\boldsymbol{\beta}, \sigma_e^2) \quad (iii)$$

The MLE values can be obtained by calculating the first derivative of the (log) likelihood with respect to the two parameters, and by then obtaining the set of parameters that will set this derivative to zero. It can then be proven that (*v*) solves (*iv*) and (*vii*) solves (*vi*), resulting in the MLE values $\widehat{\boldsymbol{\beta}}$ and $\hat{\sigma}_e^2$:

$$\frac{\delta \, log\mathcal{L}(\boldsymbol{\beta}, \sigma_e^2)}{\delta \, \boldsymbol{\beta}}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \, \sigma_e^2=\widehat{\sigma_e^2}} = 0 \quad (iv)$$

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y \quad (v)$$

$$\frac{\delta \, log\mathcal{L}(\boldsymbol{\beta}, \sigma_e^2)}{\delta \, \sigma_e^2}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \, \sigma_e^2=\widehat{\sigma_e^2}} = 0 \quad (vi)$$

$$\widehat{\sigma_e^2} = \frac{1}{N} (y - X(X^T X)^{-1} X^T y)^T (y - X(X^T X)^{-1} X^T y) \quad (vii)$$

Here, we model the genetic variant and all other known ('fixed') factors as covariates in $X \in \mathbb{R}^{\mathrm{N \, x \, C}}$. In genotype-phenotype association studies, we are, however, only interested in the effect of the genetic variant, from now on referred to as β:

$$Y = G\beta + X\boldsymbol{\alpha} + \psi, \qquad \text{with } \psi \sim \mathrm{N}(0, \sigma_e^2 I) \quad (viii)$$

Here, we focus on one phenotype *Y* that can, *e.g.*, describe the expression of one gene. The matrix $G \in \mathbb{R}^{\mathrm{N \, x \, 1}}$ contains the genotype of one genetic variant across all samples. Then, $\beta \in \mathbb{R}$ denotes the effect of the one genetic variant, and $\boldsymbol{\alpha} \in \mathbb{R}^{(\mathrm{C}-1) \, \mathrm{x} \, 1}$ denotes the effects of the remaining fixed covariates in $X \in \mathbb{R}^{\mathrm{N \, x \, (C-1)}}$.

Hereby, the genotype has to be encoded according the assumed genetic model. Here, we assume an additive genetic model that proposes a genetic effect proportional to the minor (or major) allele count η (η $\in$ {0, 1, 2}). However, some genetic effects might be modelled more accurately via the dominant or recessive genetic models that assume that either one copy of an

allele results in a phenotypic effect (dominant model), or that two alleles must be present to result in a phenotypic effect (recessive model).

The concept of the linear model can be extended to the one of a linear mixed model that includes an additional covariate type, the random covariate. This covariate describes a random effect that is the realisation of a random variable of which we only model the distribution. A random effect is hence non-deterministic as opposed to fixed effects such as the ones represented by $\boldsymbol{\alpha}$, allowing to encode intractably large numbers of fixed effects as one random effect. Random covariates have been shown to capture subtle intrinsic data structure better than deterministic fixed covariates, and to yield calibrated association studies even in the case of complex data structures (*e.g.*, Yu et al., 2006). In genotype-phenotype association studies across individuals, the population structure can be estimated based on a large number of genome-wide SNPs that have been shown to act as a reliable estimator of population structure (Kang et al., 2008). As accounting for a large number of SNPs as independent fixed effects would be intractable, we can model population structure $\boldsymbol{u}$ as a random covariate next to the fixed covariates:

$$Y = G\beta + X\boldsymbol{\alpha} + \boldsymbol{u} + \psi, \qquad \text{with } \boldsymbol{u} \sim \mathrm{N}(0, \sigma_g^2 \boldsymbol{R}), \psi \sim \mathrm{N}(0, \sigma_e^2 I) \quad (ix)$$

Here, the genetic relatedness matrix $\boldsymbol{R}$ quantifies the pair-wise genetic similarity between all individuals based on genome-wide SNPs. For example, the realised relatedness matrix $\boldsymbol{R}^*$ can be calculated based on standardised genotypes of the matrix $G \in \mathbb{R}^{\mathrm{N \times S}}$ with N individuals and S SNPs:

$$\boldsymbol{R}^* = \frac{1}{S} G G^T \qquad (x)$$

The effects in a linear (mixed) model can then be estimated by maximising the likelihood of the model. For simplicity, we firstly introduce the maximum likelihood approach applied to a linear model. In this approach, the two following hypotheses concerning the effect of a genetic variant can be tested by comparing the likelihood of their underlying models: The null hypothesis H_0 assumes no effect of the genetic variant on the phenotype (β=0), and the alternative hypothesis H_1 assumes the significant presence of an effect ($\beta \neq 0$). We can calculate the test statistic as a random variable that quantifies the evidence that H_0 can be rejected. A typically employed test statistic is the log likelihood ratio (LLR), which directly compares the likelihood of both models. If $\{\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_e^2\}$ are the MLE of the parameters of H_1, and $\{\check{\boldsymbol{\alpha}}, \check{\sigma}_e^2\}$ the MLE of the parameters of H_0, then the LLR can be calculated as follows:

$$LLR = log\mathcal{L}(\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_e^2) - log\mathcal{L}(0, \check{\boldsymbol{\alpha}}, \check{\sigma}_e^2) \quad (xi)$$

The P-value of the association between genetic variant and phenotype is then the probability that a test statistic sampled under the assumption of H_0 is greater than or equal to the observed test statistic. According to a theorem by Wilks (1938), the test statistic 2xLLR asymptotes to a χ^2 distribution with - in the case of one tested genetic variant - one degree of freedom. The P-value of the LLR can therefore be calculated via the cumulative density function of χ^2, here denoted as *F*:

$$P(LLR) = 1 - F(2LLR;\ 1) \quad (xii)$$

If the P-value is below a pre-defined significance threshold, H_0 is rejected and a statistically significant association between genetic variant and phenotype is postulated. The significance threshold hereby models the expected percentage of false-positive associations, *i.e.*, associations that reject H_0 when it is true (type 1 error). Besides the false-positive associations, also false-negative associations, *i.e.*, associations that reject H_1 when it is true (type 2 error), might occur in association studies but are not controlled for by a manually set threshold. Instead, the statistical power of an association study assesses the rate of true-positive associations, *i.e.*, the rate of accepting H_1 when it is true.

When conducting association tests across multiple genetic variants, the expected number of false-positive associations scales linearly with the number of tested variants. This problem is widely known as multiple hypotheses testing burden. Different methods to adjust nominal P-values have been developed in order to ensure that the type 1 error rates remains controlled in this setting. The most straightforward approach to adjust P-values controls the family-wise error rate (FWER), *i.e.*, the probability of having at least one false positive across all tests. One specific method that controls the FWER, the Bonferroni method, multiplies the P-values by the number of tests. This method is based on the assumption of independence of the conducted tests, an assumption that - especially in the case of complicated dependencies due to genetic structure in genotype-phenotype association studies - might lead to an overly conservative correction of P-values (Goeman & Solari, 2014). Another approach to adjust P-values for multiple testing is based on controlling the false discovery rate (FDR). Instead of controlling the probability of obtaining a false positive result at all, this approach controls the expected ratio of false positives across all tests. FDR-based multiple testing correction still assumes independence of tests but is less conservative in the magnitude of correcting P-values: *E.g.*, the Benjamini-Hochberg FDR correction method multiplies the nominal P-values by the number of statistical tests similar to the Bonferroni method, but then divides these values by the ranks assigned to the P-values (with the smallest P-value having the smallest rank), hence alleviating the increase of the P-values and taking into account the intrinsic P-value pattern across

statistical tests (Benjamini & Hochberg, 1995). Throughout this thesis, we employ the FDR correction method of Benjamini and Hochberg to avoid an overly conservative adjustment of nominal P-values.

Under the null hypothesis, P-values are uniformly distributed. For a first diagnostic visualization of the number of statistically significant associations across multiple associations, the quantile-quantile plot (QQ-plot) shows deviations from the assumption of uniform distribution of P-values by comparing the observed P-values with the expected P-value distribution (-log P, respectively). These QQ-plots give a straightforward overview of how many associations have unexpectedly small P-values (*i.e.*, are significant), and if the entire association study across all tests is calibrated, *i.e.*, does not result in inflated or deflated P-values. To ensure that statistically significant associations are true and not a result of intrinsic data structure, the data can be permuted (*e.g.*, the phenotype variable *Y*) and the resulting QQ-plot is expected to show correlation between observed and expected P-values.

In the case of linear mixed models, computations associated with obtaining the MLE scale cubically with the sample size. Various approaches have been suggested to maximise the likelihood of linear mixed models while reducing the computational complexity (*e.g.*, Kang et al. (2008); Lippert et al. (2011)). While description of all these approaches is beyond the scope of this thesis, we employ the approach by Lippert et al. (2011) that was later implemented within the LIMIX toolset (Lippert et al., 2014). Briefly, Lippert et al. (2011) compute the eigenvalue decomposition of the genetic relatedness matrix $\boldsymbol{R}$ once, and then use the decomposition to project all data into a space where phenotypic variables and covariates are uncorrelated. This transformed data can then be subjected to standard association analyses (Lippert et al., 2011).

1.2.3 Convolutional neural networks in genomics

The enormous amounts of molecular data that are now available thanks to new high-throughput technology also require more sophisticated and automated statistical tools. Different from linear regression models, machine learning methods enable the modelling of all sorts of relationships within high-dimensional data, including non-linear associations (Hastie et al., 2009). Machine learning methods can be 'supervised' or 'unsupervised'. In the case of supervised machine learning, regression and classification algorithms infer a function from labelled training data; the learnt features can then be used to predict labels for unknown samples. Examples of supervised machine learning algorithms include neural networks, support vector machines and random forests. In contrast, unsupervised machine learning algorithms do not require any prior knowledge. *E.g.*, clustering and dimensionality reduction approaches like principal component analysis (PCA) or independent component analysis explore data structure by assessing intrinsic similarities between data points.

Neural networks are especially valuable for large datasets with unknown confounding structures. A neural network first learns hidden features to then make accurate predictions.

Neural networks consist of layers of connected units, so-called neurons. Each unit represents a transformation of its input values. A neural network hereby consists of at least one input and one output layer with hidden layers in between. The depth of a neural network corresponds to the number of hidden layers, and its width corresponds to the maximum number of neurons in one of its layers. A deep neural network is a neural network with a large number of hidden layers.

Deep neural networks represent powerful applications in genomics due to two advantages (**Figure 2**): First, the DNA sequence itself can serve as input. Whereas other machine learning approaches require feature extraction based on prior knowledge (*e.g.*, k-mer counts), deep neural networks learn these features from the data. This does not only make deep neural networks applicable to large-scale data, but makes high-dimensional big data a prerequisite for training these models. Second, deep neural networks can capture any sort of dependency in the DNA sequence, not only linear relationships. This includes complex interaction effects between multiple layers of genomic data. For example, deep neural networks have already been applied to predict alternative splicing (Leung et al., 2014), binding of proteins to DNA and RNA (Alipanahi, Delong, Weirauch, & Frey, 2015) and epigenetic modifications (Kelley et al., 2016) from DNA sequence.

Multidimensional inputs such as images can be analysed by a special type of deep neural network, the CNN. CNNs were originally inspired by work on the cat's visual cortex, which has simple neurons that respond to small motifs in the visual receptive field, and more complex neurons that respond to large composite motifs (Hubel & Wiesel, 1963). Here, we apply CNNs to the DNA sequence (**Figure 2**): First, DNA is one-hot encoded: One-hot encoding converts categorical to Boolean data, thus enabling the application of machine learning algorithms that require numerical data as input. After one-hot encoding of the DNA, each nucleotide represents one channel of an image-like representation of the DNA sequence. The central class of layers in a CNN is the convolutional layer, which consists of so-called filters of a predefined length that scan the one-hot encoded DNA sequence in a sliding-window approach and learn nucleotide motifs. A trained filter is activated according to how similar the DNA sequence within a window is to its learnt motif. Hereby, the two concepts of *local connectivity* and *parameter sharing* are essential to understand the advantage of a convolutional layer as compared to a fully-connected layer typically employed by traditional neural networks. *Local connectivity* means that each neuron is only connected to neighbouring neurons of the previous layer, defining the so-called receptive field of this neuron. The size of the receptive field depends on the length of the filters, *i.e.*, the length of the sliding window. *Parameter sharing* means that each filter only learns one set of parameters that it will optimise across all locations of the previous layer. In the case of DNA sequence, each filter detects a unique sequence motif. These two concepts of a CNN ensure that it can handle high-dimensional data, what would not be feasible with fully-connected neural networks where the number of parameters would normally exceed the number of available training data.

The output of the convolutional layer is then transformed in a non-linear manner by processing it in a rectifier activation layer. This is followed by a maximum pooling step that reduces the dimensionality and provides a smoother representation of the data by setting the values of a certain pooling region to their maximum value. This assumes that the exact location and frequency of a certain feature described by a filter are irrelevant for the final prediction.

The combination of these layers is repeated multiple times, *i.e.*, the output of the maximum pooling layers is fed into another convolutional layer, and so forth. The combination of these layers and their individual architecture are described by the so-called hyperparameters of a CNN. In our example in **Figure 2**, we only illustrate one of these combined layers, reflecting the architecture of the CNN used in Section 3.2.4 of this thesis. The output of these layers is typically processed by (a) fully connected layer(s), which can reduce the dimensionality of the data. Finally, the last layer of the CNN predicts the output (either by classification or regression). In our example illustrated by **Figure 2**, this layer is a classification layer, and the output is the probability of alternative splicing, *i.e.*, of skipping an alternative exon during transcription. As has been shown before (Leung et al., 2014) and as we show in this thesis based on a single-cell data (Chapter 3), the probability of alternative splicing can be predicted with a CNN based on the DNA sequence of the alternative exon and its neighbouring regions (**Figure 2**).

While a training dataset is used to optimise a CNN with a specific architecture and hyperparameter set, an independent validation dataset is used to optimise the architecture and hyperparameters of the CNN. This includes optimising the type, number and order of the CNN layers described above. An independent test dataset is then used to assess the performance of the CNN.

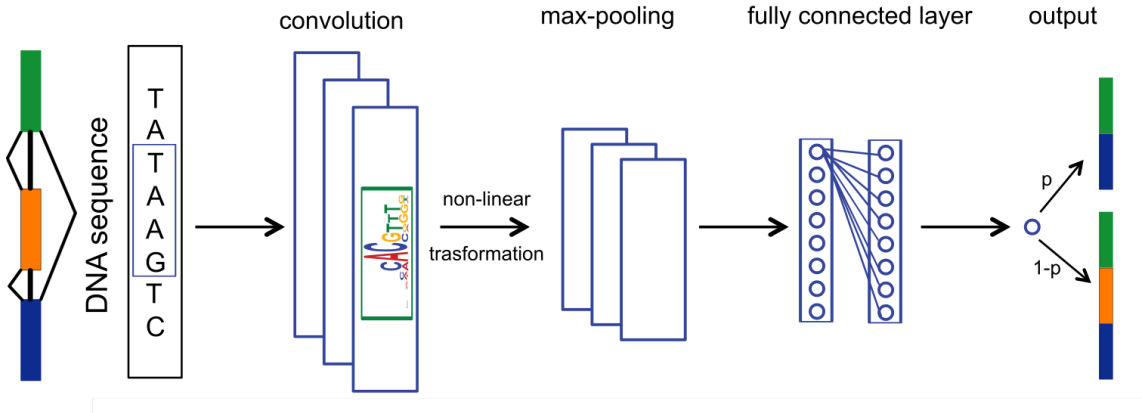

Figure 2. Scheme of a CNN applied to a DNA sequence. Typically, the DNA sequence is one-hot encoded and then processed by convolutional, non-linear, pooling and fully connected layers. A convolutional layer scans the DNA sequence with the help of filters in a sliding-window approach, and learns predictive nucleotide motifs. The output of a certain number of convolutional layers is then transformed by a non-linear activation function. The pooling layer reduces dimensionality of this output; here max-pooling summarises the values of a certain pooling region by its maximum value. The last element(s) of a CNN is/are typically (a) fully connected layer(s) that can reduce the dimensionality of the data further. The final layer of a CNN predicts the outcome. The schemes at the very left and right of this figure illustrate one of the many possible applications of CNNs to genomics, alternative splicing prediction of an alternative exon (orange genomic region in the DNA scheme on the left) that is flanked by introns and neighbouring exons both, upstream (blue genomic region) and downstream (green genomic region) of the exon itself (left scheme). This exon can either be skipped, *i.e.*, spliced-out, with probability *p* or retained, *i.e.*, spliced-in, with probability 1-*p* (right scheme). The CNN in the scheme predicts the probability of splicing based on the DNA sequence of the alternative exon and its neighbouring region.

1.3 Molecular deregulation in human cancer

Cancer describes various diseases that can affect any body part, but have in common the abnormal growth of tumour cells, which may invade adjoining body parts and metastasise to other organs. The cause of cancer is manifold; besides germline predisposition (causal for ~5-10% of all cancers), various environmental factors can lead to cancer. These environmental factors include exposure to chemical and physical agents, radiation, infection, as well as diet and lack of physical exercise. With cancer being the leading cause of human death worldwide,

its genetic and molecular characteristics have been studied in great detail, and have been leveraged for the improvement of prognosis and therapy (Weinberg, 2006).

Cancer is now understood as an 'evolutionary' disease, mostly arising from a somatic mutation in a single cell of origin that enters rapid cell proliferation, accumulating additional alterations that can be beneficial in terms of survival and proliferation and lead to expansion of so-called 'clones' of the tumour. Evolutionary forces like natural selection and genetic drift affect these clones and generate a highly heterogeneous population of cancer cells. Besides intra-tumour heterogeneity within tumour subpopulations, cancers are characterized by cancer type specific and patient-specific heterogeneity that is influenced by the germline variation of the individual patient. All this results in a highly heterogeneous regulatory landscape of cancer. Hereby, cancers harbour genetic variation from various sources: Besides patient-specific germline variants, cancers are characterised by somatic mutations that accumulate during cancer development and can stem from various different mutagenic processes, including endogenous factors like deficient DNA repair and exogenous carcinogens like viral infections, ultraviolet (UV) light or tobacco smoke. These mutations usually accumulate in specific genetic loci. *E.g.*, certain *TP53* mutations were observed across different cancer types in which they drive the activation of related oncogenic pathways (Cancer Genome Atlas Network et al., 2012). While somatic mutations can be located in or close to genes, *i.e.*, in exonic, intronic or promoter regions, they also accumulate in the vast non-coding genomic region between genes (Weinberg, 2006).

A wide range of cellular phenotypes that are relevant for cancer development and progression, including uncontrolled proliferation, immune evasion and metastasis, has been identified in human cancers (Hanahan & Weinberg, 2011). Gene expression can be leveraged to identify these cellular phenotypes and has been widely used to define molecular markers for cancer diagnosis and prognosis (*e.g.*, van ’t Veer et al., 2002). However, gene expression levels in a cancer tissue - as well as its somatic mutational load and the patients' prognosis - strongly depend on the tissue-of-origin of the tumour (Li et al., 2017). Survival analyses based on gene expression profiles of deregulated genes have hence been conducted on a per-cancer-type level (Li et al., 2017). However, similar expression patterns of certain core cancer genes can be observed across multiple cancer types; hereby, expression of so-called tumour suppressor genes protects an organism from potentially harmful cellular mechanisms and inhibits cell growth and division. The overexpression of oncogenes, on the contrary, contributes to cancer development and progression. Tumour suppressor genes and oncogenes are therefore often differentially expressed between normal and tumour tissues (Liang & Pardee, 2003). For example, the application of gene expression arrays to normal and tumour tissues allowed identifying downstream gene targets of the tumour-suppressing TF *p53* (Liang & Pardee, 2003).

Extensive somatic genetic alterations have been suggested as causative agents of these extreme gene expression alterations (Knudson, 2002; Weir et al., 2004). Somatic mutational profiles may therefore be leveraged as proxy phenotypes of molecular processes. Importantly,

mutational signatures have been shown to capture the activity of distinct mutational drivers in human cancers (Alexandrov et al., 2013; Petljak & Alexandrov, 2016). Hereby, mutational signatures delineate mutational processes by identifying genome-wide patterns of single-base mutations in a trinucleotide context. Nonnegative matrix factorization (NMF) can be used to extract these mutational signatures according to the approach by Alexandrov et al. (2013). In general, NMF is an approach to decompose and extract underlying features from complex multidimensional data (Berry et al., 2007). Briefly, in our case, a matrix **A** $\epsilon\ \mathbb{N}_0^{c\ x\ s}$ of somatic mutations in a trinucleotide context (*i.e.*, the mutated base itself and its 3' and 5' neighbouring bases; *c*=96) across patients (*s* as number of patients) can be factorised into two nonnegative matrices, the signature profile matrix **W** $\epsilon\ \mathbb{R}_{\geq 0}^{c\ x\ k}$ and the exposure matrix **H** $\epsilon\ \mathbb{R}_{\geq 0}^{k\ x\ s}$ with *k* as the number of mutational signatures (Lee & Seung, 1999). Optimal decomposition, *i.e.*, the optimal number of factors *k*, can be determined by comparison of cophenetic correlation coefficients and average reconstruction errors across possible values for *k*. Cophenetic correlation reflects how faithfully clustering can preserve pairwise distances between data points (Brunet et al., 2004).

For example, tobacco carcinogen-associated Signature 4 exhibits many CC>AA substitutions and a transcription-coupled strand bias for C>A mutations; Signature 1 has been associated with deamination of 5-methylcytosine to thymine, and has been shown to be correlated with a patient's age in a tumour- and cell type-specific manner (Alexandrov et al., 2013). In this thesis, we focus on mutational signature 3 that is associated with a failure in HR of DNA double-strand breaks and has been shown to occur in cancers with loss of *BRCA1* and *BRCA2* functionality, postulating an effect of *BRCA1* and *BRCA2* expression on mutation accumulation (Nik-Zainal et al., 2016).

There is growing evidence for the relevance of the germline background of the patient for somatic mutagenesis (Helleday et al., 2014; 2014; Nik-Zainal et al., 2014). Most famously, deleterious *BRCA1* and *BRCA2* germline variants in breast, ovarian, pancreatic and prostate cancers have been shown to contribute to characteristic mutational patterns like long indels and short SVs (Nik-Zainal et al., 2016; Lu et al., 2015; Lord and Ashworth, 2016; Alexandrov et al., 2013). Further examples include a germline CN alteration that deletes nearly the entire *APOBEC3B* gene to create an *APOBEC3A*/*APOBEC3B* fusion gene, and has been identified as a susceptibility variant in breast cancer patients. The novel gene fusion transcript has been shown to result in increased mutation load of the APOBEC-dependent mutational signature (Nik-Zainal et al., 2014). Further, increased activity of Signature 10, which is normally associated with a hypermutated phenotype in colorectal and endometrial cancers, has been linked to germline and somatic variants in the DNA polymerase epsilon, causing reduced replication fidelity and increased mutagenesis (Alexandrov et al., 2013). Moreover, patients harbouring specific germline variants in genes related to DNA mismatch repair (MMR) pathways show significantly increased indel rates in repetitive regions, a typical hallmark of MMR-deficient individuals (Karran, 1996).

To identify associations between germline and somatic variants and gene expression, previous integrative efforts (The Cancer Genome Atlas (TCGA; Cancer Genome Atlas Research Network et al., 2013); The International Consortium of Cancer Genomes (ICGC; Zhang et al., 2011)) obtained gene expression and exonic genetic variation data across patients and cancer types. First approaches identified genes that are recurrently altered by somatic mutations and whose mutation state is associated with gene expression levels. *E.g.*, Masica and Karchin (2011) correlated mutation state and expression levels across genes to detect potential cancer driver genes. Ding et al. (2015) extended this approach to assess the impact of individual mutations within genes by analysing somatic mutations and gene expression across 12 cancer types: Their Bayesian hierarchical model predicts both, the probability that a recurrently mutated gene influences gene expression across patients, and the probability that an individual mutation influences expression in a specific patient. Like this, they identified 30 novel *cis*-modulated tumour suppressor genes, and somatic mutations in 150 genes that show *trans* associations with specific cancer-related expression networks.

Non-coding genetic variation has been out of scope for these studies due to their restriction to whole-exome sequencing data. The role of the large number of non-coding somatic variants for cancer development therefore remains to be fully understood (Cancer Genome Atlas Research Network et al., 2013). The assessment of their functional consequences is rendered even more difficult by the fact that most somatic mutations are very rare or even private to individual tumours due to restricted sample sizes of existing cancer cohorts, making standard association studies (Section 1.2) impossible. Fredriksson et al. (2014) screened 14 cancer types for associations between somatic mutations in gene regulatory regions and gene expression changes. While they found many recurrently mutated promoter regions, they were only able to confirm *TERT*'s promoter to be strongly associated with (increased) expression. This suggests that the effect of non-coding variation on gene expression has to be queried beyond the promoter regulatory region. Weinhold et al. (2014) have already shown that other genome-wide regulatory regions like 5' and 3' untranslated regions (UTRs) and distal enhancer regions as annotated by Ensembl (Hubbard et al., 2002) are recurrently mutated across different cancer types. Functional somatic mutations are, however, expected to be located across the entire genome, beyond Weinhold's functionally annotated genomic regions; their role in the molecular deregulation of cancerous cells remains to be determined (PCAWG Transcriptome Core Group et al., 2018).

2 The gene regulatory landscape in human cancer

Contributions

The Pan-Cancer Analysis of Whole Genomes (PCAWG) study (Campbell et al., 2017) is an international initiative that aims to identify pan-cancer mutational patterns in more than 2000 whole genomes (n=2,658) from 39 cancer types. In this section, I describe my contributions to the research of the PCAWG working group 3 for the integration of transcriptomic and genomic data. This working group focuses on the analysis of a subset of 1,188 whole genomes of the PCAWG cohort for which gene expression data is available.

I led the studies to investigate associations between gene expression and mutational signatures, and the associated role of germline determinants. Besides, I collaborated on the PCAWG analyses concerning allele-specific expression, the effect of germline genetic variants on gene expression and the effect of somatic genetic variants on gene expression. While the entire PCAWG consortium and particularly PCAWG working group 3 were involved in this project, crucial contributions to the presented work have been made by Claudia Calabrese, Kjong Lehmann, Fenglin Liu, Roland Schwarz and Nuno Fonseca. Oliver Stegle, Alvis Brazma, Gunnar Raetsch, Zemin Zhang and Angela Brooks guided and supervised the analyses.

Here, I present the results of my main project and my key collaborative efforts in the somatic variation and allele-specific expression studies. Other PCAWG-3 analyses I contributed to can be found in Calabrese et al. (2017), PCAWG Transcriptome Core Group et al. (2018), Waszak et al., (2017) and Campbell et al. (2017).

All scripts will be available on github after peer-reviewed publication. The manuscript is currently under late-stage revision at Nature.

2.1 The functional relevance of somatic mutagenesis in cancer

Cancer is characterised by extensive somatic genetic alterations that result in cellular phenotypes which are often relevant for disease, including uncontrolled proliferation, immune evasion and metastasis (Knudson, 2002; Weir et al., 2004). Such variations comprise SNVs, indels, and SVs such as SCNAs. Somatic mutagenesis is in part explained by environmental and intrinsic risk factors that are independent of the cancer itself, with growing evidence that somatic mutations have an early onset and occur during healthy ageing (*e.g.*, Kennedy et al., 2012). The prevalence and the consequences of somatic mutations are also increasingly linked to germline factors of the patient (Alexandrov et al., 2013; Karran, 1996; Nik-Zainal et al., 2014; Pleasance et al., 2010; Waszak et al., 2017). The relationship between germline mutations in MMR-related genes and increased somatic indel load in repetitive regions of the genome has been known for some time (Karran, 1996). Since then, several studies have implicated both rare and common germline variants in mutational processes and in the patterns of somatic variation. Alexandrov et al. (2013) have found associations between germline and somatic variants in the DNA polymerase epsilon and specific genome-wide somatic mutational patterns associated with hypermutation in colorectal and endometrial cancers. A very strong relationship between germline predisposition and somatic mutagenesis has been unravelled by Nik-Zainal et al. (2014) and concerns a germline deletion that removes nearly the entire *APOBEC3B* gene and fuses its remaining region to its neighbour, *APOBEC3A*. This germline variant leads to increased activity of the APOBEC-related mutational processes and to increased susceptibility to breast cancer (Nik-Zainal et al., 2014). Further, patients with breast, ovarian, pancreatic or prostate cancer that carry pathogenic germline variants in the *BRCA1* and *BRCA2* gene exhibit characteristic mutational patterns like long indels and short SVs (Nik-Zainal et al., 2016; Lu et al., 2015; Lord and Ashworth, 2016; Alexandrov et al., 2013). In pediatric medulloblastoma, germline *TP53* mutations lead to massive chromosome rearrangements known as chromothripsis (Rausch et al., 2012). Most recently, Waszak et al. (2017) have integrated rare and common germline and somatic variants across 39 cancer types to uncover previously unknown relationships between these genetic factors. This study has emphasised the relevance of germline variants in cancer predisposition and DNA repair genes for mutational processes; *e.g.*, protein-truncating germline variants in the *MBD4* gene have been shown to cause increased activity of clock-like mutational processes, an effect that can be observed across cancer types.

While these findings emphasise the relevance of associations between germline and somatic genetic factors, their combined interplay with molecular and cellular functions is not yet fully understood. Approaches that can shed light on this question include association analyses between genetic variation and molecular readouts such as gene expression levels. Previous efforts using exome-sequencing and transcriptomic data from TCGA (Cancer Genome Atlas

Research Network et al., 2013) and ICGC (Zhang et al., 2011) have identified associations between somatic variants in coding regions and gene expression. Although these studies have helped identify and characterise regulatory drivers, the role of the much larger number fraction of non-coding somatic variants is not yet fully understood (Cancer Genome Atlas Research Network et al., 2013; Kanchi et al., 2014). In addition, most somatic changes are rare or private to individual tumours and their functional consequences are difficult to study. Recent studies have begun to address this by identifying genomic loci that are recurrently altered by somatic mutations and associated with gene expression alterations (Section 1.3). Most work has, however, focused on the effects of variation in promoters of established cancer-genes, including *TERT* and *BCL2* (Ding et al., 2015; Fredriksson et al., 2014; Smith et al., 2015; Weinhold et al., 2014; Section 1.3).

So far, a comprehensive analysis of associations between (non-coding) somatic variation and gene expression is missing. In addition, the potential regulatory effect of germline variants on the link between somatic mutations and gene expression remains to be fully understood. In the PCAWG transcriptome study, we assessed the local (*cis*) and genome-wide (*trans*) effect of these variants on gene expression across cancer types (**Figure 3a**). Our approach combined complementary strategies, including allele-specific expression (ASE) analyses, somatic and germline eQTL mapping and the analysis of gene expression associations with global somatic signatures (**Figure 3b**). This integrative analysis allowed us to derive a pan-cancer regulatory map of genetic and gene expression variation (PCAWG Transcriptome Core Group et al., 2018).

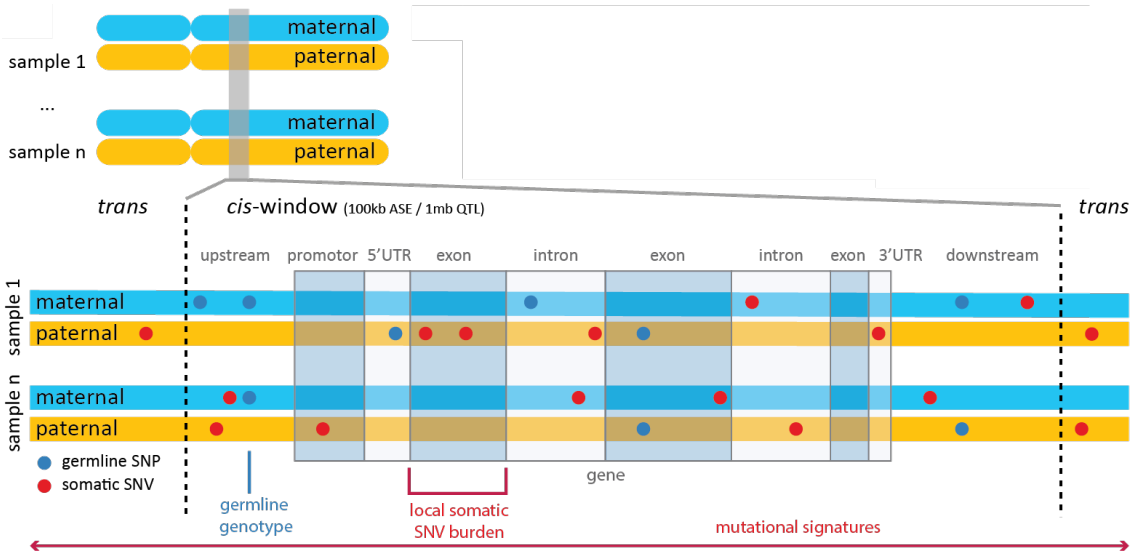

Figure 3. Integrative analysis of the gene regulatory landscape in human cancers. Overview of the sources of genetic variation considered in the analysis. For analyses of *cis* regulation, germline SNPs (blue) were individually tested for association with total gene expression using standard eQTL approaches. Due to their low recurrence in the cohort, somatic SNVs (red) were aggregated in burden categories depending on their position relative to the gene tested (*e.g.*, promoter, 5' UTR, intron). Local SNV burdens were then tested for association with ASE globally across all genes, as well as with total expression on a per-gene level using eQTL approaches. *Trans* effects were estimated by testing total gene expression for association with mutational signatures. Window sizes were 1M base pairs for all somatic *cis* eQTL analyses and 100k base pairs for ASE and germline *cis* eQTL.

2.2 The PCAWG study

The PCAWG initiative has derived a comprehensive map of genetic aberrations in 2,658 human cancers across 39 different cancer types (Campbell et al., 2017). We analysed the tumour whole-genome sequencing (WGS) and matched RNA-Seq data that were available for a subset of 1,188 patients, and carried out joint genetic analyses that integrate coding and non-coding somatic variation with germline variants to investigate regulatory effects on gene expression levels across altogether 27 cancer types.

SNP, SNV and allele-specific SCNA calls from the 1,188 patients were obtained from PCAWG working group 1. Briefly, SNPs were derived from WGS data of the normal (blood) tissue of the respective patients; local SNVs, allele-specific SCNAs and mutational signatures were then derived from the tumour in comparison to the normal WGS samples. The 1,188 patients are spread across 27 cancer types and 29 project codes and include 899 carcinomas. The project code defines sequencing centre and cancer type of the respective sequencing study. All samples are primary tumours, except for 34 metastatic and 13 recurrent ones.

Gene expression values (measured in FPKM; Section 1.1.2) were obtained from other members of the PCAWG working group 3 (Tophat2/Star gene expression; see PCAWG Transcriptome Core Group et al. (2018) for details). Gene expression data pre-processing was realised in collaboration with various members of the PCAWG working group 3. Genes with FPKM ≥ 0.1 in at least 1% of the patients (12 patients) were retained, resulting in 47,730 genes. Out of these genes, a subset of 18,898 protein-coding genes (based on Gencode version 19) was used for the subsequent analyses. We subjected the $\log_2$-transformed expression values to peer analysis to account for hidden covariates in this highly heterogeneous dataset (Stegle et al., 2012). In order to balance number of covariates, statistical power and available sample size per cancer type, we followed the GTEx protocol and estimated 35 hidden peer factors to be used (Lonsdale et al., 2013). Peer residuals were then rank-standardised across patients.

Building on these 1,188 consistently processed genomes and transcriptomes we derived a detailed regulatory map that considers different associations between germline and somatic genetic factors and gene expression (PCAWG Transcriptome Core Group et al., 2018).

2.3 Associations between mutational signatures and gene expression

Global variations in mutational patterns across individuals, for example due to generic environmental factors or exogenous damage, can be quantified using mutational signatures. These mutational signatures tag mutational processes specific to their tissue-of-origin and environmental exposure (Alexandrov et al., 2013). However, the pan-cancer relationship

between genome-wide mutational signatures and gene expression levels has not been studied yet, and the regulatory origin and/or effect of these signatures remains poorly understood.

2.3.1 Methods

NMF was performed by the PCAWG working group 7 per cancer type and across cancer types (Alexandrov et al., 2018). The final set of signatures were manually determined by the PCAWG working group 7 by including additional signatures from other datasets, validating particularly dominant signatures, and, where available, supplementing prior experimental evidence (Alexandrov et al., 2018). Alexandrov et al. (2018) extracted a vast range of different signatures, including globally distributed and clustered signatures, single base substitution, doublet base substitution, indel and composite signatures, and signatures derived with the help of different statistical tools. In this thesis, we focus on the canonical mutational signatures that are based on single base mutations and were extracted by SigProfiler, a tool that determines mutational signatures and their contribution in accordance with the signatures defined by the Catalogue of Somatic Mutations in Cancer (COSMIC; Forbes et al., (2008); Forbes et al. (2017); Alexandrov al. (2013)). With SigProfiler, Alexandrov et al. (2018) extracted 49 high-confidence single base substitution signatures across cancer types.

In this analysis, we focused on a subset of 1,159 of our altogether 1,188 patients for which mutational signature profiles were available. Altogether, we used the 39 mutational signatures that were labelled according to the COSMIC system by Alexandrov et al. (2018) (version of the mutational signature set: PCAWG-7 beta 2 release). Gene expression data of 18,831 protein-coding genes with FPKM ≥ 0.1 in at least 1% of the 1,159 patients was retained, and corrected for hidden peer factors as described above (Section 2.2).

Signatures with zero variance and prevalence below 1% in their exposure across the 1,159 patients were removed. Like this, we obtained 28 signatures. We then applied linear mixed models implemented in the LIMIX package (Lippert et al., 2014) to test for associations between the exposure to these signatures (from now on just referred to as 'signatures') and expression of all quantified protein-coding genes across all 1,159 patients. This entailed 28 x 18,831 association tests between each signature and each gene. We queried the same associations in a subset of 877 carcinoma patients or a subset 891 European patients to assess robustness of the associations (Section 2.3.2 for details).

In the linear mixed models, we accounted for known confounding factors by modelling them as fixed effects, and for population structure that we modelled as a random effect. Specifically, we accounted for sex, project code, per gene CN status, total somatic mutational burden (number of SNVs and indels) and sample purity (PCAWG Transcriptome Core Group et al., 2018). Per gene CN alterations were derived as the average CN across all CN alterations called within the annotated gene boundaries. Sample purity was obtained from PCAWG working group 11 who based their purity estimates on CN segmentation. The population structure was assessed via a

kinship matrix. This kinship matrix was calculated as an empirical patient-by-patient covariance matrix based on every 20th germline variant (PCAWG Transcriptome Core Group et al., 2018).

The inclusion of the total somatic mutational burden per patient as fixed effect covariate allows for studying the relative effect of each mutational signature. The signatures are therefore negatively correlated with each other, a problem our model did not account for due to independent association testing of each signature. The problem of interdependencies between relative proportions of the whole data is widely known in the statistical analysis of compositional data (Aitchison, 1982). Aitchison (1982) suggested representing compositional data in a log-ratio space by relating individual compositions to one representative, arbitrarily chosen composition. This relies on the assumption that various possible compositions are equivalent to each other when projected to a simplex, a space defined by the Aitchison geometry that compresses all equivalent compositions to unique compositions. When these representative compositions are projected into a two-dimensional space, independent ratios of the initial compositions are obtained - in our case ratios of mutational counts of different signatures - which can be used in independent association studies. As we are, however, not interested in the effect of ratios but in the effect of individual signatures onto gene expression, this approach is not valuable for a gene expression-signature association study. However, we emphasise that only an approach that handles relative signatures as compositional data ensures independence between signatures. The signature association studies presented in this thesis might therefore lead to spurious negative correlations introduced by the inherently negative correlations between relative signatures. Accordingly, Temko et al. (2018) have already discussed previously that analyses of a mutational signature in patients who also show evidence of other signatures can only approximate the contributions of this signature.

2.3.2 Gene expression deregulation linked to mutational signatures

We firstly investigated if the linear mixed model was calibrated and did not result in any spurious associations. We employed QQ plots to compare observed and expected P-value distributions. Under the null hypothesis of no associations, the resulting P-values of an association analysis should be uniformly distributed and follow the diagonal in a scatterplot of the ranked observed over the ranked expected P-values ($-\log_{10}$ P, respectively; **Figures 4a-c** for examples; see Section 1.2.2 for details). In our case where we tested associations between mutational signatures and many protein-coding genes (the majority of which are expected to have generic functions independent of mutational processes), we expected some significant associations that deviate from the diagonal in the QQ plot, but also many non-significant associations that follow the assumptions of the null hypothesis and are therefore located close to the diagonal in the QQ plot.

Besides assessing the distribution of the nominal P-values of our association studies, we applied permutations tests by permuting the expression values per gene across patients.

Permutation (or randomisation) tests are statistical tests that obtain the distribution of the test statistic under the null hypothesis by calculating all possible values of the test statistic after randomisation of the data points (in our case, after rearrangement of patient IDs in our gene expression matrix). As we used the permutation test to study the calibration of our association analysis and not to assess an exact P-value, we restricted our permutation test to n=1,000 random permutations. As permutations break up the true associations, the P-values of the permutated tests are expected to follow the assumptions of the null hypothesis, which assumes the absence of associations. Therefore, the P-values resulting from each permutation are expected to follow the diagonal in a QQ plot. Deviations from this expectation suggest that the statistical test is not well calibrated and either inflated or deflated due to un-accounted data structure (Section 1.2.2).

The QQ plot of permuted P-values obtained from our model showed calibration of the analysis and the presence of true significant associations (n=1,000 permutation of the patient IDs in the gene expression matrix; only the result of one representative permutation test is shown; **Figure 4a**). The nominal P-values, however, appeared to be inflated; hereby, cancer type is a known strong driver of mutational signature variability, and was therefore a likely candidate for creating spurious associations. Hence, we tested a different permutation scheme that accounts for cancer type related structure by permuting only patients within each cancer type. As the P-values permuted according to this scheme also closely followed the diagonal in the QQ plot (assessed within the analyses of Section 4.2.2; see **Figure 13e** with a per cancer type permutation scheme versus **Figure 13a** with an across all patients permutation scheme), we however concluded that the associations detected in **Figure 4a** and in subsequent signature association studies were statistically valid.

A volcano plot that compares significance and effect size of associations shows that several mutational signatures were strongly associated with the expression of multiple genes (**Figure 4e**).

Our cancer cohort is a heterogeneous dataset composed of patients with various different cancer types and from multiple ethnicities. Although we accounted for known confounding factors in our analyses and for hidden structure in our gene expression data, we wanted to ensure that our association analysis was robust and did not result in spurious associations due to more complex data structure. We therefore tested if the detected associations could be replicated in more homogeneous subsets of our cancer cohort. To still retain majority of the samples for each replication association study, we decided to focus on the subset of European patients to understand the potential confounding effect of ethnicity, and on the subset of carcinoma samples to understand the potential difference between carcinoma and lymphoma tumours. Both association studies were calibrated and resulted in statistically significant associations (**Figures 4b-c**). While both replication association studies resulted in slightly less significant associations than the original association study (FDR ≤ 10%; **Figure 4d**), the P-values across individual signature-gene pairs were strongly correlated (Figure **4f** for a

comparison between carcinoma and full association study, Figure **4g** for a comparison between European based and full association study). The reduced power might therefore be attributable to the reduced sample size.

After confirming the robustness of the linear mixed model analyses in these data subsets, we restricted our downstream analyses to the full dataset.

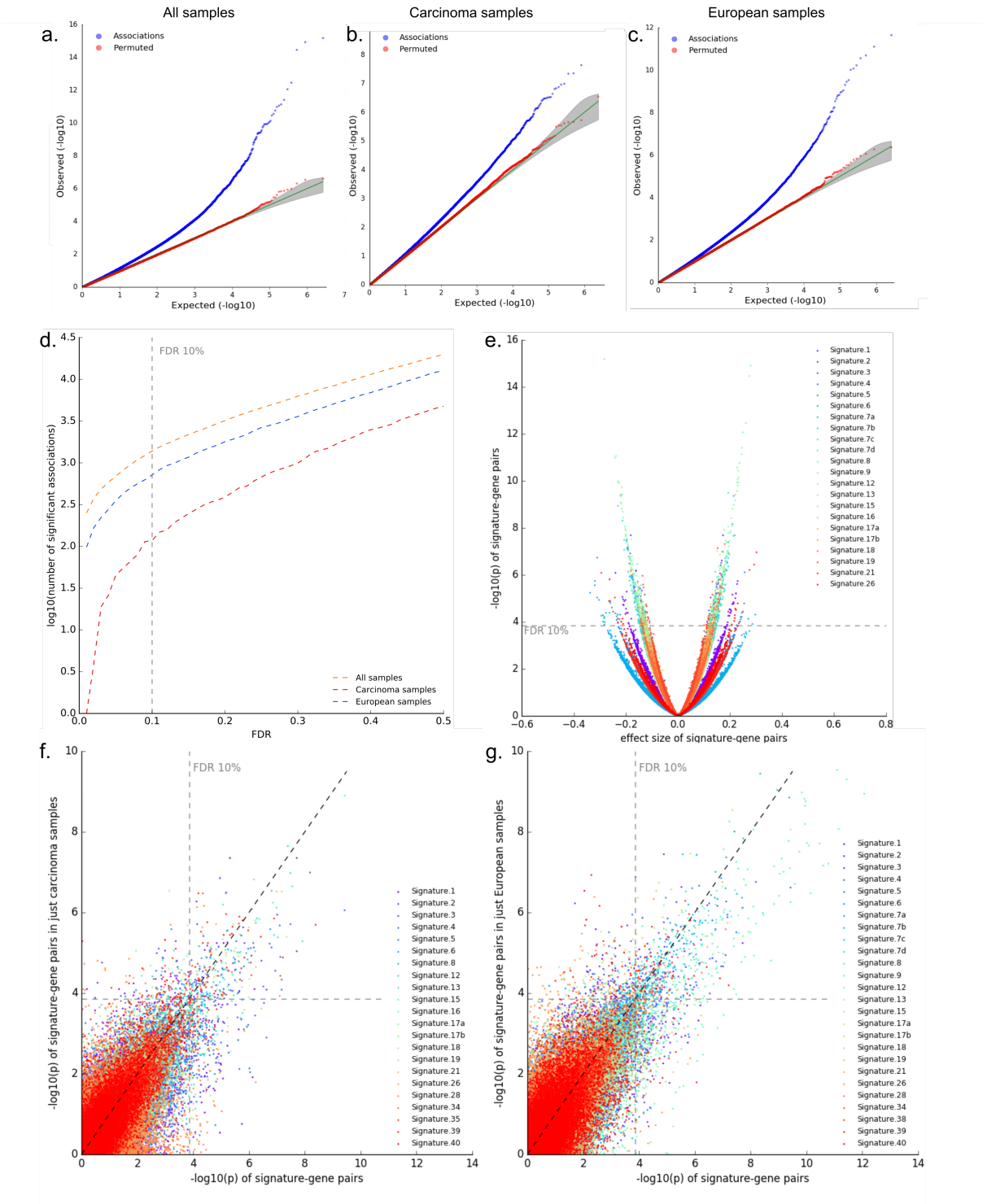

Figure 4. See next page for figure caption.

Figure 4. Quality control of the gene expression-mutational signature association studies. **a-c.** QQ plots of the P-values ($-\log_{10}$ P) of the linear mixed model associating expression of 18,831 genes with 28 signatures across **a.** all 1,159 samples, **b.** 877 carcinoma samples, or **c.** 891 European samples. Nominal P-values are shown in blue, permuted P-values are shown in red. Gene expression values are corrected for 35 peer factors. **d**. Plot of number of significant associations ($\log_{10}$ P) at different FDR thresholds (across all, carcinoma and European samples). **e.** Volcano plot of P-values ($-\log_{10}$ P) over effect sizes across gene-signature associations in the full dataset. Data points are coloured according to the mutational signature involved in the respective gene-signature association. **f-g.** Scatterplot of P-values ($-\log_{10}$ P) across gene-signature associations to comparison the full dataset analysis with the analyses performed on only **f.** carcinoma (Pearson correlation coefficients r=0.763), or **g,** European samples (Pearson correlation coefficients r=0.789), respectively. Data points are coloured according to the mutational signature involved in the respective gene-signature association.

Across all patients, this association study identified 1,176 genes significantly associated with at least one signature in altogether 1,388 unique associations after Benjamini-Hochberg multiple testing correction (FDR ≤ 10%, multiple testing was applied across all signature-gene pairs; see Supplementary Table 15 of PCAWG Transcriptome Core Group et al. (2018) for statistical results per signature-gene pair).

The lymphoma/leukaemia-typical Signature 9 showed the largest number of associations (n=354), followed by the smoking-related Signature 4 (n=119). Whereas many of the remaining mutational signatures only affected a small number of genes (6 to 95 genes, with 6 genes in the case of Signature 34 and 95 genes in the case of Signature 7a, median number of genes across all signatures: 34.5), 18 signatures showed particularly prominent effects and affected the expression of more than 20 genes. While the vast majority of genes (85.8%) were associated with only a single signature (1,009 genes), 129 genes were associated with two, 32 with three, five with four and one with five signatures.

Notably, the set of identified genes constituted a markedly different set of genes compared to associations with total mutational burden alone (total number of SNVs): When we correlated the P values ($-\log_{10}$ P) of the associations of the mutational burden with the respective signature-specific P values, the absolute Pearson correlation coefficients always remained below 0.1 (see Supplementary Table 15 of PCAWG Transcriptome Core Group et al. (2018) for exact correlation coefficients). This confirmed that we discovered signature-specific molecular deregulation, which was independent of the total number of SNVs.

While some signatures have clear aetiologies and are well known in the cancer research field, others have not been fully characterised yet. To annotate these ambiguous signatures *de novo*, we considered the 18 signatures that were associated with at least 20 genes, and assessed functional enrichment using both, Gene Ontology and Reactome Pathways categories (Fabregat et al., 2018; Milacic et al., 2012). The gene enrichment was performed with the Bioconductor packages biomaRt (Durinck et al., 2009), clusterProfiler (Yu et al., 2012) and ReactomePA (G. Yu & He, 2016), considering the set of 18,831 protein-coding genes that were tested in the linear mixed models as background gene set. Benjamini-Hochberg multiple testing correction was applied across all categories per signature (FDR ≤ 10%; see Supplementary Table 15 of PCAWG Transcriptome Core Group et al. (2018) for all significant enrichments).

We found that 11 signatures were enriched for at least one category (FDR ≤ 10%), revealing associations consistent with known aetiologies (**Figure 5a**). For example, lymphoma Signature 9 was associated with 354 genes enriched for lymphocyte/leukocyte-related and immune response-related processes ($P=1x10^{-6}$ for both categories), including *TCL1A*, *LMO2* and *TERT* ($P=1x10^{-10}$, $7x10^{-10}$, $2x10^{-9}$, respectively). The smoking Signature 4 was associated with 119 genes enriched for biological oxidation processes ($P=5x10^{-4}$). Processing of benzo[a]pyrene is a typical biological oxidation process that results in breaking down of this tobacco carcinogen. One of the associated genes was *CYP24A1* (P=0.001, β=-0.193), a gene that is known to be down-regulated in tobacco-smoke exposed tissue (**Figure 5b**; Woenckhaus et al., 2006). The 70 genes associated with APOBEC-related Signature 2 were significantly enriched for DNA deamination pathways ($P=6x10^{-6}$; **Figure 5a**).

Among signatures with unknown aetiology, our identified associations linked Signature 38 to 39 genes enriched for melanin processes, vitamin D response and pigmentation ($P=3x10^{-5}$; **Figure 5a**). While Signature 38 has not been functionally annotated yet, it is strongly correlated with the canonical UV Signatures 7 across the patients of our cohort (*e.g.*, correlation with Signature 7a: $r^2=0.375$, $P=5\cdot10^{-40}$). Melanin synthesis causes oxidative stress to melanocytes (Denat et al., 2014; Kvam & Tyrrell, 1999) and we found Signature 38 associated with the oxidative stress promoting gene *TYR* ($P=1x10^{-4}$, Jimbow et al., 2001). A hallmark of Signature 38 are C>A mutations, also a typical product of ROS mediated by activity of 8-hydroxy-2′-deoxyguanosine (Valavanidis et al., 2009). This suggests that Signature 38 may capture DNA damage indirectly caused by UV after direct sun exposure due to oxidative damage (Premi et al., 2015), with *TYR* as a possible mediator of the effect (PCAWG Transcriptome Core Group et al., 2018).

Signature 8 has been found to be prevalent in medulloblastoma (Forbes et al., 2017); our association studies linked this mutational signature to 25 genes enriched for ABCA-transporter pathways (P=0.011; **Figure 5a**). Whereas the PCAWG dataset does not contain any medullablastoma samples, this association might point towards a cancer type-independent relevance of Signature 8 that could be leveraged for drug repurposing: Drugs targeting the ABC(A) transporter pathways have already undergone clinical trials for treating medulloblastoma (Ingram et al., 2013), and preliminary results point towards a relationship

between ABC(A) transporter inhibition and decreasing radiation therapy and chemotherapy resistance. Specifically, Ingram et al. (2013) discovered that cells from medullablastoma cell lines that survived radiation exposure and were still carcinogenic *in vivo* showed elevated levels of different ABC(A) transporters. One of them, ABCA1, was also shown to be elevated in independent cell cultures obtained from medullablastoma patients. Ingram et al. (2013) then discovered that drugs like verapamil that inhibit multiple ABC(A) transporters led to radiation sensitization and had anti-proliferative effects in patient-derived medullablastoma cells. Ingram et al. (2013) also found that subtypes of ABC(A) transporters were associated with subtypes of medullablastoma, with, *e.g.*, ABCA8 being associated with Sonic Hedgehog-dependent subtypes. The deregulation of ABC(A) transporters and with it the genome-wide pattern of Signature 8 might therefore constitute a possible genetic marker that could be queried for future diagnostic classification of medullablastoma subtypes. As a side note: Othman et al. (2014) later confirmed the potentially beneficial application of ABC(A) inhibitors to high-risk medullablastoma, and proposed a combinatorial treatment of chemotherapy and ABC(A)-inhibiting drugs especially for metastasing tumours. Interestingly, the anti-proliferate effect of verapamil on medulloblastoma, pinealoblastoma, glioma, and neuroblastoma cell lines was already observed by Huber et al. (1988) although they were not yet able to understand the causative molecular mechanisms. Verapamil is commonly known as a calcium channel blocker that is, *e.g.*, used for the treatment of high blood pressure; however, Huber et al. (1988) did not observe any alterations in calcium influx or efflux, suggesting - already at that time - a different underlying molecular process leading to anti-proliferation.

In summary, there is strong evidence for an interplay between ABC(A) transporters and medullablastoma progression and metastisation. Given our association between these transporters and Signature 8 across cancer types and the previously known increased prevalence of Signature 8 in medullablastoma tumours, Signature 8 might constitute a biomarker for the potentially beneficial treatment with ABC(A) transporter inhibitors. However, as the key ABC transporter detected by both, Ingram et al. (2013) and Othman et al. (2014), were not contained in our list of protein-coding genes due to low gene expression levels across patients (Section 2.3.1), we were not able to state a direct relationship between the expression of these transporters and Signature 8.

The enrichment of genes associated with other mutational signatures remained even more elusive. Age-related Signature 1 was associated with translational elongation and peptide chain elongation events ($P=5x10^{-5}$ for both pathways; **Figure 5a**). While translational elongation describes a quite broad biological process, its role in oncogenesis has already been investigated. *E.g*, the eukaryotic translation elongation factor 1 alpha (*eEF1A*) gene was found to be overexpressed in metastatic compared to non non-metastatic cells and also plays a role in apoptosis regulation across cancer types (Lamberti et al., 2004). Our finding might therefore point towards an age-dependent role of regulatory mechanisms in protein synthesis for cancer progression.

Signature 18 has been commonly found in neuroblastoma (Alexandrov et al., 2013; Petljak et al., 2019) and was associated with damage due to reactive oxygen species (ROS; Viel et al., 2017). We found Signature 18 to be associated with peptide cross-linking ($P=6x10^{-6}$; **Figure 5a**) that in turn is known to be a product of excessive ROS production (Sharma et al., 2012). This might confirm the suggested link between Signature 18 and ROS on the gene expression level.

Alexandrov et al. (2018) recently extended the catalogue of single base substitution mutational signatures to more than the 30 ones defined by Alexandrov et al. (2013) and Forbes et al., (2017). This catalogue links Signature 35 to chemotherapy treatment with platinum drugs (Boot et al., 2018). We found this mutational signature to be associated with platelet adhesion ($P=6x10^{-4}$; **Figure 5a**). Platelet adhesion is known to create a favourable microenvironment for chemoresistance in gastric cancers (Saito et al., 2017). We therefore hypothesised that chemotherapy resistance after first treatments might not only arise from the selective survival of resistant clones, but possibly also from the *de novo* introduction of Signature 35 mutations. This hypothesis would however have to be supported by more detailed and extensive experimental evidence.

2.3.3 Towards causality of associations between mutational signatures and gene expression

While the analysis in Section 2.3.2 identified associations between somatic genetic variation and gene expression across cancer types, it is - unlike in association studies based on germline genetic variation - not clear *a priori* what is cause and what is consequence in these relationships. However, germline variants can act as an anchor of such analyses since they are inherently the causal part of any genetic association. To reduce the search space for genetic anchors and simultaneously identify likely causal germline players in these associations, we identified lead genetic variants of germline eQTL of signature-associated genes. The lead variant is the most significant SNP per eQTL and assumed to be likely causal for the variable gene expression, or at least in LD with the actually causal variant. By testing for associations between these variants and the mutational signatures that were associated with the respective genes, we could gain directed mechanistic insight into the relationship between gene expression and mutational signatures. This eQTL-based approach entailed substantially fewer tests than genome-wide germline analyses (*e.g.*, Waszak et al., 2017), providing our analyses with a significantly reduced multiple testing burden.

We firstly set out to perform *cis* germline eQTL mapping of all genes. Briefly, we employed linear mixed models in LIMIX (Lippert et al., 2014) to associate gene expression with local germline genetic variants, using a random effect to account for population structure and accounting for additional fixed effect covariates as described in the Methods (Section 2.3.1). In terms of gene expression, we used the gene expression residuals after correction for peer

factors (see Section 2.2 for a description of the peer factor analysis). In terms of local germline genetic variants, we considered common SNPs (MAF ≥ 1%) that were located proximal to individual genes (±100k base pairs from the gene boundaries). For each gene, we used the linear mixed model to test for statistical significance between these germline variants and the respective gene's expression. We used Bonferroni correction to adjust for the number of independent tests per gene; the number of independent tests was estimated based on local LD. Thus, we determined significant eQTL and their lead variants (P_{adj} ≤ 5%). After applying this approach across all genes, we performed a global multiple testing correction according to Benjamini Hochberg across all lead variants, identifying 3,509 genes with a significant eQTL (FDR ≤ 5%).

Among our 1,176 signature-linked genes, 197 genes had a germline eQTL, and these 197 genes were uniquely associated with altogether 26 of our 28 tested mutational signatures. If a lead variant of a germline eQTL was also significantly associated with the signature the respective gene was associated with, it could be used to query the directionality of the gene expression-signature association by causal association analysis. We therefore tested the association between the lead variants of the 197 genes and the respective mutational signatures across 2,507 patients. These 2,507 patients constituted the subset of PCAWG patients for which mutational signature profiles, germline variant calls and all necessary covariates were available. We accounted for the same covariates as in the mutational signature-gene expression association studies (Section 2.3.1). This entailed 197 association tests between 197 genetic variants and their corresponding signatures; P-values were adjusted for multiple testing following the Benjamini Hochberg procedure.

Only the *APOBEC3B* eQTL lead variant rs12628403 was significantly associated with its corresponding signature, Signature 2, after multiple testing ($P=5x10^{-7}$, FDR ≤ 10%, **Figure 5c**). This germline variant is a known risk variant for Signature 2 prevalence (Middlebrooks et al., 2016).

To investigate if we had found the shared and true causal germline variant for *APOBEC3B* expression and Signature 2, we performed proportional colocalisation analysis with Bayesian Model Averaging (BMA) (Wallace, 2013). This analysis tests if two molecular phenotypes, here gene expression and mutational signature, share a common causal genetic variant. Briefly, it tests the null hypothesis of colocalisation that assumes that the two phenotypes that share a causal variant have proportional regression coefficients against any set of variants that are close to the causal variant. An inherent problem of this analysis is the choice of the putatively causal variant. In our germline eQTL analyses, we had focused on the lead variant of the eQTL, *i.e.*, the variant most strongly associated with gene expression. Using this lead variant as potentially causal variant, might, however, introduce bias in the regression coefficients of the proportional colocalisation analysis and lead to an increased likelihood of falsely rejecting the null hypothesis of colocalisation. We therefore decided to use the BMA approach that models various genetic variants that are close to our lead variant as potential causal variants and then

generates P-values averaged across all models. We used the implementation of proportional colocalisation analysis by the R-package coloc (version 3.1; Wallace, 2013). An intrinsic caveat of proportional colocalisation analysis is that the non-rejection of the null hypothesis can not be attributed to true colocalisation or to a lack of power (Wallace, 2013). Hence, while we were able to state that our colocalisation analysis suggested germline variant rs12628403 as a plausible genetic determinant of both, *APOBEC3B* expression and Signature 2 prevalence (P=0.398, *i.e.*, the null hypothesis of colocalisation cannot be rejected), we were not able to rule out a lack of power as underlying cause.

To study the relationship between gene expression and mutational signature in more detail, and to assess directionality of the effect between the germline variant rs12628403, gene expression of *APOBEC3B* and Signature 2, we performed causal mediation analysis (Baron & Kenny, 1986; Preacher & Hayes, 2004). Briefly, mediation analysis hypothesises a causal chain between three variables and mathematically describes them as a set of linear regression models, which when fitted can be leveraged to estimate the mediation effect (Baron & Kenny, 1986). We modelled the genetic variant as the causal variable, and considered two models that either assumed signature or gene expression to be the mediating variable and the respective remaining variable to be the outcome variable.

We firstly used a structural equation model (function *sem*) from the R package lavaan (Rosseel, 2012) to compare the statistical effects predicted by these two possible models. The total effect of the causal variable on the outcome variable (*t*) was modelled as the sum of the direct effect between the two variables (*c*) and the indirect effect between the two variables via the mediating variable. This indirect effect can be calculated as the product of the effect of the causal variable on the mediating variable (*a*) and the effect of the mediating variable on the outcome variable (*b*). Only if the following requirements of the fitted model are met, is mediation of the proposed mediating variable statistically possible: (i) The effects *t* and *a* have to be statistically significant, and (iia) either the effect *c* has to be non-significant or (iib) the effect size of *c* has to be smaller than the effect size of *t*. In the case of (iia), full mediation of the effect via the mediating variable can be assumed, whereas in the case of (iib) partial mediation can be assumed (Preacher & Hayes, 2004). Whereas requirement (i) was met by both our models at a significance level of α=5%, (iia) did not apply to either of them and (iib) only to the model that assumed gene expression to be the mediating variable. This result provided first evidence that gene expression acts as potential partial mediator between the germline variant and mutational signature. The partial character of the mediating effect was expected since mediating effects are rarely - if ever - fully mediating (Tingley et al., 2014).

Next, we subjected this model to the *mediate* function of the R package mediation (Tingley et al., 2014) to assess significance of the average causal mediation effect and estimate its proportion by nonparametric bootstrapping (n=1,000 simulations). Briefly, the average causal mediation effect is the difference between total effect *t* and direct effect *c* which equals to the indirect effect defined by the product *a x b*. Further, by bootstrapping, *i.e.*, repeated random

sampling of data points with replacement and computation of the desired statistic on the randomly sampled data points, an approximate distribution of the average causal mediation effect can be obtained to determine its significance (Preacher & Hayes, 2004).

Our mediation analysis revealed a significant average causal mediation effect ($P \leq 2x10^{-16}$) and estimated that a remarkable fraction of 87.11% of the effect of the germline variant rs12628403 conferred to the mutational signature by *APOBEC3B* expression. We were therefore able to state with high confidence that the effect of the *APOBEC3B* eQTL modulates Signature 2 accumulation via *APOBEC3B* expression, *i.e*., that in this particular case, gene expression acts a molecular regulator of a mutational signature (PCAWG Transcriptome Core Group et al., 2018).

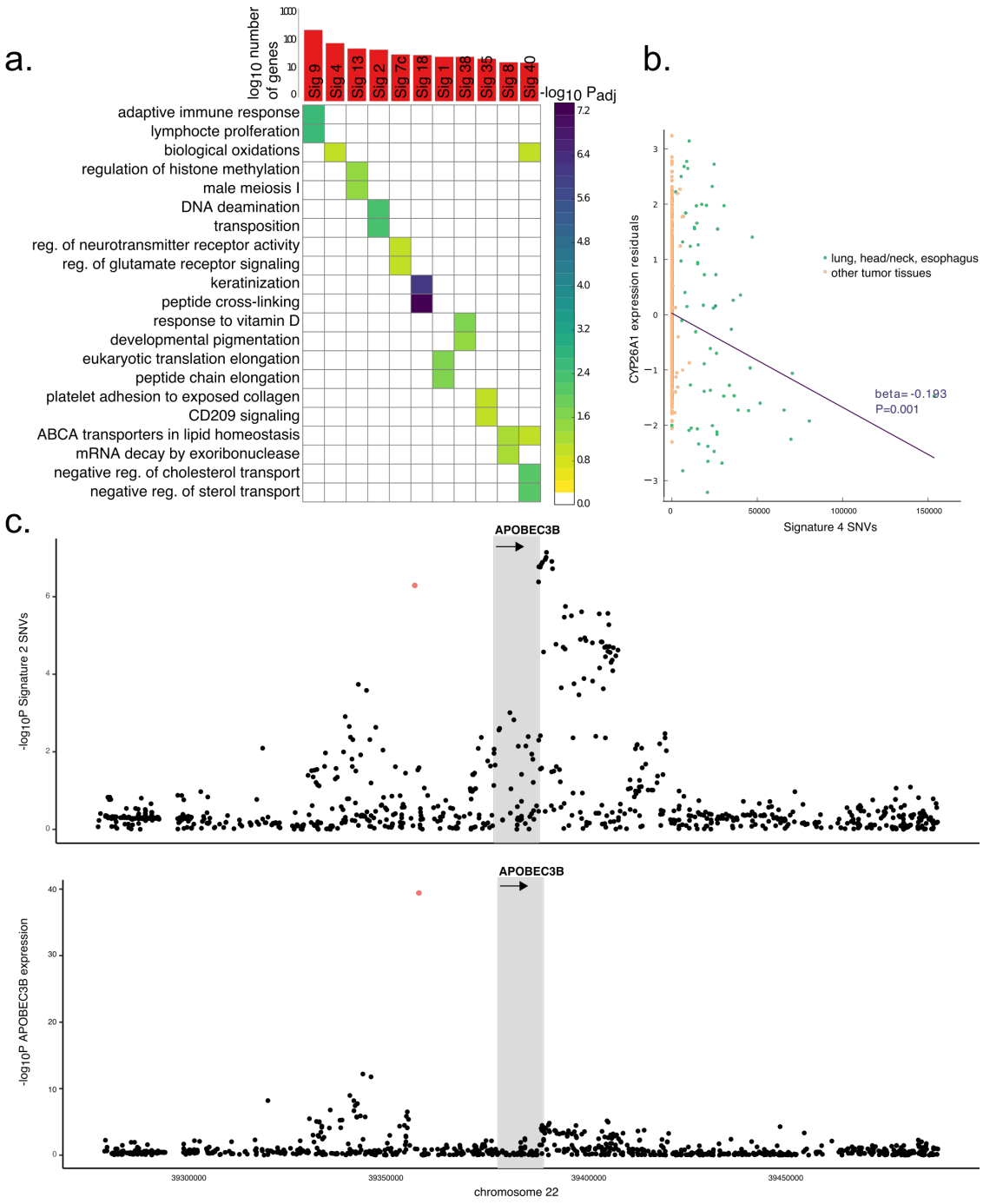

Figure 5. See next page for figure caption.

Figure 5. Associations between mutational signatures, gene expression and germline variation. **a.** Summary of significant mutational signature-gene expression associations. Top panel: Barplot of total number of associated genes per signature (FDR ≤ 10%). Bottom panel: Heatmap of enriched Gene Ontology and Reactome Pathways categories for genes associated with each signature ($-\log_{10} P_{adj}$ of enrichments; associations with FDR > 10% are coloured in white). **b.** Representative signature-gene association, depicting a negative association between *CYP26A1* expression and Signature 4. **c.** Manhattan plots of associations between *cis* germline variants proximal to *APOBEC3B* (plus or minus 100k base pairs from the gene boundaries) and Signature 2 (top panel) and *APOBEC3B* gene expression level (bottom panel), respectively. The grey region denotes the gene body, the orange dot is the eQTL lead variant rs12628403.

2.4 Wide-spread associations in somatic eQTL mapping

While the mutational signature analyses revealed links between gene expression and somatic mutations, they were based on genome-wide accumulated somatic SNVs and could therefore not capture the effect of local somatic mutations on close-by genes. These local effects might, however, be important for gene expression alterations and hence for larger-scale phenotypic consequences like cancer progression and metastasis if the gene in question modulates these phenotypes. Therefore, as a complementary strategy, we considered the effect of local somatic variation (*cis*) on gene expression. To mitigate the low allele frequency of somatic SNVs, we aggregated somatic variants into local burdens in genic and non-genic regions, to then follow an approach similar to established germline eQTL analyses.

2.4.1 Quantification of somatic mutation burden

We used the set of high-quality consensus somatic calls provided by the PCAWG Technical Working group based on several core caller pipelines and MuSE (Campbell et al., 2017; Fan et al., 2016). On average, each patient harboured 22,144 somatic SNVs across the whole genome. Due to this low frequency of somatic SNVs and the resulting low recurrence of specific SNVs across the cohort, we aggregated the variants into somatic burdens of different genomic regions defined by gene annotations according to Gencode (version 19, Harrow et al., 2006). To map local effects of somatic mutations within the gene body, we generated a set of disjoint gene exons (*cis* exonic regions) by collapsing overlapping exon annotations into single collapsed exonic regions using bedtools (Quinlan & Hall, 2010). The set of disjoint introns was generated using bedtools by subtracting the collapsed exonic regions from the gene regions (*cis* intronic regions). To map local effects of somatic mutations in flanking features outside the gene body,

we binned the surrounding regions (±1M base pairs from the gene boundaries) into 2k base pairs windows overlapping by 1k base pairs ('*cis* flanking regions').

We considered alternative strategies for aggregating somatic SNVs in burden regions: (i) a binary value that indicates presence or absence of SNVs, (ii) the aggregated burden as sum of SNVs, or as (iii) weighted burden, *i.e.,* sum of variant allele frequencies (VAFs) of the SNVs. VAF is an important concept in cancer genomics since it can be used to estimate the fraction of a tumour that contains certain mutations. As a concept of intra-tumour heterogeneity, VAFs can be used to reconstruct the cancer evolutionary tree and tumour subpopulations (Hajirasouliha et al., 2014). While all burden definitions (i, ii, iii) resulted in calibrated association analyses, the latter (from now on referred to as weighted burden) had most power in mapping somatic eQTL and was therefore used for downstream analyses (see PCAWG Transcriptome Core Group et al. (2018) for details). This somatic burden was standardised across patients (to mean zero and standard deviation one).

2.4.2 Variance decomposition of gene expression

We firstly performed variance decomposition of the gene expression data in LIMIX. Specifically, we used variance decomposition to quantify the variability of gene expression explained by local and distal germline and somatic genetic variation. To do so, we modelled all genetic effects as random effects in a linear mixed model, accounting for known confounding factors as fixed effect covariates as in the mutational signature analysis (Section 2.3.1) and in addition for local somatic burden per gene. The following different measurements of germline variants and somatic burden were hereby modelled as genetic factors in the variance decomposition model: *cis* somatic intronic variants (weighted burden in introns), *cis* somatic exonic variants (weighted burden in exons), *cis* somatic flanking variants (weighted burden in regions of 2k base pairs overlapping by 1k base pairs within 1M base pairs from gene boundaries), somatic intergenic variants (weighted burden in regions of 2k base pairs overlapping by 1k base pairs outside the window of 1M base pairs), *cis* germline variants (germline variants within 100k base pairs from gene boundaries), genome-wide germline variants (*i.e.*, genome-wide population structure), and local SCNAs. For each of these random effects, the data was mean-centered and standardised, and a linear kernel was computed and used as covariance matrix. The random effect model was fit to decompose gene expression variation across individual genes into the defined genetic components. The resulting variance components were normalized to a sum of one.

This analysis identified SCNAs as the major driver of expression variation (27.3% on average, **Figure 6a**), followed by flanking somatic and germline variants. Notably, *cis* germline effects, although exhibiting smaller effects on individual genes, explained the largest proportion of variance for 11,905 genes, compared to 3,568 genes, for which somatic factors explained most variation (PCAWG Transcriptome Core Group et al., 2018).

.

2.4.3 Cis somatic eQTL mapping

We then tested for associations between the somatic burden and gene expression levels. Specifically, we associated the weighted somatic burden (Section 2.4.1) with gene expression (of 18,898 protein-coding genes corrected for 35 peer factors, Section 2.2) by employing linear models in LIMIX (Lippert et al., 2014). We accounted for the same confounding factors as fixed effects as in the variance decomposition analysis (Section 2.4.2). This analysis was conducted on all 1,188 patients, and on a subset of 899 carcinoma patients to replicate the analysis on a more homogeneous set of tumours. Only the somatic burden of recurrently mutated genomic intervals, *i.e*., with a burden frequency ≥ 1% across all patients (presence of burden in at least 12 patients in the full cohort and 9 patients in the carcinoma cohort), was taken into account. Altogether, 18,708 of the genes had at least one mutated interval at that frequency. Each of these genes was tested for association with the weighted burden of all *cis* burden regions, *i.e*., *cis* exonic, *cis* intronic and *cis* flanking regions. After the analysis, Bonferroni correction was applied to correct for multiple *cis* windows tested within the same gene. Then, Benjamini-Hochberg correction was applied to adjust the P-values of the lead genomic regions across genes.

Genome-wide, this identified 649 somatic eQTL (FDR ≤ 5%) that were associated with the somatic burden of 567 unique regions. Among these, only a few (n=11) somatic eQTL were explained by somatic burden in exonic and intronic regions. These included genes with known roles in the pathogenesis of certain cancers, *e.g*., *CDK12* in ovarian cancer (Bajrami et al., 2014), *PI4KA* in hepatocellular carcinoma (Ilboudo et al., 2014), *C11orf73* in clear cell renal cancer (Bhalla et al., 2017) and *BCL2* and *SGK1* in lymphoma (Hartmann et al., 2016).

In general, the majority of associated genes (68.4%) were associated with flanking non-coding intervals (272 intergenic, 172 intronic regions of different genes). In contrast to known germline eQTL, the detected somatic burden-gene expression associations tended to be located distal to the transcription start site (TSS) of the respective gene (88% at ≥ 20k base pairs from the TSS), with on average larger effects of distal associations than associations proximal to the TSS (|β|=3.3 for distal associations *versus* |β|=1.4 for proximal associations). This finding points towards relevance of somatic mutations at distal regulatory elements that have probably been underestimated in the analysis of somatic regulatory networks in human cancers so far (**Figure 6b**).

We discovered that so-called cancer-testis genes were marginally more frequent among genes with somatic eQTL than expected (45/982, P=0.06, Fisher's exact test). Cancer-testis genes are of interest for their known immunogenic properties (Scanlan et al., 2002) and exhibit high expression in sperm and some cancers but are repressed in healthy tissues (Simpson et al., 2005).

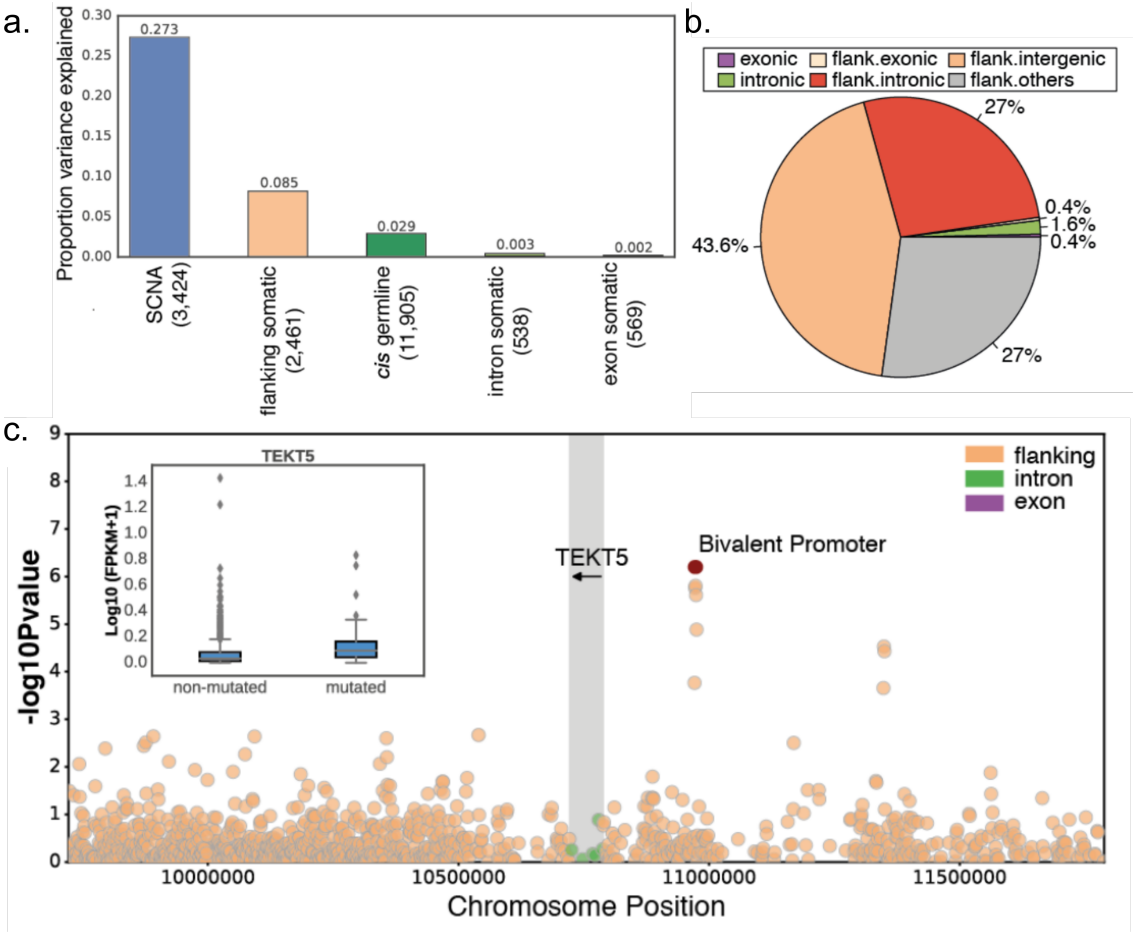

Figure 6. Somatic eQTL analysis. **a.** Variance composition of gene expression. Average proportion of variance (y-axis) explained by different genetic factors (x-axis: SCNAs, somatic variants in flanking regions, germline variants, intronic somatic and exonic somatic variants). The numbers in parentheses behind the genetic factors indicate the number of genes for which the respective factor is the largest variance component, respectively. **b.** Breakdown of 567 genomic regions that underlie the observed *cis* somatic eQTL by genomic region (flank. = 2k base pairs flanking region within 1M base pairs from the; flank.intergenic = flanking region in a genomic location without gene annotations; flank.intronic = flanking region overlapping an intron of a nearby gene; flank.others = flanking region partially overlapping exonic and intronic annotations of a nearby gene). **c.** Manhattan plot of nominal P-values of associations between *TEKT5* expression (gene body highlighted in grey) and the somatic burden of flanking, intronic and exonic genomic regions. The somatic burden of the leading genomic region is associated with increased *TEKT5* expression (inlet boxplot: $\log_{10}$FPKM+1 over binarised somatic burden of the genomic region; $P=2x10^{-6}$, $\beta=0.221$). The genomic region overlaps with an upstream bivalent promoter (red dot).

We further observed an enrichment of somatic eQTL in bivalent promoters for cancer-testis genes (P=0.04, Fisher's exact test). Re-activation of poised promoters is known to be one mechanism of upregulation of developmental genes - including cancer-testis genes - in cancer (Bernhart et al., 2016), and our findings might point towards a re-activation by somatic mutation. One example is *TEKT5,* an integral component of sperm, that has been found to be aberrantly expressed in a variety of cancers (Hanafusa et al., 2012). We observed a positive association between *TEKT5* expression and somatic mutational burden in a bivalent promoter site close to the 5' end of the gene (**Figure 6c**; PCAWG Transcriptome Core Group et al., 2018).

2.5 Cancer-specific deregulation of allele-specific expression

While the somatic eQTL analysis establishes associations between proximal (likely *cis* acting) genetic variants on gene expression, allele-specific effects allow for directly investigating haplotype-specific deregulation. We therefore extended the QTL-based somatic analysis to study the effect of somatic mutations in various genomic contexts on ASE, which provides an internally controlled readout that enables assessing differential regulation between haplotypes in the same patient (Korir & Seoighe, 2014).

2.5.1 Allelic expression quantification

To enable an allele-specific analysis, we firstly obtained phased germline and somatic variants from the PCAWG working group 11. Briefly, for assembling phased germline genotypes, the PCAWG working group 11 applied IMPUTE2 to the Sanger 1000G output for phasing of heterozygous germline variants (Howie et al., 2012). The IMPUTE2 output was corrected using results from the Battenberg CN calling algorithm (Davies et al., 2017), ascertaining that no haplotype switches occurred within regions of consecutive CN gain (Nik-Zainal et al., 2012). The resulting phased germline variants were arranged such that haplotype 1 always corresponded to the amplified alleles in regions with SCNAs. In cases where both co-occurred on the same read (10M variants, *i.e.*, ~20% of all SNVs), individual somatic variants were phased to the nearest germline heterozygous site. We only considered SNVs that were phased to the respective germline variant by at least three reads (resulting in ~6M variants).

We quantified ASE in tumour and the corresponding normal tissue (available for 150 patients across 13 cancer types) based on heterozygous germline variants in exonic regions, using the GATK ASEReadCounter algorithms for counting ASE reads (Castel et al., 2015). We considered RNA-Seq reads with a minimum mapping quality of 20 and a minimum base quality of ten. Only heterozygous variants with a minimum coverage of eight RNA reads were

considered for further analyses. The raw ASE read counts were then post-processed as follows: ASE sites were converted to BED files and aligned against the ENCODE 50mer mappability track to extract mappability scores for all sites. All sites with mappability scores unequal to one were removed. All sites with allelic read counts less or equal to one were removed to prevent genotyping error to influence ASE quantification. To maximise detection power, we then aggregated ASE counts across heterozygous sites within genes to gene-level ASE by leveraging phased germline variant maps. First, gene mapping was performed against ENSEMBL release 75 using the pyEnsembl library (Hubbard et al., 2002; Zerbino et al., 2018). Second, for each gene, we summed up the read counts on the respective haplotypes over all gene-specific ASE sites to gene-level haplotype-specific read counts. To allow for a robust assessment of gene-level ASE we only considered genes with at least 15 reads total, yielding 4,379,378 gene-patient pairs across 1,120 patients and 17,009 unique genes across 12,441,502 ASE sites. Per patient, this allowed us to quantify between 588 and 7,728 genes for both haplotypes (median=4,112).

We then identified potential functional somatic and germline variants on each of the haplotypes that could be tested for *cis* regulatory associations. Due to the low frequency of somatic SNVs and the resulting low recurrence of specific SNVs across the cohort, we aggregated all phased SNVs into functional categories based on their genomic regions (see somatic eQTL analysis (Section 2.4.1)). Using the Variant Effect Predictor (VEP) tool (McLaren et al., 2016), we defined regulatory regions of each gene as functional categories (flanking (upstream, downstream), promoter, 5' UTR, exon, intron, 3' UTR). We additionally classified exonic SNVs as synonymous, missense, or stop gain variants, depending on their functional annotation by VEP. We integrated 'splice donor' and 'splice acceptor' variants into the general 'splice region' variant category and mapped 'stop retained' variants to the 'synonymous' variant category. We then mapped each SNV to the respective regulatory region of its nearest gene. Hereby, we only mapped SNVs that were located in a *cis* window of 100k base pairs from the boundaries of a gene; the remaining SNVs were discarded. We further divided the flanking (upstream and downstream) categories into disjoint categories using windows of the size of 10k base pairs from 10 to 100k base pairs from the gene boundaries. We used the relationship between the VAF of SNVs and SCNAs at the same genetic locus to determine whether SNVs occurred before (‘early’) or after (‘late’) the corresponding SCNA. As outlined in Section 2.4.1, VAFs can be used to estimate the fraction of a tumour that contains certain mutations, and can therefore be used to reconstruct the cancer evolutionary tree (Hajirasouliha et al., 2014). For final quantification of SNVs per functional category, we therefore computed a weighted *cis* mutational burden per functional category by aggregating SNVs to a total localised burden weighted by the VAF of each SNV (see somatic eQTL analysis (Section 2.4.1)).

Finally, we calculated allelic expression imbalance (AEI) for each gene using a binomial test against the expected CN ratio modified by tumour purity. The CN-corrected P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure. Significant AEI was called at FDR ≤ 5%. Importantly, this definition of AEI notably prevents the assessment of the

effect of copy neutral loss of heterozygosity that has been observed in various cancer types (Maciejewsky & Mufti, 2008) and should therefore be studied in future analyses. We found substantial differences in the fraction of genes with AEI between cancer types (median percentages between 14.2% in prostate adenocarcinoma and 46.8% in squamous cell carcinoma of the lung; $P=2.2x10^{-13}$, Mann-Whitney-U-Test; **Figure 7a**), and between AEI in cancer and the corresponding normal tissue. Cancers with extensive chromosomal rearrangements, including lung, breast and ovarian cancers, were associated with most frequent AEI events (**Figure 7a**), which is consistent with previous reports that have implicated SCNAs in allelic deregulation in cancers (Ha et al., 2012; PCAWG Transcriptome Core Group et al., 2018).

2.5.2 Decomposition of allele-specific expression determinants

Motivated by the substantial difference in the fraction of genes with AEI between cancer types and between cancer and corresponding normal tissues (Section 2.5.1), we considered AEI to robustly identify the genetic elements that contribute to somatic deregulation of gene expression. We employed a generalised linear model to identify the determinants of AEI, accounting for known imprinting status, the germline eQTL genotype, local allele-specific SCNAs and the weighted mutational burden of proximal somatic SNVs stratified into functional categories (upstream, downstream, promoter, 5'UTR, intron, synonymous, missense, stop gain, 3'UTR; Section 2.5.1). We accounted for genomic imprinting since it results in parental-dependent gene expression; for doing so, we used a census of known imprinted genes collated by Morison et al. (2005). We additionally corrected for sample purity, local CN ratio, sex, cancer type, total somatic burden, gene length, length of the canonical transcript, the number of accessible ASE sites per gene and both, gene-level and sample-level read depth. Briefly, gene-level CN ratio was obtained by averaging haplotype-specific CN states to mean haplotype-specific CN ratio per gene and computing the major over total ratio of those averages.

In aggregate, SCNAs accounted for 84.3% of the total explained variation, confirming our findings of the somatic eQTL analysis, followed by germline eQTL lead variants (9.1%), somatic SNVs (4.9% as sum of non-coding and coding SNVs) and imprinting status (1.7%; **Figure 7b**). While cumulatively, non-coding variants were more relevant than coding variants, somatic protein truncating variants ('stop gain' variants), which trigger nonsense-mediated decay, were the most predictive individually (**Figure 7c**). This observation confirms the importance of nonsense-mediated decay in cancer gene regulation (Lindeboom et al., 2016). SNVs within splice regions, 5' UTR and promoters were also strongly associated with AEI presence and we observed a global trend of decreasing relevance of variants with increasing distance from the TSS (**Figure 7c**).

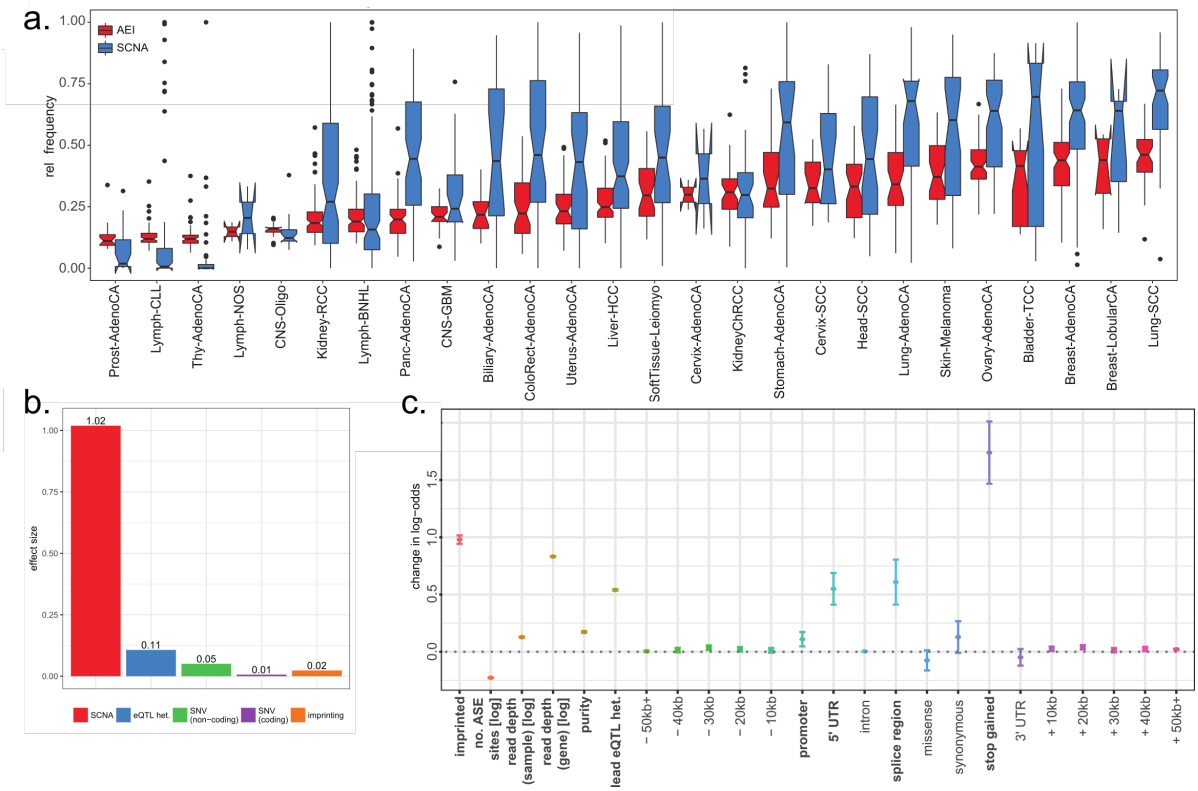

Figure 7. ASE analysis. **a.** Distribution of the proportion of genes with significant AEI (FDR ≤ 5%) (red) and SCNAs (blue) across patients for different cancer types. Cancer types with prevalent chromosomal instability (frequent SCNAs) exhibit most frequent AEI. **b.** Standardised effect sizes of a generalized linear model modelling the presence of AEI, taking SCNAs, germline eQTL, coding and non-coding mutations and imprinting into account. **c.** Relative contribution of different types of somatic mutational burden and other covariates to the likelihood of observing AEI. Significant covariates (FDR ≤ 5%) are highlighted in bold.

Our model allowed for attributing AEI to germline SNPs, SCNAs and somatic SNVs by computing average scores derived from predicting AEI individually from SNPs, SCNAs and SNVs. Using the trained model for AEI, we hence set out to characterise sets of genes with strong allelic deregulation that can be attributed to these different genetic factors. We ranked genes according to average scores across the cohort, based on (i) the predicted AEI from the germline component and (ii) the predicted AEI from somatic components (SCNAs and SNVs). For comparison, we also considered (iii) the empirical AEI frequency in the cohort, and (iv) the burden of loss-of-function and gain-of-function mutations derived from genetic data only. When overlaying these ranked lists with known cancer genes (COSMIC catalogue; Forbes et al. (2017)), we found cancer genes to be enriched among genes with high somatic AEI score (P ≤ 0.005, Gene Set Enrichment Analysis; Subramanian et al., 2005), but we observed no

enrichment amongst genes with recurrent AEI (P=0.99). As expected, genes with AEI due to germline eQTL were depleted for cancer genes (negative enrichment, P ≤ 0.001). Consistent with the traditional definition of cancer genes based on recurrent mutations (Forbes et al., 2006), genes with recurrent loss/gain-of-function mutations were most strikingly enriched for the COSMIC genes (P ≤ 0.001). In addition, the top 10% of genes of the somatic AEI scores were enriched for Gene Ontology categories with relevance to cancer, including chemotaxis, cell motility, locomotion and cell migration, which notably were absent when considering loss/gain-of-function mutations. These results suggest that somatic AEI could be used to prioritise regulatory variants to identify genes with roles in cancers, which extends previous observations in a single cancer (Ongen et al., 2014) to a pan-cancer setting (PCAWG Transcriptome Core Group et al., 2018).

2.5.3 Assessment of allele-specific expression of potential cancer genes

As we found cancer genes to be enriched among genes with high somatic AEI (Section 2.5.2), we specifically investigated genes that were primarily deregulated by SNVs or SCNAs. Of the 4,007 genes in the upper quartile based on the prevalence of overall AEI across all tumours, 1,843 genes exhibited SNV-linked AEI. When we ranked these genes based on the predicted AEI from SNVs, the top 10 genes were *FBXO5*, *ASPM*, *PSCA*, *CDKN1A*, *KIF20B*, *TP53* and *CLDN4,* which have previously been linked to cancer (*e.g.*, Wang et al., 2014) but also *SYNE1* and *EXO1*, genes that have not been prominently linked to cancer.

SYNE1 controls nuclear polarity and spindle orientation, and acts upstream of NOTCH signalling in squamous lineage development (Lasorella et al., 2014). To understand the effects of the regulatory SNVs in detail, we applied a Combined Annotation Dependent Depletion (CADD) analysis according to Kircher et al. (2014), which scores deleteriousness of SNVs based on multiple annotations like conservation and functional effect. Notably, three melanoma cases in our cohort preferentially expressed deleterious missense mutations in *SYNE1* (elevated CADD scores > 25), likely leading to a relative decrease of gene expression in these tumours. Based on the GEPIA (Gene Expression Profiling Interactive Analysis) webserver (Tang et al., 2017), we also found that low *SYNE1* expression was associated with worse overall survival in TCGA melanoma patients (log-rank P=0.002, hazard ratio=0.57), providing further support for its relevance in disease.

EXO1 is known to be involved in MMR and recombination, and exhibited significant AEI for a deleterious missense mutation in a colorectal adenocarcinoma patient (CADD score 34). Similarly, TCGA colorectal adenocarcinoma patients with lower expression of *EXO1* showed worse overall survival (log-rank P=0.022, hazard ratio=0.57), implicating *EXO1* as a potential tumour suppressor in human colorectal cancer. Consistent with this finding, *EXO1* knockout

mice exhibited defects in DNA damage response and increased carcinogenesis (Schaetzlein et al., 2013; PCAWG Transcriptome Core Group et al., 2018).

Notably, we found cancer-testis genes (see Section 2.4.3) to be depleted when considering the full somatic score including SCNAs (25/476 cancer-testis genes in the top 10% of genes, 48 expected, Chi-squared test, $P=6x10^{-4}$), but enriched in the AEI score based on SNVs only (66/476 cancer-testis genes in the top 10% of genes, 48 expected, Chi-squared test, $P=6x10^{-3}$). One potential explanation is that repressed cancer-testis genes have to undergo somatic re-activation by SNVs before CN amplification. To elucidate this, we used the stratification of SNVs into the categories 'early' and 'late' (Section 2.5.1) and found strong over-representation of 'early' SNVs in 329 out of 7,525 cancer testis gene-patient pairs (216 expected, Chi-squared test, $P=4x10^{-14}$), what confirmed our hypothesis.

In summary, we identified somatic and germline variation associated with allele-specific deregulation of genes across cancer types. We were able to demonstrate the power of ASE as an integrator of different sources of transcriptional deregulation in *cis*. Further, genes that showed ASE as a result of somatic mutagenic processes were enriched for both, known cancer genes and cell migration pathways, suggesting the utility of ASE analyses for the identification of candidate driver genes involved in metastasising processes. In particular, we found recurrent somatic reactivation of cancer testis genes, warranting further investigation into their role in carcinogenesis and tumour progression (PCAWG Transcriptome Core Group et al., 2018).

2.6 Summary and discussion

The regulatory landscape of cancer is highly heterogeneous, cancer type specific and influenced by the germline background. The prevalence of perturbations due to somatic mutations makes cancer a perfect model for studying regulatory dependencies, including regulatory dependencies of genes that are not subject to germline perturbations. Collectively, our analyses of the PCAWG genomic and transcriptomic data provided a comprehensive picture of this regulatory landscape, and of how different germline and somatic variations alter gene expression levels in a pan-cancer setting.

First, we studied associations between gene expression association and mutational signatures. As mutational signatures capture global variations across individuals, they describe variation patterns distinct from local somatic variations. We investigated the utility of associations between mutational signatures and gene expression levels, thereby deriving *de novo* annotations of signatures with previously unknown roles. We produced comprehensive across-tissue germline eQTL maps to carry out analyses that integrate germline and genome-wide somatic variation with gene expression. This allowed us unpick the molecular chain of events for the common APOBEC mutational process and its germline component. The APOBEC signature was however the only mutational process that we were able to associate with germline variation.

This might be due to the tissue specificity of mutational signatures that inherently reduces the effective sample size per signature analysis and might hence render the detection of the effect of (rare) germline variants impossible. In general, the strong tissue specificity of mutational signatures did not allow for an entirely cancer type-independent study of the molecular consequences of the signatures: Although we accounted for differences between cancer types, if in an extreme case a signature only occurs in one cancer type, associations will inherently be confounded with tissue type effects. It will therefore be important to conduct similar analyses for individual cancer types and in cohorts with more WGS samples.

In terms of local somatic variation, we considered the effect of somatic mutational burden in different genomic regions on gene expression changes by building a systematic map of somatic eQTL. We accounted for variation in clonality as well as local hypermutations, thereby identifying likely causal associations between somatic burden and gene expression. However, this approach has limitations and we cannot rule out that a fraction of the associations we identified are due to technical or biological factors that jointly affect gene expression and local mutation rates. We also note that an analysis of cancer type specific somatic eQTL is currently not feasible with the given sample size. However, the somatic eQTL map will be a valuable resource to address a wide range of downstream analyses, providing a comprehensive overview of gene expression determinants in cancer and insights into the underlying biology. The systematic assessment of regulatory non-coding genetic variation significantly improves our understanding of the aetiology and functional implications of intra- and inter-tumour heterogeneity and the selective forces applied to these heterogeneous genomes.

We then used ASE readouts for integrated modelling of local genetic variation to characterise genetic elements that have the largest regulatory effects. We demonstrated the extent to which AEI on the expression level follows allelic imbalance on the genomic level. A problem was that the considered phased somatic mutation set was based on read phasing, which has only been possible for around 20% of all SNVs and hence substantially reduced the size of our available dataset. Further, ASE readouts can only be derived in cases with at least one heterozygous germline variant in the gene in question, what reduced overall sample size once more and additionally hindered gene-level associations.

Altogether, we showed that co-regulation of the same genes by multiple different types of variants is common in cancer, and we simultaneously reported the relative magnitudes of these different effects. Previous studies have been limited by the lack of WGS data, which is essential for identifying contributions of non-coding variants to gene expression variability. Indeed, our analyses of the currently largest cohort of matched tumour WGS and RNA-Seq data of 1,188 patients identified previously underappreciated associations between somatic regulation in distal regulatory elements and gene expression, as well as *de novo* functional annotations of genome-wide mutational signatures. Finally, our results pointed towards non-coding somatic deregulation as an important functional driver of carcinogenesis beyond mutations of the coding sequence (PCAWG Transcriptome Core Group et al., 2018).

3 Combined single-cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity

Contributions

Together with Marc-Jan Bonder and Oliver Stegle, I supervised Bachelor student Stephanie Linker in this project. While I supervised Stephanie in analysing splicing ratios and applying deep learning modelling, my own work focused on understanding splicing variation across cells and on splicing category switches during stem cell differentiation. All work has been realised in close collaboration with Stephanie Linker, who pre-processed the majority of the data, and Marc-Jan Bonder, who pre-processed the methylation data and co-supervised the project. Davis McCarthy was involved in gene expression data processing and cell-level feature calculations.

Mariya Chhatriwala, Stephen Clark, Shradha Amatya, Ludovic Vallier, and Wolf Reik generated the scM&T-Seq data.

All scripts are available as jupyter notebooks here: https://github.com/PMBio/scmt_splicing. The manuscript was published in Genome Biology (Linker, Urban et al., 2019). The deep learning component of this project was published as part of the kipoi project (Avsec et al., 2019) in Nature Biotechnology.

3.1 Introduction

RNA regulation is not limited to changes in total gene expression levels, and in particular post-transcriptional changes affect the pool of gene products and proteins in the cell. One such modification is alternative splicing of RNA, which enables efficient gene encoding and contributes to gene expression variation by alternative exon usage (Black, 2003; Section 1.1.2). Alternative splicing is pervasive and affects more than 95% of human genes (Barbosa-Morais et al., 2012). Splicing is known to be regulated in a tissue-specific manner (Kelemen et al., 2013; Revil et al., 2010) and individual splicing events have been implicated in human diseases (Xiong et al., 2015). Bulk RNA-Seq data of cell populations has been used to identify and quantify different splicing events (Sammeth et al., 2008). These studies mostly focused on exon skipping at cassette exons, the most frequent alternative splicing event (Black, 2003; Section 1.1.2).

Different genetic factors have been linked to splicing of cassette exons, including sequence conservation (Wainberg et al., 2016) and genomic features such as local sequence composition and length of the exon and of the flanking introns (Huang & Sanguinetti, 2017; Xiong et al., 2015). Although there is some evidence for an epigenetic component of splicing regulation, this relationship is not fully understood and alternative models for the role of DNA methylation in splicing have been proposed (Maunakea et al., 2013; Shukla et al., 2011b; Yearim et al., 2015). The transcriptional repressor CTCF has been shown to slow down RNA polymerase II, resulting in increased exon inclusion rates. By inhibiting CTCF binding, DNA methylation can cause reduced exon inclusion rate (Shukla et al., 2011). Alternatively, increased DNA methylation of the MeCP2 pathway has been associated with increased exon inclusion rates. MeCP2 recruits histone deacetylases in methylated contexts; histone deacetylases are enzymes that remove acteyl groups from histones and therefore allow the histones to wrap the DNA more tightly. This interplay between MeCP2 and DNA methylation slows down RNA polymerase II and leads to an increased exon inclusion rate (Maunakea et al., 2013). Finally, the heterochromatin protein HP1, which serves as an adapter between DNA methylation and TFs, increases the exon inclusion rate if it is bound upstream of the alternative exon (Yearim et al., 2015). This might be due to heterochromatin formation that again slows down RNA polymerase II. However, when HP1 binds to alternative exons, this results in increased exon skipping (Yearim et al., 2015). These alternative mechanisms point towards complex regulation of alternative splicing via combined effects of DNA sequence and DNA methylation, both proximal as well as distal to the alternative exon. These links are, however, difficult to study since this would require a model system with constant genetic background but epigenetic variation, which is hard to realise experimentally.

While epigenetic variation can be found between different tissue types, the intrinsic variation between individual cells has not been exploited yet. However, single cells with the same genetic background constitute the perfect study system that enables analysing molecular single-cell variability as perturbation at the epigenetic level. While technological advances to perform RNA-Seq in single cells have recently enabled analyses of splicing variation at single-cell resolution

(Huang & Sanguinetti, 2017; Song et al., 2017; Welch et al., 2016), no studies have linked the epigenome to the transcriptome yet. Here, we leveraged technological advances that enable joint profiling of gene expression and DNA methylation in the same cell (single-cell methylation and transcription sequencing (scM&T-Seq) by Angermueller et al. (2016)). Considering induced pluripotent stem (iPS) cells and differentiated endoderm cells, we studied associations between single-cell splicing variation and DNA methylation while taking into account the effect of local genetic variation across genomic splice sites (Linker, Urban et al., 2019).

We considered iPS and endoderm cells since alternative splicing has been shown to change substantially between pluripotent and differentiated cells, and to play an important role in maintaining pluripotent homeostasis (He et al., 2015). Importantly, iPS cells resemble embryonic stem (ES) cells in terms of their molecular states including alternative splicing patterns (Chen et al., 2015). They can be obtained by reprogramming differentiated cells to stem cells: While the pluripotency of cells depends on a vast network of genes and signalling molecules, the TFs OCT4, SOX2, and NANOG are known to control majority of these processes and modulation of their activity has allowed reprogramming of various cell types across species (Boyer et al., 2005). Similar to ES cells, iPS cells can then be differentiated into derivatives of any of the three primary germ layers (Medvedev et al., 2010). iPS cells therefore allow the study of molecular changes during the development of organs and whole organisms (Yamanaka, 2012), and will potentially play an important role in cell therapy of human diseases (Medvedev et al., 2010). A thorough understanding of the molecular mechanisms of pluripotency and differentiation is hereby essential for any clinical application of iPS cells (Chen et al., 2015).

3.2 The determinants of single-cell splicing variation

3.2.1 Single-cell splicing variation during differentiation

We applied parallel scM&T-Seq to differentiating iPS cells from a single donor ('joxm_1') of the Human Induced Pluripotent Stem Cell Initiative (HipSci) (Kilpinen et al., 2017; Streeter et al., 2017). We profiled 93 cells in the iPS state, as well as following three days of differentiation towards definitive endoderm. After quality control (see Section 3.5 for details) this resulted in 86 iPS and 59 endoderm cells, which were used for further analysis. In each cell we quantified cassette exon inclusion rates (scheme of a cassette exon with the central alternative exon 'A' in **Figure 8a**). The alternative splicing rate (PSI) was estimated as the fraction of reads that include the alternative exon *versus* the total number of reads of the cassette exon. We detected and quantified splicing for between 1,386 and 14,434 exons per cell (minimum coverage of five reads; see Section 3.5 for Methods, and Linker, Urban et al. (2019) Table S1/Additional File 1 for read distributions of the transcriptome and methylation data across cells). We considered 6,282 (iPS) and 4,096 (endoderm) cassette exons that were quantified by at least five reads in at least ten cells.

Initially, we investigated to what extent single-cell data is consistent with existing models of splicing variation. One canonical splicing model assumes that individual cells express only a single splice isoform ('cell model'), whereas the other existing splicing model postulates the presence of multiple isoforms in a given cell ('gene model'; **Figure 8b**). The question which splicing model describes splicing variation across cells best can not be modelled with bulk data (Faigenbloom et al., 2015; Shalek et al., 2013), and no study has tried to answer this question on a single-cell level yet. We therefore modelled the expected distribution of PSI assuming either the cell or the gene model. For the cell model, we modelled PSI variation as a bimodal distribution at the single cell level. In the case of the gene model, the limited number of transcripts per gene had to be taken into account. Assuming that each time a gene is transcribed, the probability of exon inclusion equals mean PSI; however, the limited number of transcripts leads to fluctuations in the actually observed PSI and the binomial distribution of isoforms is therefore restrained by an upper boundary of the standard deviation. To obtain this boundary, we calculated this standard deviation across cells by simulating the PSI (n=400 simulation) of each cell as a binomial distribution.

A comparison of mean and standard deviation of PSI in our data with these expected distributions ruled out the cell model, but we also observed deviations from the gene model, in particular for exons with intermediate levels of splicing ($0.2 < PSI < 0.8$; **Figure 8c**). An intermediate model might therefore describe our data best, *e.g.*, a model that assumes differential alternative splicing behaviour between individual chromosomes. We also observed that iPS cells have splicing patterns that are more consistent with the gene model (percentage of cassette exons that follow the gene model) as compared to endoderm cells ($P=8.x10^{-12}$, Mann-Whitney-U-test), suggesting an impact of cell differentiation on the splicing model and a general complex regulation of alternative splicing across cells (Linker, Urban et al., 2019).

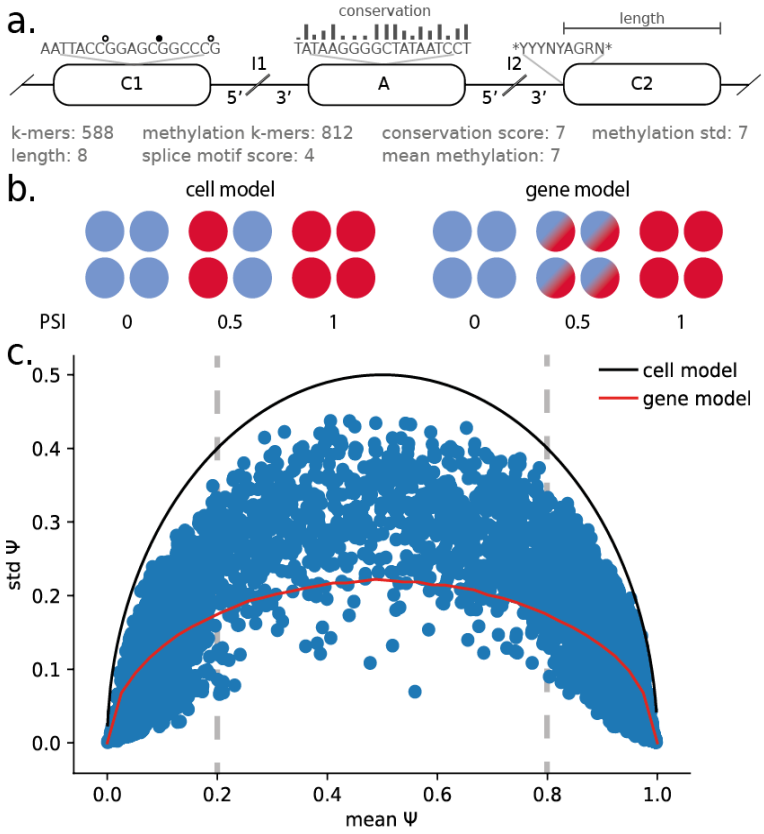

Figure 8. Modelling of alternative splicing in single cells. **a.** Illustration of the sequence contexts per cassette exon (top) and the total numbers of genomic and epigenetic features extracted per sequence context (bottom). 'A' denotes the alternative exon, 'I1' and 'I2' correspond to the upstream and downstream flanking introns and 'C1' and 'C2' to the upstream and downstream flanking exons, respectively. Hereby, the 5' and 3' ends of the flanking introns are considered separately. **b.** Illustration of two canonical splicing models. The 'cell model' assumes that splicing variation is due to differential splicing between cells, with each cell expressing one of two splice isoforms. The 'gene model' corresponds to the assumption that both splice isoforms are expressed in the same cells. **c.** Observed cell-to-cell variation in PSI across cells (standard deviation) as a function of the average PSI across cassette exons in 86 iPS cells. Solid lines model the expected trend under the assumption of a cell model (black line) or a gene model (red line). Grey vertical lines mark intermediate ranges of splicing (0.2 < PSI < 0.8).

3.2.2 Splicing variability and methylation heterogeneity across single cells

To identify locus-specific correlations between DNA methylation heterogeneity and variation in splicing across cells, we tested for associations between differences in DNA methylation levels and splicing rates across cells. Specifically, for each cassette exon, we tested for associations between splicing and variation in DNA methylation in each of seven sequence contexts (based on Spearman's rank correlation coefficient and probability; see Section 3.5.4 for details): the alternative exon, the 5' and 3' end of the two flanking introns, and the two flanking exons (**Figure 8a**). Genome-wide, this identified 424 cassette exons with methylation-splicing associations in iPS cells (out of 5,280 tested cassette exons, FDR ≤ 5%) and 253 associations in endoderm cells (out of 2,622 tested cassette exons, FDR ≤ 5%; Section 3.5). DNA methylation variation in the upstream alternative exon was most frequently associated with splicing variation (~60% of associations included the alternative exon), with approximately equal numbers of positive and negative associations. In iPS cells, 58% of all correlations were positive, in endoderm cells 55%. We identified positive associations as associations where increased DNA methylation was linked to increased alternative exon inclusion; a negative association meant that increased DNA methylation was linked to decreased exon inclusion. Most associations could be detected in more than one sequence context of a given cassette exon, always with consistent effect directions. Similarly, we observed largely concordant associations across the two cell types in our data (again based on Spearman's rank correlation coefficient and probability; Section 3.5). Among exons that were expressed in both, iPS and endoderm cells (n=3,743 cassette exons), 77% of the associations identified in iPS could be replicated on a nominal level in endoderm cells (P ≤ 5%, with a consistent effect direction), and 89% of the associations identified in endoderm were also observed in iPS cells (P ≤ 5%, with a consistent effect direction). Genes with negative associations between DNA methylation in the three upstream regions and PSI were enriched for HOXA2 transcription factor binding sites (TFBSs; enrichment with G:Profiler (Reimand et al. (2016)) and P-value correction across categories according to Benjamini Hochberg; iPS: 78/118 query genes linked to HOXA2, $P_{adj}=6x10^{-4}$; endoderm: 60/90 query genes linked to HOXA2, $P_{adj}=9x10^{-3}$) (Linker, Urban et al., 2019).

3.2.3 Prediction of splicing at single-cell level

To gain insights into the genetic determinants of splicing, we trained regression models that related genomic to splicing rates in single cells (**Figure 9a**). Briefly, we pooled splicing information from cassette exons across cells and trained separate regression models for iPS and endoderm cells. Initially, we considered 607 features that explain sequence composition (based on k-mers), sequence length and sequence conservation per sequence context of the

respective cassette exon and per cell ('genomic' features). We additionally considered an additional set of up to 826 features derived from DNA methylation, including an extended k-mer alphabet that takes the methylation status into account (that results in a five-letter code which differs between methylated ('M') and un-methylated ('U') cytosines), as well as DNA methylation average and variance (across CpG sites) in each of the seven sequence contexts of the cassette exon per cell. Methylation features were either incorporated on a pseudo-bulk average level ('genomic & mean methylation' features) or individual cell level ('genomic & cell methylation' features; see Section 3.5 for details).

Notably, the model to predict single-cell splicing based on genomic features yielded comparable performance to previous attempts to predict splicing using bulk data (Xiong et al., 2015) and single-cell data (Huang & Sanguinetti, 2017; R^2=0.706 and R^2=0.670 for iPS and endoderm cells, respectively; assessed using ten-fold cross validation; **Figure 9a**). To facilitate the comparison with previous results using bulk RNA-Seq, we also predicted aggregate splicing rates across cells ('pseudo-bulk PSI'; bPSI), which resulted in similar prediction accuracies (R^2=0.747 and R^2=0.732 for iPS and endoderm cells, respectively). The inclusion of DNA methylation features increased the prediction accuracy slightly, with larger gains for cell-matched DNA methylation information (**Figure 9a**, *e.g.*, for iPS cells R^2=0.719 and R^2=0.729 for 'genomic & cell methylation' *versus* 'genomic & mean methylation', respectively). In combination with our previous results regarding the relationship between alternative splicing and DNA methylation, these findings suggest that DNA methylation is most predictive of cell-to-cell variation in splicing at the same locus, whereas genomic features capture variation across different loci. While we found these observations to be consistent across iPS and endoderm cells, we replicated these findings in a second independent scMT-Seq dataset of mouse ES cells (n=80 cells; Angermueller et al., 2016; Section 3.5) where we yielded similar splicing prediction performance that also increased with the inclusion of methylation information.

Next, to explore the relevance of these genomic and methylation features in individual cells, we used the prediction accuracy (Pearson R^2) of single-feature regression models. Consistent with previous bulk studies (Wainberg et al., 2016; Xiong et al., 2015), this identified features derived from the alternative exon and its neighbouring sequence context, *i.e.*, the 3' end of the upstream intron and the 5' end of the downstream intron, as most informative for alternative splicing. Within these sequence contexts, sequence conservation of the alternative exon was most informative (**Figure 9b**). Other high-ranking features included the k-mers CT, CTC and CCT of the alternative exon (**Figure 9b)**, sequence patterns that show close resemblance to CTCF binding motifs. Although CTCF or CTCF-like motifs have previously been implicated in alternative splicing, these identified motifs were located upstream or downstream of the alternative exon and were associated with an increased splicing ratio (Brooks et al., 2011; Shukla et al., 2011b; Section 3.1). Here, we were the first to discover the importance of these motifs within the alternative exon, and their association with a decreased splicing rate. In the case of the previously known phenomenon, CTCF-like motifs are hypothesised to lead to increased exon inclusion because CTCF binding to intronic regions slows down the RNA

Polymerase II (Section 3.1). We might see a different direction in our dataset for multiple reasons; CTCF might not bind to exonic regions as efficiently as to intronic regions, or the RNA Polymerase II might not be slowed down by CTCF bound to the alternative exon, or our observed splicing patterns only occur at the particular cell differentiation stages we have studied. Our dataset does not allow us to confirm the molecular caused of this relationship between splicing and exonic CTCF-like motifs. Our hypotheses could, however, be tested, by ChIP-Seq profiling of the TF CTCF, or by single molecular analysis measuring RNA Polymerase II activity and dynamics (Weiss, 1999). Alternatively, the motifs we discovered might differ from the canonical CTCF binding motifs (in terms of extended motifs or orientation), a possibility that we could not rule out based on our k-mer data and that might explain the reverse effect of our motifs on exon inclusion.

The relevance of the cell-specific features for splicing prediction was markedly consistent across iPS and endoderm cells (average R^2 between weights across all iPS and endoderm cells R^2=0.79). Despite the overall consistency in feature relevance, PCA applied to the feature relevance matrix across all (iPS and endoderm) cells identified subtle differences in feature relevance between the two cell types (**Figure 9c**): The first two principal components (PCs) clearly separate iPS from endoderm cells. These components primarily reflect variation in (un)methylated cytosine-containing k-mer composition of the downstream intron ('I2') (**Figure 9d**). Consistent with this, a single-cell splicing model trained on genomics and methylation data from endoderm cells yielded only moderate prediction accuracy in iPS cells (R^2=0.52), highlighting the relevance of these parameters for splicing models that account for DNA methylation information. Altogether, these findings pointed towards a combination of DNA methylation and sequence composition as main determinant of cell type-specific splicing regulation (Linker, Urban et al., 2019).

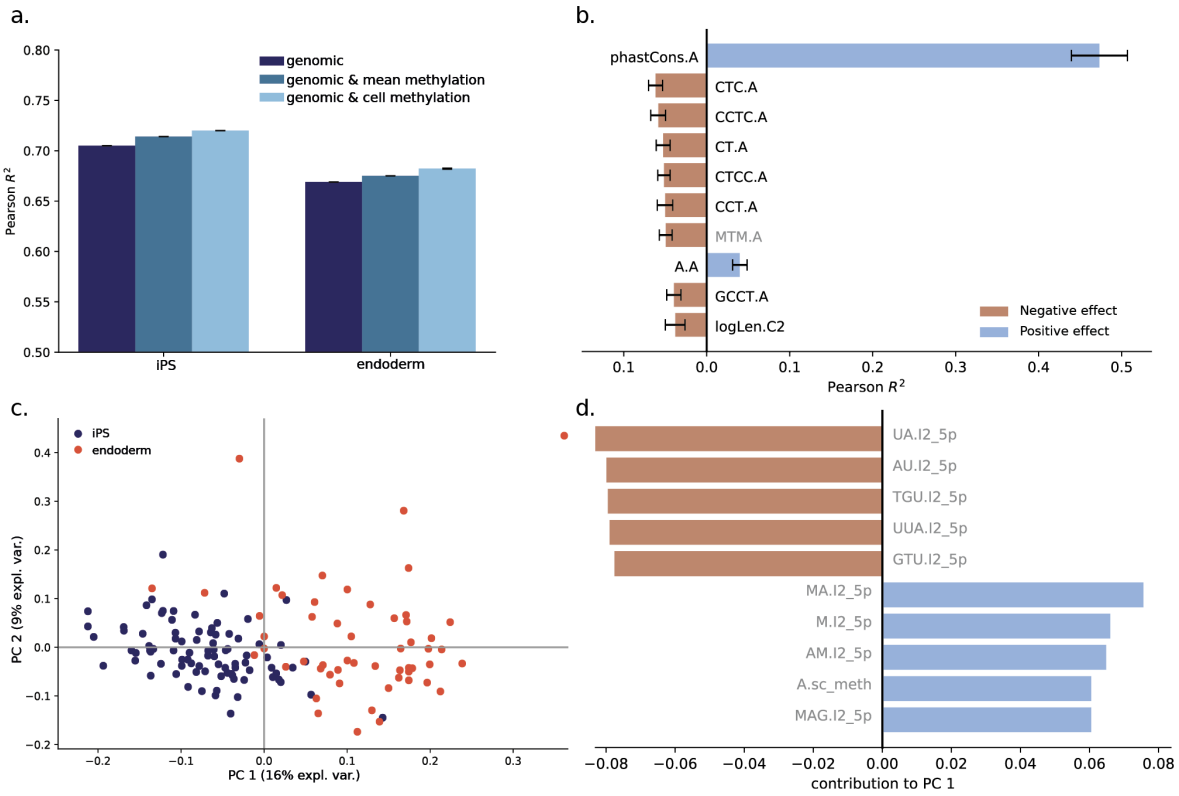

Figure 9. Prediction of single-cell splicing variation. **a.** Prediction accuracy (R^2 based on ten-fold cross validation, y-axis) of different regression models for predicting single-cell splicing rates in iPS cells and endoderm cells (x-axis). The genomic model is based on sequence k-mers, conservation scores and sequence lengths of sequence contexts (genomic features, dark blue). Other models account for average methylation rates across cells (genomic & mean methylation features, median blue), or cell-specific methylation rates (genomic & cell methylation features, light blue). Error bars denote $\pm$1 standard deviation across four repeat experiments. **b.** Features ranked by relevance for predicting splicing in iPS cells as determined by single-feature regression models. The most relevant features are features of the alternative exon, and include a methylated k-mer. Error bars denote $\pm$1 standard deviation of the feature relevance across cells. Methylation features are indicated in grey. **c.** PCA on all cell-specific feature weights as shown in (b). The first PC captures differences between differentiation states. **d.** The ten features with the largest contribution to the first PC of (c). Negative correlations are shown in red, positive ones in blue. Methylation features are shown in grey.

3.2.4 Deep learning modelling of splicing

Whereas linear models require *a priori* definition of features that we here averaged per sequence context (*e.g.*, average k-mer load of the alternative exon), other models can use more complex data like raw DNA sequence as input. We therefore considered if the application of CNNs (Section 1.2.3) to predict single-cell splicing based on raw DNA sequence would add new insights to our analyses, and improve our understanding of the complex relationship between DNA sequence, DNA methylation and splicing rates. When encoding our DNA sequences, we either used the standard four-letter code for the four DNA bases or the previously used five-letter code that differs between methylated ('M') and un-methylated ('U') cytosines (Section 3.2.3).

CNNs have already been used to predict methylation rate, TFBSs and alternative splicing from DNA sequence (Alipanahi et al., 2015; Angermueller et al., 2017; Wainberg et al., 2016; Xiong et al., 2015; Section 1.2.3). However, no deep model for predicting splicing from DNA sequence (or DNA methylation) in single cells has been developed so far. Therefore, we set out to construct our own CNN that can predict splicing from DNA sequence (and DNA methylation) in a single-cell setting.

For each cassette exon, a sequence window of $\pm$400 base pairs around the centre of the alternative exon was used to train a CNN. This region was selected since the linear model predictions revealed the alternative exon to be the most relevant region for splicing predictions. In addition, most exons are shorter than 200 base pairs (Sakharkar et al., 2005), *i.e.*, the region is chosen to include the whole exon as well as regions from the adjunctive introns. The DNA sequence was one-hot encoded, and subsequently multiplied with the conservation score of the region, as previously derived from BRIE (Huang & Sanguinetti, 2017; see Section 3.5 for details). The sequence conservation was taken into account since previous studies (*e.g.*, Wainberg et al., 2016) and our first analyses pointed towards relevance of this feature for splicing regulation.

In the CNN analyses, the same set of cassette exons was used as described in Section 3.2.1; one CNN per cell type was trained. Here, I only describe the analysis of the iPS cells, but the results in endoderm cells were comparable. Briefly, the transformed DNA sequences were used as input for a CNN to predict binary splicing, *i.e.*, inclusion (PSI $\geq$ 1/3) or exclusion (PSI < 1/3) of the alternative exon. Hereby, the PSI threshold of 1/3 was chosen (instead of a mean threshold of 1/2) since it approximately balanced our positive and negative samples; imbalanced datasets have been shown to have a detrimental effect on the prediction performance of deep models (Buda et al., 2018). Our final CNN consisted of one convolutional layer (58 filters of length 10), a bias and rectification layer that obtains non-negative values by applying the rectified linear unit function, a maximum pooling step that compresses the data (pooling length of 8 base pairs) and an additional fully connected layer (50 nodes; Section 1.2.3 for details about CNNs). The convolutional layer is hereby the core component of a CNN, which acts as a feature extractor to

identify predictive DNA sequence motifs. When the filters of the convolutional layer slide over the DNA sequence in an overlapping manner, they assign a score to each position and hence create a list of scores that can be thought of as a position weight matrix (Stormo et al., 1982): The higher the score, the higher the similarity between the respective DNA sequence and the motif learned by the filter (Section 1.2.3).

For training the CNN, we randomly split our dataset into training (60%), validation (20%) and test (20%) datasets considering cassette exons in a given cell. We did not split the dataset on the cell level since multiple cells shared multiple alternatively spliced genes, which may lead to inflated prediction performance. The model was trained until the weights converged (patience=3 iterations; maximum number of iterations=100) using the training dataset. We then optimized across some sets of hyperparameters, using the validation dataset. Importantly, we chose the number of described hidden layers (see above) since other models with more hidden layers did not converge after 100 iterations. The weights of the model did also not converge when we tried to predict quantitative PSI. This might be due to a too small sample size that did not allow attributing quantitative PSI changes to base-level changes in the DNA sequence. The binary PSI, however, yielded a successfully trained model.

For the four-letter encoded DNA, we obtained a CNN with an AUC (area under the receiver operating characteristic curve) of 0.908 using the test dataset (see **Figure 10a** for the receiver operating characteristic (ROC) curve). This performance is comparable to state-of-the-art CNNs that predict alternative splicing in bulk data (Wainberg et al., 2016), which was a motivating result given that single-cell data is inherently noisier than bulk data. The five-letter code CNN that distinguishes methylated and un-methylated bases achieved an AUC of 0.912 using the test dataset (see **Figure 10b** for the ROC curve). Inclusion of methylation information therefore seemed to improve predictability of splicing slightly, similar to what we had observed in the linear models.

To our knowledge we here present the first CNN trained on single-cell splicing data, which additionally allows for inclusion of DNA methylation information. We showed that CNNs predict single-cell alternative splicing with good performance. Opposed to linear models that are based on *a priori* defined features, the CNN can further be queried for sequence motifs relevant for splicing modulation. These sequence motifs can be detected by studying which motifs of the input DNA sequences lead to the highest activation of a filter and by then using these motifs to build a consensus sequence according to Ding et al. (2017).

To make the (five-letter code) model accessible to the scientific community we contributed it to the kipoi model zoo (kipoi.org), a framework for deep learning applications in genomics (Avsec et al., 2019). Within this framework, our model can now be employed to make regulatory gene predictions using genomic and epigenetic variation data of any single-cell dataset. Finally, our model can be combined with other deep learning models provided by kipoi or the user to improve prediction accuracy.

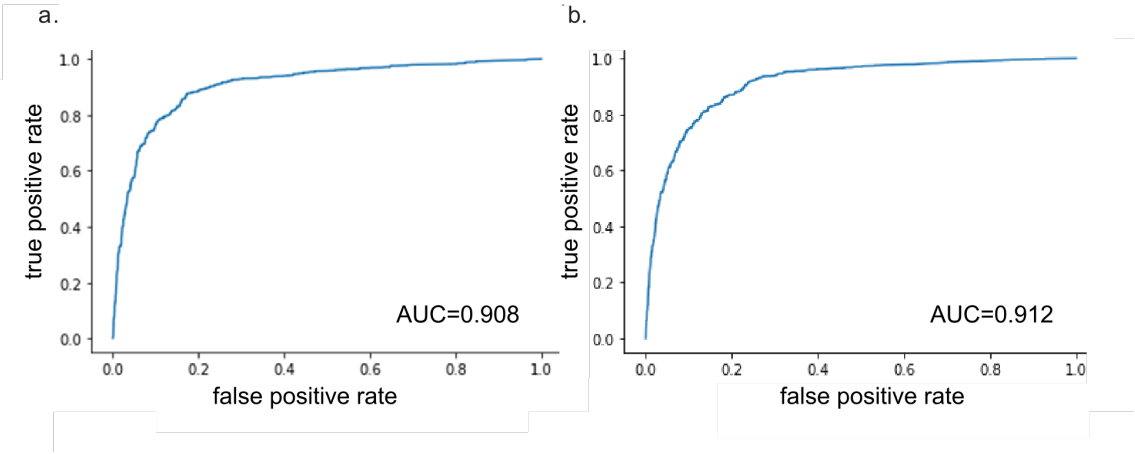

Figure 10. Prediction of splicing rates with CNNs based on genomic sequences. **a.** Performance of the CNN based on only genomic information (four-letter code model), assessed by ROC curve and AUC. **b.** Performance of the CNN based on genomic and methylation information (five-letter code model), assessed by ROC curve and AUC.

3.3 Prediction of splicing categories

3.3.1 Prediction of splicing categories of individual exons

Next, we set out to study variability in alternative splicing patterns across cells. To describe cassette exons by both, their splicing rate and the variability of this splicing rate across cells, we classified them into five distinct categories, using a scheme similar to the one described by Song et al. (2017): We considered 1) excluded, 2) included, and three intermediate splicing categories: 3) overdispersed, 4) underdispersed and 5) multimodal categories (**Figure 11a**; **Figure 11b** shows the category distribution according to **Figure 8c**). We then trained multinomial regression models to classify individual exons using features sets analogous to the ones considered for the quantitative regression models that predict splicing rates (see Section 3.2.3). A model based on genomic features yielded a macro-average AUC across all five categories of 0.85 (**Figure 11c**), where again sequence conservation in different contexts was the most informative individual feature. Interestingly, we observed differences in the feature relevance across splicing categories: i) included and excluded exons, where the most relevant features were located in the alternative exon, and ii) the intermediate splicing categories, where features of the flanking exons were most informative. Predictions of included and excluded exons were most accurate on the individual category level (AUC=0.96 for included and excluded in iPS, AUC=0.94 for included in endoderm, AUC=0.96 for excluded in endoderm cells; see **Figure 11d** for individual and macro-average ROC curves). These prediction

accuracies exceeded previously reported results in bulk data (Xiong et al., 2015). Even higher accuracies were achieved when training a model to differentiate between included and excluded exons alone (AUC=0.99), whereas lower prediction accuracies were achieved for differentiating just the intermediate splicing categories from each other (AUCs=0.7 to 0.9 for all *one-vs-one* predictions). The inclusion of methylation features did not improve the prediction performance of these categorical models (**Figure 11d**).

Consistent with this, we also found that a model based on DNA methylation alone did not yield accurate predictions although methylation contained some information for identifying underdispersed cassette exons (**Figure 11d**). We therefore investigated the distribution of DNA methylation patterns across splicing categories, observing distinct distributions of DNA methylation in the upstream exon of underdispersed cassette exons (**Figure 11e**). This effect was consistent, although less pronounced, in other sequence contexts (decreasing from the upstream to the downstream exon).

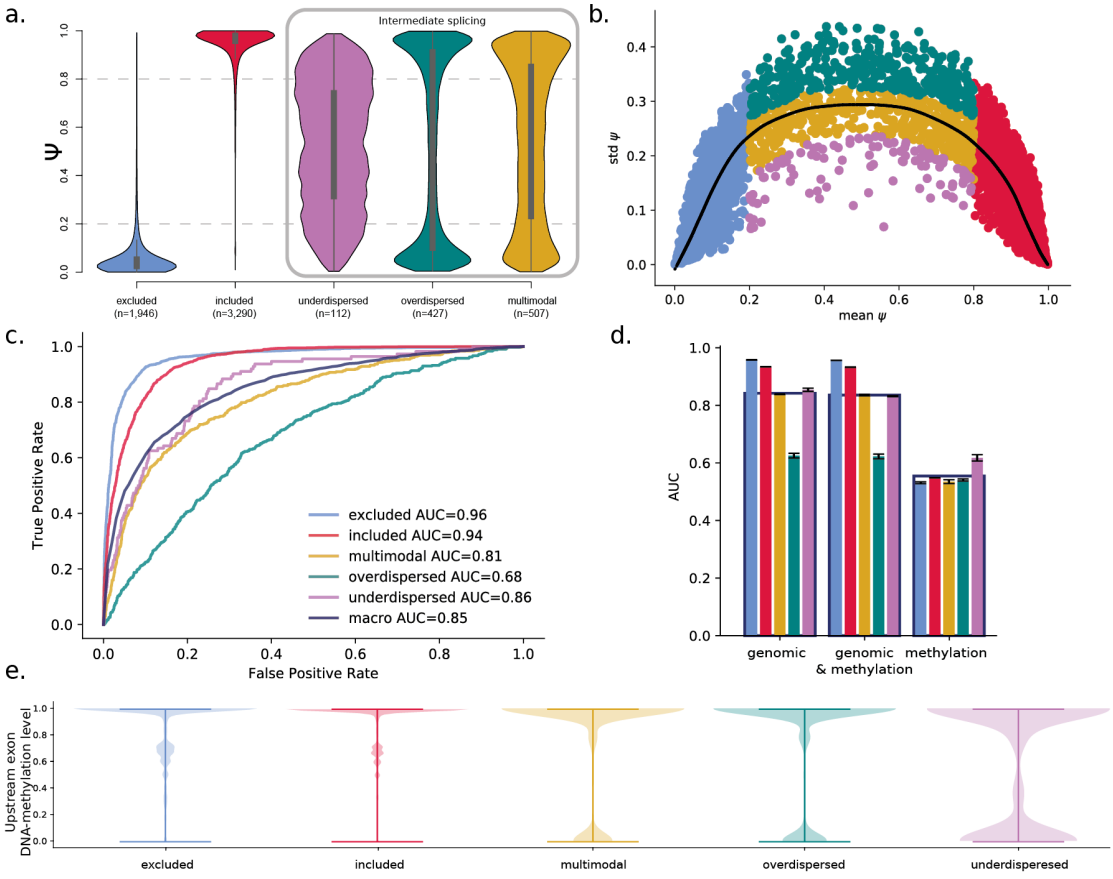

Figure 11. Classification of cassette exons based on their single-cell splicing patterns. **a.** PSI distributions for five splicing categories, inspired by Song et al. (2017). The intermediate splicing categories that can only be distinguished if single-cell data is provided are highlighted by a grey frame. **b.** Variation of PSI (standard deviation) across cells as a function of the average

inclusion rate of cassette exons across 86 iPS cells, coloured according to their respective splicing category as defined in (a). The solid black line denotes the LOESS (locally weighted scatterplot smoother) fit across all cassette exons. **c.** Prediction performance of logistic regression for predicting splicing categories based on genomic features. The ROC curves and AUCs for each splicing category and for the macro average across all categories are shown. Splicing categories are coloured as in (a). **d.** Prediction performance of the logistic regression models for all splicing category, based on either genomic features ('genomic', left), genomic and all DNA methylation features ('genomic & methylation', centre) or only DNA methylation features ('methylation', right). The genomic model includes k-mers, conservation scores and sequence lengths of the sequence contexts. The genomic and methylation model additionally includes all DNA methylation features. The methylation model only includes average DNA methylation features per sequence context. Splicing categories are coloured as in (a). Error bars denote ± 1 standard deviation across four repeat experiments. **e**. Distribution of DNA methylation levels in the upstream exon ('C1') per splicing category. Methylation is relatively decreased in underdispersed exons. Splicing categories are coloured as in (a).

We again investigated the consistency of these models between iPS and endoderm cells. To do so, we trained the genomic model on endoderm cassette exons and assessed this model's predictions on iPS-specific cassette exons, which resulted in a prediction accuracy that was comparable within-cell-type prediction (macro AUC=0.82). However, inclusion of the DNA methylation features into the model resulted in a decline in the cross-cell-type prediction performance (macro AUC=0.54). Similar to the cross-replication analysis using models trained on quantitative splicing rates (Section 3.2.3), this emphasised the relevance of cell type-specific DNA methylation for accurately predicting splicing. The relevance of genomic information for splicing, on the contrary, seems to be more conserved and applies across cell types.

Next, we assessed if the same splicing category analyses applied to the scM&T-Seq dataset of mouse ES cells would yield similar results. We observed that the performance for splicing category prediction was very similar to the performance in the endoderm and iPS cells (macro AUC=0.82, for both, the genomic and the genomic and methylation model). We observed the same distinct distribution of DNA methylation in the upstream exon of underdispersed cassette exons. However, the relationship between the DNA methylation levels and underdispersed category could not be confirmed by the mouse ES cells (Linker, Urban et al., 2019).

3.3.2 Splicing category switches during cell differentiation

Finally, we assessed changes in the splicing category switches between cellular differentiation stages. Similar to previous observations in the context of neuronal iPS differentiation (Song et

al., 2017), we observed that a majority (88%) of the cassette exons retained their category during differentiation (**Figure 12a**). Also, no cassette exon switches from included to excluded or *vice versa* were observed. Instead, the majority of the switching events (55%) were observed within the three intermediate splicing categories. The most prevalent event was switching to the multimodal category: 51% of the underdispersed and close to 45% of the overdispersed cassette exons in iPS cells switched to be multimodal after differentiation to the endoderm state.

After observing these interesting patterns in category switches between the two differentiation stages, we set out to build a set of logistic ridge regression models based on genomic and methylation features to predict category switching ability of cassette exons during differentiation (**Figure 12b** for prediction performance). The model based on genomic features had limited power to predict category switches, and DNA methylation did not significantly improve the prediction. However, moderately better prediction performance was achieved for the switching behaviour of over- and underdispersed cassette exons.

Lastly, we explored if DNA methylation had an impact on the switching behaviour between splicing categories. We assessed if DNA methylation changed within switching cassette exons during differentiation. The DNA methylation levels of cassette exons that switched category only changed slightly during differentiation. However, we observed that DNA methylation of the alternative exon of switching cassette exons differed from non-switching cassette exons at the iPS stage (**Figure 12c**). DNA methylation of cassette exons that would switch from an included or excluded state was increased in the upstream exon ('C1') in comparison to their non-switching counterparts. In the case of switching overdispersed cassette exons, we observed higher DNA methylation levels within and in the vicinity of the alternative exon, again compared to their non-switching counterparts (Linker, Urban et al., 2019).

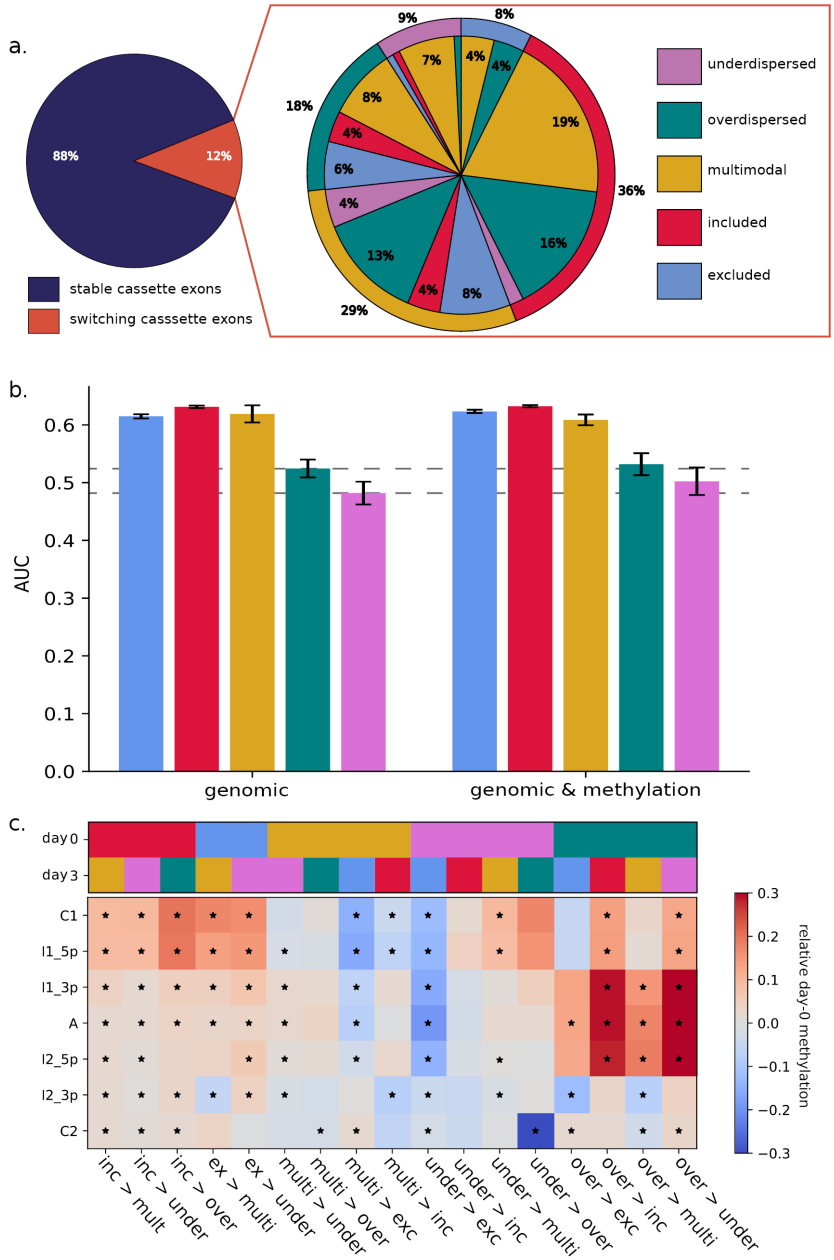

Figure 12. See next page for figure caption.

Figure 12. Comparison of splicing category distributions between iPS and endoderm cells. **a.** Pie chart showing the number of category switches between iPS and endoderm cells (left panel). The zoom-in (right panel) shows details of different category switches. The outer pie chart shows the splicing category of each cassette exon at the iPS state and the internal pie chart shows the respective category at endoderm state. Non-annotated slices in the pie chart reflect <1% of the data. **b**. Performance of logistic ridge regression models that predict absence/presence of switching splicing categories between iPS and endoderm states. DNA methylation information improves prediction of the under- and overdispersed cassette exons. The categories are coloured according to (a). Error bars denote ± 1 standard deviation across four repeat experiments. The horizontal dashed grey lines highlight slight differences in prediction accuracy of the under- and overdispersed categories between the models based on genomic or on genomic and methylation data. **c.** DNA methylation changes associated with the observed category switches (x-axis). The top panel shows the iPS and endoderm splicing categories coloured according to (a). The bottom panel shows a heatmap of DNA methylation levels within the seven sequence contexts of a cassette exon at iPS state (y-axis). The colour of the cells denote the relative DNA methylation level, relating DNA methylation levels of switching to the DNA methylation levels of non-switching cassette exons. Significant changes (FDR ≤ 5%) in DNA methylation are marked with a star. DNA methylation of the alternative exon ('A') and of its vicinity is increased in cassette exons that switch from the overdispersed category. Cassette exons that switch from either included or excluded to any other splicing category show increased DNA methylation of the upstream exon ('C1').

3.4 Discussion and conclusion

This chapter describes the analysis of alternative splicing in single cells and the impact of both genomic and epigenetic factors. Our study focused on variation of splicing in cassette exons at two different stages of cell differentiation, and we used data from mouse ES cells to replicate the most relevant results. This chapter extends the scope of this thesis from general associations between genomic variation and gene expression (Section 2) to (i) the functionally relevant phenotype of alternative splicing, (ii) the study of single-cell instead of bulk data, (iii) the impact of epigenetic variation on molecular functions, and (iv) the application of more complex machine learning models like CNNs.

First, we demonstrated that splicing events do not strictly follow either of the canonical cell or gene models of splicing variation across cells. Instead we found a substantial proportion of exons that were better described by an intermediate model (**Figure 8a**). We therefore

suggested a more complex model that involves both, within- and across-cell variability in splicing.

We showed that genomic features of the cassette exon modulated single-cell splicing, as previously assessed in bulk data. We additionally identified predictive epigenetic features derived from DNA methylation profiles. Hereby, DNA methylation was most strongly linked to single-cell splicing ratios. When studying splicing variation in bulk populations ('pseudo bulk'), most of the information encoded in DNA methylation got lost. A reason may be that average methylation primarily reflects genomic features, in particular due to the strong correlation between DNA methylation and cytosine-related genomic features. However, the cell-to-cell heterogeneity of DNA methylation for the same loci does appear to contain predictive information of splicing beyond these bulk properties in the same cells. These results indicate that the relationship between splicing and DNA methylation is highly locus-specific. This might explain why DNA methylation improved prediction performance to a very limited extent when we predicted average splicing rates across single cells, or when we quantified splicing based on pseudo bulk data.

Sequence conservation has already been described as an important predictor of alternative splicing in bulk transcriptome studies (Wainberg et al., 2016). We were able to confirm this relationship in single-cell data. Besides sequence conservation, the most relevant features to accurately predict splicing in our dataset were the k-mer motifs CTC, CT and CCT within the alternative exon (**Figure 9b**). These k-mers point towards involvement of CTCF. Previous work has already shown that CTCF motifs are linked to splicing by slowing down RNA Polymerase II, thereby leading to a higher chance of exon inclusion (Shukla et al., 2011). However, according to existing literature, these motifs are located in the flanking introns, and not in the alternative exon that we identified as a functional source of CTCF motifs. Moreover, there is also a known link between DNA methylation and CTCF motifs (Shukla et al., 2011): Methylation of CTCF binding sites can block CTCF and thereby result in decreased inclusion rates of an exon. In our analyses, we however found that the methylated equivalents of these k-mers were less predictive of splicing than the k-mers that were just based on genetic and not on epigenetic information. We therefore suggest a more complex involvement of DNA methylation in alternative splicing, which our models were not able to capture (Section 3.2.3).

In addition to modelling splicing ratios, we considered categorical models of splicing to gain insights into variability of splicing across cells (**Figure 11**). The splicing categories considered in our model reflect both, overall splicing rate and splicing variability across cells. The splicing state of exons with included *versus* excluded splicing states could be predicted with high accuracy. In contrast, the intermediate splicing categories that are reflective of single-cell variability could only be predicted with lower accuracy. This might be due to the lower number of cassette exons assigned to these categories in our dataset and therefore smaller sample sizes when training regression models (multimodal n=506, overdispersed n=427, underdispersed n=110, *versus* included n=3,278 and excluded n=1,944 in iPS cells). Alternatively, this

decreased accuracy might reflect increased vulnerability to assay noise when studying intermediate splicing categories, as well as biologically more complex regulatory dependencies. As in the cell-specific regression models, we observed that DNA sequence conservation scores were the most informative features for predicting splicing. Interestingly, for intermediate categories, features of the genomic regions in the vicinity of the alternative exon rather than of the exon itself seemed to be predictive of splicing variability. Whereas DNA methylation did not contribute to improving the splicing prediction, we observed that DNA methylation levels of underdispersed cassette exons were significantly reduced in all genomic contexts, most significantly in the upstream exon. We hypothesised that the lower DNA methylation levels of underdispersed cassette exons give the sequence motifs more power to control splicing levels, *i.e*., that increased DNA methylation levels lead to more stochasticity in splicing. This hypothesis is supported by the effect direction of methylation features, which was opposed between overdispersed and underdispersed cassette exons. We finally observed that the methylation k-mers were on average less informative of splicing than non-methylation features, potentially further supporting this hypothesis.

By leveraging data from two cell types, we were able to assess the consistency of both, splicing prediction performance and the associated relevance of the genomic and methylation features, across cell types. The differences between features predictive of splicing between iPS and endoderm cells were primarily observed within the (methylated) k-mers, which is consistent with the known alterations of TF activity and DNA methylation between cell types. We were able to confirm the findings from Song et al. (2017) that only few cassette exons switch splicing categories during differentiation (**Figure 12a**). Additionally, as previously described in context of neural differentiation (Song et al., 2017), switches between included and excluded categories, and *vice versa*, were not observed. Instead, most of the splicing category switches occurred between the three intermediate splicing categories. Hereby, DNA methylation differences seemed to predate the switching ability. Using ridge regression, we were able to predict if a cassette exon would switch its splicing category between the cell types, but only with limited power (**Figure 12b**). Again, DNA methylation only seemed to be relevant for intermediate splicing categories: It slightly improved the predictability of switching in over- and underdispersed categories.

The novelties of our analyses simultaneously revealed their main limitations. Single-cell sequencing approaches intrinsically deliver fewer reads than bulk data to quantify both, gene expression and DNA methylation levels. In our dataset, especially the genome coverage of the bisulphite-treated DNA remained low due to the low quantities of starting material. Using computational imputation, we were able to mitigate this effect to some extent. However, imputation strategies suffer from inherent limitations and especially our decision to set cytosines with no methylation information in their genetic proximity to an un-methylated state might have biased our analyses.

The intrinsic properties of single-cell data also affected the accuracy of our estimates of splicing ratios across cassette exons. We opted for a lenient threshold on read depth to determine splicing ratios in order to obtain more cassette exons to train our models, but this simultaneously rendered splicing ratios less accurate in comparison to deep-sequenced bulk data: The low read depth increases the chance of missing an isoform or cassette exon, an effect known as a dropout. Dropouts in single-cell RNA-Seq data can have a strong impact on fitting the cell- or gene-model: If an isoform were completely unobserved, this would decrease the fit of the gene model. On the contrary, erroneously sequencing multiple cells at once would decrease the fit of the cell model. Given, however, that our results were robust across cassette exons, cell types and species, the overall findings we reported are unlikely to be affected by these properties of single-cell data.

In summary, we showed for the first time that alternative splicing and splicing variability across cells can be predicted from genomic and DNA methylation variation in single cells. We assessed the impact of DNA methylation and cellular features on cassette exon splicing, and were able to replicate our findings in two human cell types and mouse ES cells. We investigated stability and variance of splicing between the two cell types and, importantly, we showed that DNA methylation primes splicing switches during cell differentiation (Linker, Urban et al., 2019).

3.5 Detailed experimental and statistical approaches

3.5.1 Experimental procedures

Single-cell transcription and methylation data was generated from a single donor from the Human Induced Pluripotent Stem Cells Initiative (HipSci) (Kilpinen et al., 2017; Streeter et al., 2017), using the protocol for scM&T-Seq (Angermueller et al., 2016). Line joxm_1, an iPS cell line derived from male fibroblasts cells from the HipSci project, was cultured and triggered into differentiation towards endoderm. scM&T-Seq data was generated for 93 cells (together with one empty well as negative control and two 15-cell and 50-cell positive controls) at iPS and endoderm stage, yielding 186 cells for analysis.

The joxm_1 IPSC line was cultured in LifeTechnologies Essential 8 media. For dissociation and plating, cells were washed with Dulbecco's Phosphate-Buffered Saline and dissociated using StemPro Accutase at 37°C for 3-5mins. Colonies were fully dissociated through gentle pipetting. Cells were washed with MEF medium (Hannan et al., 2013) and pelleted gently by centrifuging at 285x*g* for 5mins. Cells were re-suspended in E8 media, passed through a 40µm cell strainer, and plated at a density of 60k cells per well of a gelatin/MEF coated 12-well plate in the presence of 10µM Rock inhibitor – Y27632 [10 mM]. Media was replaced with fresh E8 free of Rock inhibitor every 24 hours post plating. Differentiation into definitive endoderm started 72 hours post plating as previously described (Hannan et al., 2013).

During all staining steps, cells were protected from light. Cells were dissociated into single cells using accutase and washed with MEF medium as described above. Approximately 1M cells were re-suspended in 0.5mL of differentiation stage specific medium containing 5µL of 1mg/mL Hoechst 33342 (Thermo Scientific). Staining with Hoechst was carried out at 37°C for 30 min. Unbound Hoechst dye was removed by washing cells with 5mL FACS buffer (PBS + 2% BSA + 2 mM EDTA with nuclease-free BSA and PBS). For staining of cell surface markers Tra-1-60 (BD560380) and CXCR4 (eBioscience 12-9999-42), cells were re-suspended in 100µL of FACS buffer with enough antibodies to stain 1M cells according to the manufacturer's instructions, and were placed on ice for 30mins. Cells were washed with 5mL of FACS buffer, passed through a 35µM filter to remove clumps, and re-suspended in 250µL of FACS buffer for live cell sorting on the BD Influx Cell Sorter. Live/dead marker 7AAD (eBioscience 00-6993) was added just prior to analysis according to the manufacturer's instructions and only living cells were considered when determining differentiation capacities. Living cells stained with Hoechst but not Tra-1-60 or CXCR4 were used as gating controls (Linker, Urban et al., 2019).

scM&T-Seq

As previously described in Angermueller et al. (2016), scM&T-Seq library preparation was performed following the protocols for G&T-Seq (single-cell genome and transcriptome sequencing; see Macaulay et al., 2016) and scBS-Seq (single-cell bisulfite sequencing; see Clark et al., 2017), with minor modifications as follows. G&T-Seq washes were performed with 20µl volumes, reverse transcription and cDNA amplification were performed using the original Smart-Seq2 volumes (Picelli et al., 2014) and Nextera XT libraries were generated from 100-400pg of cDNA, using 1/5 of the published volumes. RNA-Seq libraries were sequenced as 96-plexes on a HiSeq 2000 using v4 chemistry and as paired end reads (with a respective length of 125 base pairs). BS-Seq libraries were sequenced as 24-plexes using the same machine and settings, which yielded a mean of 7.4M raw reads after trimming (Linker, Urban et al., 2019).

DNA methylation processing and quantification

For DNA methylation data, scBS-Seq data were processed as described (Clark et al., 2017). Reads were trimmed with Trim Galore (Andrews, 2010; Krueger, 2011; Martin, 2011), using default settings for DNA methylation data and additionally removing the first 6 base pairs. Subsequently, Bismark (version 0.16.3) was used to map the bisulfite data to the human reference genome (version 38), in single-end, non-directional mode, which was followed by de-duplication and DNA methylation calling using default settings (Krueger & Andrews, 2011). All but two single-cell libraries (alignment rate < 15%) yielded good alignment rates (mean alignment rate of 43%), with negative control wells having very low mappability (mean of 2%). Additionally, seven samples with a library size of less than 1M reads were removed.

To mitigate typically low coverage of scBS-Seq profiles (20-40% according to Angermueller et al., 2017), we applied DeepCpG to impute unobserved methylation states of individual CpG sites (Angermueller et al., 2017). DNA methylation profiles in iPS and endoderm cells were imputed separately, using the default settings of the method. Predicted methylation states were binarised according to DeepCpG probabilities as follows: CpG sites with a probability equal to or lower than 0.3 were set to 0 (un-methylated base), all methylation sites with a probability greater than 0.7 were set to 1 (methylated base). Intermediate methylation levels were handled as missing. After imputation the methylation data was mapped to the human reference genome version 37 to match the expression data (see below), using the UCSC lift-over tool (Kent et al., 2002).

We integrated the imputed methylation information into the DNA sequence by distinguishing methylated ('M') and un-methylated ('U') cytosines. Cytosines without methylation information after imputation were assigned the value of the closest cytosine with methylation information. If there was no methylation information within 900 base pairs around the cytosine, its state was set to un-methylated (Linker, Urban et al., 2019).

Gene expression and splicing quantification

For single-cell RNA-Seq data, adapters were trimmed from reads using Trim Galore using default settings. Trimmed reads were mapped to the human reference genome build 37 using STAR (version 020201) in two-pass alignment mode, using the defaults proposed by the ENCODE consortium (Dobin et al., 2013).

Expression quantification was performed per cell type using Salmon (version 0.8.2), using the 'SeqBias', 'gcBias' and 'VBOpt' options on transcripts derived from ENSEMBL 75 (Patro et al., 2017). Transcript-level expression values were summarized at gene level (estimated counts) and quality control of RNA-Seq data was performed using scater (McCarthy et al., 2017). Cells with the following transcriptomic features were retained for analysis: at least 50k gene-level counts from endogenous genes; at least 5k genes with non-zero expression; less than 90% of counts come from the 100 most highly expressed genes in the cell; less than 20% of counts come from ERCC spike-in sequences; a Salmon mapping rate of at least 40%.

Of the 192 cells, 86 (iPS) and 59 (endoderm) cells passed quality control in terms of both DNA methylation and gene expression data. Exon splicing rates in individual cells were quantified using BRIE (Huang & Sanguinetti, 2017). BRIE calls splicing at predefined cassette exons and quantifies splicing using exon reads in single-cell data. By default BRIE combines a prior based on sequence features with a likelihood calculated from RNA-Seq reads within a mixture-modelling framework. As our aim is to model the local and global determinants of splicing, we used splicing rate estimates based on the observed data at individual exons only. We quantified splicing for between 1,386 and 14,434 exons per cell (minimum coverage of five reads per cassette exon), and finally considered 6,282 (iPS) and 4,096 (endoderm) cassette exons that were detected in at least ten cells for further analysis.

We quantified PSI in three different ways, namely via the actual single-cell splicing rate (PSI), thepseudo bulk splicing rate (bPSI) or the variance of the splicing rate. To derive bPSI values per cassette exon, we aggregated the PSI values per cassette exon. The variance of splicing ratios per cassette exon was defined as the standard deviation of PSI across cells (Linker, Urban et al., 2019).

Replication cohort

To replicate our results, we processed the mouse ES single-cell scM&T-seq data (n=80 cells) presented in Angermueller et al. (2016). We preprocessed the aligned RNA and DNA methylation data to quantify splicing following the same protocols that were applied to the human data, with the following changes: GRCm38 was used as a reference for imputation, genome and transcriptome annotations were based on Gencode v18 ('GRCm38.p6.genome.fa' as genomic, 'gencode.vM18.annotation.gff3' as transcriptomic reference, available at ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/release_M18/ [August 2018]), and conservation scores were taken from 'mm10.60way.phastCons.bw' [August 2018] downloaded from UCSC (Kent et al., 2002).

Out of the 80 cells, 12 cells did not pass quality control on the transcriptome data. Cells with less than 500k sequenced reads and cells that had less than 80% of the reads aligned to the genome were removed. Additionally, four cells did not pass quality control on the DNA methylome data. Cells with less than 1M reads aligned and a Bismark (Krueger & Andrews, 2011) mapping efficiency below 7% were discarded. These filters yielded 68 cells that were used for the splicing analysis and 64 that were used for DNA methylation-related analyses. In these cells, we quantified between 649 and 1,433 cassette exons per mouse ES cell (minimum coverage of 5 RNA-Seq reads); in the replication analysis, we considered 2,194 exons that were supported by at least one cell.

3.5.2 Definition of sequence features and splicing categories

The genomic features used to predict the splicing ratios and its variance were based on the features described by both, BRIE (Huang & Sanguinetti, 2017) and Xiong et al. (2015). The following features were calculated for five sequence contexts of each cassette exon (*i.e.*, the alternative exon, the two neighbouring exons and the two introns between the exons): the (log) length, and the strength of the splice site motifs at the exon-intron boundaries. The strength of a splice site motif was defined as the similarity between this splice site motif and known splice motifs. Additional features were calculated for seven sequence contexts (*i.e.*, the alternative exon, the two neighbouring exons and the 5’ and 3’ boundaries of both introns): conservation scores according to PhastCons (Siepel et al., 2005) and k-mer frequencies (with $k \leq 3$). The k-mers reflect the percentage of nucleotides in the context that match the respective specific

motif. For these features, only the two boundary regions of the introns (300 base pairs length) were used since intron length is highly variable and the boundaries are the most relevant contexts for splicing.

In addition to the genomic features, we defined DNA methylation features for each of the seven sequence contexts. We considered cell-specific methylation levels per context and extended the k-mer features by considering un-methylated ('U') and methylated ('M') cytosine. For the bPSI model, we included the mean frequencies of the k-mers that contained 'M' or 'U' across cells and the mean and standard deviation of the seven sequence contexts across cells per cell type.

In bulk RNA-Seq data, splicing events can be broadly categorised into two major categories: included and excluded. Leveraging the single-cell information, we defined more fine-grained splicing categories that reflect both, splicing rates and splicing variability across cells (inspired by Song et al. (2017): 1) excluded (mean PSI < 0.2), 2) included (mean PSI > 0.8), 3) overdispersed, 4) underdispersed and 5) multimodal. The later three categories describe the extent of splicing variation across single cells, since cassette exons with intermediate average splicing rates (here 0.2 ≤ mean PSI ≤ 0.8) exhibit substantial differences in splicing variance. To assign intermediate cassette exons to these three splicing categories, we calculated the distribution of the distance between the observed and the expected variation in PSI per cassette exon and cell type. The expected variation was calculated by a scaled binomial standard deviation, using mean PSI as scaling factor (Faigenbloom et al., 2015). We then defined cassette exons with PSI deviating from the expected PSI beyond the upper boundary of the third quartile + 1.5x interquartile range (IQR) as overdispersed. This boundary corresponded to PSI > 0.016 in iPS and PSI > 0.022 in endoderm cells. Likewise, we defined underdispersed cassette exons as exons with a PSI smaller than the lower boundary of the first quartile - 1.5x IQR. This corresponded to PSI < -0.032 in iPS and PSI < -0.039 in endoderm cells. The remaining intermediate cassette exons were assigned to the multimodal category (Linker, Urban et al., 2019).

3.5.3 Prediction of splicing ratios and categories

We applied linear ridge regression to model PSI and bPSI and (multi-class) logistic ridge regression to model the splicing categories. The models are based on only the genomic features or on both genomic and DNA methylation features. The performance of linear models was evaluated using Pearson R^2 between predicted and observed splicing rates. For the multi-class prediction models, we applied a *one-versus-rest* scheme and reported the per-category and macro-average AUCs. To determine the most relevant individual features, we additionally trained regression models with a single feature at a time. In the case of the linear models, we reported Pearson correlation (both, R and R^2) per feature. In the case of the logistic models, we reported the absolute weight multiplied by the standard deviation of the feature, and the AUC per feature. We assessed performance and parameters of models by using a ten-fold cross-

validation with fixed training-validation splits across experiments. To assess variability of prediction performances, we repeated the cross-validation procedure four times with different fixed training-validation splits. Error bars indicate ± 1 standard deviation of the statistic in question (AUC or R^2) (Linker, Urban et al., 2019).

3.5.4 Relating DNA methylation heterogeneity to splicing

We applied Spearman correlation to link splicing at a single locus to variation in DNA methylation across cells. The test was performed and corrected per sequence context within the cassette exon. We only considered cassette exons where both, variation in splicing and DNA methylation, were observed in the relevant sequence contexts. In total, 5,280 cassette exons were tested for iPS and 2,622 for endoderm cells. The P-values obtained from the test were adjusted for multiple testing using the Q-value package in R (Bass et al., 2015; Storey, 2010). The gene enrichment of cassette exons which were significantly associated with DNA methylation was then performed with G:Profiler (Reimand et al., 2016), using all observed cassette exons per cell type as background gene sets (Linker, Urban et al., 2019).

All Software and scripts of this project are available as jupyter notebooks at https://github.com/PMBio/scmt_splicing.

4 Transcriptome-guided decomposition of homologous recombination repair deficiency

Contributions

This project was realised under supervision of Sebastian Waszak and Oliver Stegle. While I was responsible for data pre-processing, statistical and downstream analyses, Sebastian Waszak contributed key ideas and advice on statistical analyses. The work took place in regular exchange with him and Oliver Stegle.

4.1 Introduction

4.1.1 Homologous recombination repair deficiency

DNA double-strand breaks are one of the most serious damages for genomic integrity (Jackson & Bartek, 2010). The two distinct mechanisms of non-homologous end joining (NHEJ) and homologous recombination repair (HR) repair the majority of DNA double-strand breaks in eukaryotic cells. NHEJ repairs in the following way: So-called Ku proteins recognise double-strand break fractions of genes and activate the DNA dependent protein kinase catalytic subunit, leading to activation of more enzymes, polymerases and of the DNA ligase IV. While this mechanism is already prone to error due to the lack of a template that could be used for faithful repair, the Ku-independent NHEJ, called microhomology-mediated end joining (MMEJ), always results in sequence deletions (Jackson & Bartek, 2010).

HR is, on the contrary, able to achieve faithful repair. As it uses sister chromatid sequences as templates, it is however restricted to the S/G2 cell cycle states, while both NHEJ and MMEJ are operative during the entire cell cycle. After double-strand break generation, various proteins of the MRE11-RAD50-NBS1 complex generate single-stranded DNA at the location of the break, what in turn initiates the invasion of the undamaged template, catalysed by RAD51, BRCA1 and BRCA2. This is followed by the activity of polymerases, nucleases, helicases and ligases that faithfully repair and close the double-strand break. Besides direct DNA repair, HR also contributes, in interaction with the Fanconi Anaemia protein complex, to the de-stalling of replication forks and the repair of interstrand crosslinks (Jackson & Bartek, 2010).

In the last decade, HR deficiency (HRD) has received increasing attention in cancer research. Tumours with HRD have been found to be particularly responsive to cisplatin therapy and Poly(ADP-Ribose)-Polymerase (PARP) inhibition (PARPi; Fong et al., 2009). On a molecular level, PARP is responsible for the repair of single-strand breaks; in its absence, these breaks are not repaired before replication, leading to the accumulation of double-strand breaks during cell cycle progression. As double-strand breaks cannot be faithfully repaired in HR-deficient tumours, cancer cells accumulate more and more damage due to unfaithful repair until the genomic damage ultimately leads to cell death. As *BRCA1* and *BRCA2* are core genes of the HR pathway, cancers with loss-of-function mutations or promoter methylation of *BRCA1/2* are very sensitive to platinum and PARPi therapy. So far, mainly breast, ovarian, and pancreatic cancers with a *BRCA1/2* mutation have been treated with these therapies. However, also cancers without *BRCA1/2* alterations that share key signatures of HRD such as mutations in other HR-involved genes have shown similar sensitivity to these therapies (Waddell et al., 2015). Here, we refer to the presence of HRD in a tumour with *BRCA1/2* wildtype as 'BRCAness', describing a phenocopy of *BRCA1/2* alterations that renders a tumour HR-deficient and therefore sensitive to cisplatin and PARPi treatments (Lord & Ashworth, 2016).

While around 20% of high-grade serous ovarian carcinoma tumours carry *BRCA1/2* alterations, a further 30% show HRD with some of them having mutations in alternative HR-involved genes like *PALB2* and *RAD51*. Several brain, breast and prostate cancers that harbour mutations in the DNA repair gene *PTEN* also showed signs of HRD. Further, patient-derived xenografts from *ATM*-mutated tumours were found to respond to PARPi (Evans et al., 2017); *ATM* has been of particular significance since it is frequently mutated in many cancer types, including gastric, bladder, lung, renal and endometrial cancers (Lord & Ashworth, 2016). Based on this collective evidence, Nik-Zainal et al. (2016) suggested that more comprehensive HRD biomarkers that go beyond *BRCA1/2* mutations are needed for the identification of HRD in human cancers (Nik-Zainal et al., 2016), hypothesising that this might increase the potential number of patients that could benefit from treatment with cisplatin/PARPi (Polak et al., 2017).

While mutations in multiple genes have been associated with BRCAness, most importantly *PTEN*, *PALB2*, *RAD51B*, *BRIP1*, *CHEK2*, *FANCI*, *RAD51C*, and *ATM*, the molecular mechanisms that underpin BRCAness are not understood. In particular, it is not known if all mutations of BRCAness genes have similarly strong effects, and whether some of these mutations are conditional for BRCAness. In addition, we don't understand the cancer type specificity of these genes' effects, and how they interact with *BRCA1/2* alterations. Finally, previously unknown genes may be associated with HRD and their perturbation may lead to sensitivity to cisplatin/PARPi treatment (Lord & Ashworth, 2016; Nik-Zainal et al., 2016).

4.1.2 Molecular signatures of homologous recombination repair deficiency

HRD has been linked to genetic signatures, like accumulation of mutational signature 3 (Alexandrov et al., 2013; Nik-Zainal et al., 2016). The link with Signature 3 has recently been confirmed by Polak et al. (2017) who showed that bi-allelic inactivation of *BRCA1/2* leads to accumulation of this mutational signature. However, not all HRD- and Signature 3-positive tumours harbour these mutations. Polak et al. (2017) showed that these *BRCA1/2*-wildtype patients with Signature 3 mutations partially harbour *RAD51C* genetic and epigenetic mutations, and *PALB2* germline lesions. Therefore, Polak et al. (2017) proposed these alterations as alternative sources of BRCAness, and Signature 3 as a molecular readout of HRD.

Nik-Zainal et al. (2016) assessed the relationship between HRD and signatures of genome-wide rearrangement patterns; in this thesis, we refer to these signatures as SV signatures as opposed to the canonical single base substitution signatures (Section 2.3). Nik-Zainal et al. (2016) found that almost all tumours (91%) of their breast cancer study with loss-of-function mutation or hypermethylation of *BRCA1* harboured an increased amount of SV Signature 3. SV Signature 3 is strongly associated with relatively short tandem duplications (<10k base pairs),

what suggests inactivation of *BRCA1* to be associated with this mutational phenotype (Nik-Zainal et al., 2016).

In addition, Nik-Zainal et al. (2016) found that many cancers with *BRCA1* and/or *BRCA2* mutations showed accumulation of SV Signature 5 that is defined by small deletions (<100k base pairs) . However, some tumours without *BRCA1/2* alterations were also enriched for Signature 3 and SV Signature 5 and further resembled *BRCA1/2* mutants in gene expression patterns (as assessed by hierarchical clustering), possibly pointing towards a contribution of other genes to an HRD-like phenotype. Some tumours without *BRCA1/2* mutations, however, showed different patterns of gene expression, despite extensive signs of Signature 3 and SV Signature 5. An alteration in other genes involved in DNA double-strand break pathways like *ATM*, *ATR*, *PALB2*, *RAD51C*, *RAD50*, *TP53*, *CHEK2* and *BRIP1* could not explain these mutational patterns. Due to these observations, a combined diagnostic score of (SV) signatures and indel patterns was suggested as an alternative or extension to the existing assessment of HRD (Nik-Zainal et al., 2016). Indel patterns were involved since long indels (>10 base pairs) have been shown to be enriched in HR-deficient tumours due to increased NMEJ and MMEJ (see Davies et al., 2017).

Besides (SV) signatures, other molecular features have been used to define HRD. Most recently, Knijnenburg et al. (2018) defined a HRD signature across 33 cancer types in TCGA, taking into account various genomic features like loss of heterozygosity, large-scale state transitions, and number of telomeric allelic imbalances scores using SCNAs. Using Bayesian ridge regression on gene expression data to predict their HRD score, they found contributions of various gene mutations in *BRCA1*/2, *RAD51B*, *RAD51C*, *TP53*, *MLH1*, *MSH2*, and *TCEB3* to HRD, and proposed a potential cisplatin/PARPi sensitivity of tumours from various cancer types. Mutations in one of those genes, the tumour suppressor gene *TP53*, have already been shown to co-occur with *BRCA1* mutations (Lord & Ashworth, 2016). Another approach, HRDetect, used lasso logistic regression and identified six signatures that can distinguish *BRCA1/2*-deficient from sufficient cases, including microhomology-mediated indels, Signature 3, and SV Signatures 3 and 5 (Davies et al., 2017).

While the mutational and epigenetic factors that underpin HRD have been studied in detail (*e.g.*, Alexandrov et al., 2013; Davies et al., 2017; Nik-Zainal et al., 2016; Polak et al., 2017), the relationship between HRD and transcriptional profiles in the corresponding tumours remains elusive. An understanding of the molecular processes that are associated with HRD could improve the precision of detecting HR-deficient tumours and therefore increase the potential number of patients that can be successfully treated with cisplatin/PARPi. In addition, a better understanding of protein pathways linked to HRD could help develop new targets for therapy; this has become particularly important considering the development of PARPi resistance in originally HR-deficient tumours through demethylation of *BRCA1* or *RAD51C* promoters (Montoni et al., 2013).

Among the few existing studies that have considered transcriptional profiles associated with HRD is Larsen et al. (2013), who derived a transcriptional signature that is able to predict *BRCA1/2* germline mutation status in breast cancers. This approach is based on a support vector machine classifier trained on the expression data of gene sets that maximally segregate between *BRCA1/2* and sporadic tumours. A classification of all breast cancer samples, however, distinguished breast cancer subtypes according to the PAM50 classification (Parker et al., 2009) rather than BRCA mutation state. This makes sense since *BRCA1/2* mutation states are strongly correlated with breast cancer subtypes. For example, a vast majority of *BRCA1* mutants exhibit the basal phenotype and are 'triple-negative', *i.e.*, they test negatively for oestrogen and progesterone receptors, and human epidermal growth factor receptor type 2 (HER2), which normally mediate cell growth (Lord & Ashworth, 2016). While subtype-specific *BRCA1/2* state predictions achieved better accuracy and were replicable across datasets, Larsen et al. (2013) did not (i) study the functional relevance of the maximally separating genes to explain the molecular basis of HRD, (ii) check for cancer subtype-specific effects of these genes, nor (iii) search for genes that might be associated with HRD beyond canonical *BRCA1/2* germline cases.

Later, Wang et al. (2017) used the same dataset to propose a BRCA score per patient based on the expression of 600 genes that segregated *BRCA1/2* mutants. This score accounted for clinical covariates like breast cancer subtype, and pointed towards the importance of CN alterations rather than expression of HRD-related genes in BRCA wildtype tumours.

Another transcriptomic study of HRD by Peng et al. (2014) artificially introduced HRD in breast cancer cell lines by RNA-mediated inactivation of known HR-related pathways, and then defined a transcriptional signature based on the top genes separating between parental and HR-deficient cell lines. While Peng et al. (2014) showed that these genes were enriched for cell cycle and DNA replication control, the resulting transcriptomic changes were later shown to probably be a general consequence of the reduced proliferation rate of the cell lines after RNA-mediated inactivation (Wang et al., 2016).

These existing transcriptomic approaches for analysing HRD have in common that they used the expression level of genes that are highly predictive of *BRCA1/2* mutation state or HRD, in order to investigate patterns that differ between HRD-positive and -negative tumours. However, no molecular mechanisms of *BRCA1/2*-independent HRD, *i.e.*, BRCAness, have been suggested, and, to the best of our knowledge, no transcriptome-wide pan-cancer studies of both, HRD and BRCAness, have been conducted so far.

4.2 A PCAWG gene expression-mutagenesis screen

As the role of gene expression in understanding and detecting HRD and BRCAness across cancer types might have been underappreciated, we set up an expression-mutagenesis screen

to analyse molecular aberrations of HR-deficient tumours in the PCAWG dataset. We first assessed the importance of gene expression for HRD detection in tumours while taking differences between cancer types into account. We then studied these gene expression-HRD associations to explore the molecular basis of BRCAness by adjusting for mutational and epigenetic alterations and germline predisposition of the BRCA genes.

This gene expression-HRD study is based on linear mixed models (Section 1.2.2) and is guided by the gene expression-mutational signature screens of Section 2.3. The study is based on the WGS and RNA-Seq datasets of 1,159 patients of the PCAWG cohort that we used in Chapter 2, and that span 27 human cancer types (see Sections 2.2 and 2.3.1 for a detailed description of the dataset).

4.2.1 A molecular marker of homologous recombination repair deficiency

First, we had to define putatively HR-deficient samples in the PCAWG dataset. We hereby considered the following potential molecular markers of HRD that were quantifiable in the PCAWG dataset: Signature 3 (see Alexandrov et al., 2013; Nik-Zainal et al., 2016; Polak et al., 2017); long indel (>10 base pairs) count that is increased in HR-deficient tumours due to increased NMEJ (see Davies et al., 2017), and short deletions and tandem duplications (<10k base pairs) that are the characteristics of HRD-associated SV Signatures 3 and 5 (see Nik-Zainal et al., 2016).

Briefly, we designed a combined biomarker to postulate HRD presence or absence based on evidence of an increased number of long indels and short SVs. We chose these molecular markers since their relationship with HRD can be explained in a mechanistic manner: Long indels are the result of non-faithful repair of DNA double-strand breaks due to the lack of HR, and short SVs are strongly correlated with *BRCA1* down-regulation (Nik-Zainal et al., 2016). Whereas Signature 3 has previously been used as a marker for BRCA deficiencies, it does not provide mechanistic insights into the origin of the mutational patterns due to its definition by statistical decomposition (NMF; see Section 2.3.1), and is therefore less robust against potential correlated factors that are not directly linked to HRD. We defined patients with an increased ratio of long indels of more than 15%, and at least one increased ratio of short SVs of more than 10% (either tandem duplications or deletions) as HRD-positive. With these criteria, we identified 102 HRD-positive patients amongst our cohort of 1,159 patients. We encoded these HRD-positive patients as '1' in a binary HRD score; HRD-negative patients were encoded as '0'.

An explorative analysis of this set of 102 putative HRD-positive patients revealed known HRD-related signatures: The majority of our HR-deficient cases were ovarian or breast cancer patients, with 46 ovarian and 30 breast tumours. Breast and ovarian, together with prostate and pancreas, cancers are known to be the cancers that show HRD most frequently. Of the breast

tumours, the majority (n=25) showed the basal signature subtype; this subtype is known to be particularly prone to HRD (Lord & Ashworth, 2016; Chartron et al., 2019; Telli et al., 2018). The remaining patients had pancreatic (n=7), liver (n=3), bladder (n=3), stomach (n=3), uterine (n=2) cancer or lymphoma (n=2).

Amongst the 102 patients, 14 HR-deficient patients had a deleterious germline *BRCA1* mutation (out of altogether 16 patients with a deleterious germline *BRCA1* mutation; $P \leq 2x10^{-16}$, hypergeometric test), 6 a deleterious germline *BRCA2* mutation (out of 12; $P=2x10^{-4}$, hypergeometric test), 13 a hypermethylated *BRCA1* promoter (out of 13; $P \leq 2x10^{-16}$, hypergeometric test), 7 a hypermethylated *RAD51C* promoter (out of 19; $P=7x10^{-4}$, hypergeometric test), 23 a somatic *BRCA1* mutation (out of 23; $P \leq 2x10^{-16}$, hypergeometric test) and 4 a somatic *BRCA2* mutation (out of 4; $P=6x10^{-5}$, hypergeometric test). As somatic mutations, we considered missense, frameshift, stop gained, splice acceptor, splice donor, start lost, and inframe deletion mutations. In addition, 36 patients showed significantly decreased *BRCA1* expression (FPKM < 0.5, out of 289; P=0.009, hypergeometric test). Altogether, known HRD-related genetic, epigenetic and gene expression alterations are strongly enriched in HR-deficient patients as compared to all patients.

Deleterious *BRCA1* germline mutations are known to have a strong effect on HRD development (Lord & Ashworth, 2016). We had, however, detected two patients with *BRCA1* germline mutations that were HRD-negative according to our HRD score. One explanation might be that a secondary alteration has re-established *BRCA1* functionality early on. Both patients did not show any evidence of Signatures 3 mutations either, strengthening our hypothesis. Indeed, the two patients (one liver, one cervical cancer) did not show remarkably decreased *BRCA1* expression (FPKM < 0.5).

This first exploration gave us confidence that we were able to distinguish HR-deficient from HR-sufficient patients with the help of our newly defined HRD score.

4.2.2 Molecular deregulation in homologous recombination repair deficient tumours

Leveraging our newly defined HRD score, we assessed associations between HRD and gene expression levels. Initially, we trained a linear mixed model (implemented in LIMIX (Lippert et al., 2014)) to associate the binary HRD score with expression levels of 10,959 protein-coding genes (filtered for minimum expression FPKM ≥ 1 in at least 50% of the patients), accounting for genetic relatedness of the patients, cancer type, sex, age, and sample purity. As previously (Section 2.2), we considered 35 peer factors in the analysis, thereby adjusting for heterogeneity in the gene expression data (see **Figure 13a** for quality control). As stated beforehand (Section 2.3.2), the P-values of a permutation analysis within each cancer type delivered uniformly

distributed P-values, suggesting a calibrated analysis and no strong effect of cancer type on our detected associations (**Figure 13e**).

Altogether, this association analysis identified 834 genes as differentially expressed between HRD-positive and HRD-negative tumours (FDR ≤ 10%). We repeated the same analyses considering the presence or absence of Signature 3; P-values between the two analyses were strongly correlated (Pearson correlation coefficient R=0.59; **b**), indicating consistent associations with gene expression profiles between the two molecular HRD classifications despite the divergent classification of 7% of the samples (12 patients classified as HR-deficient by our HRD score, and as HR-sufficient by Signature 3 presence; 66 patients classified as HR-sufficient by our HRD score, and as HR-deficient by Signature 3 presence).

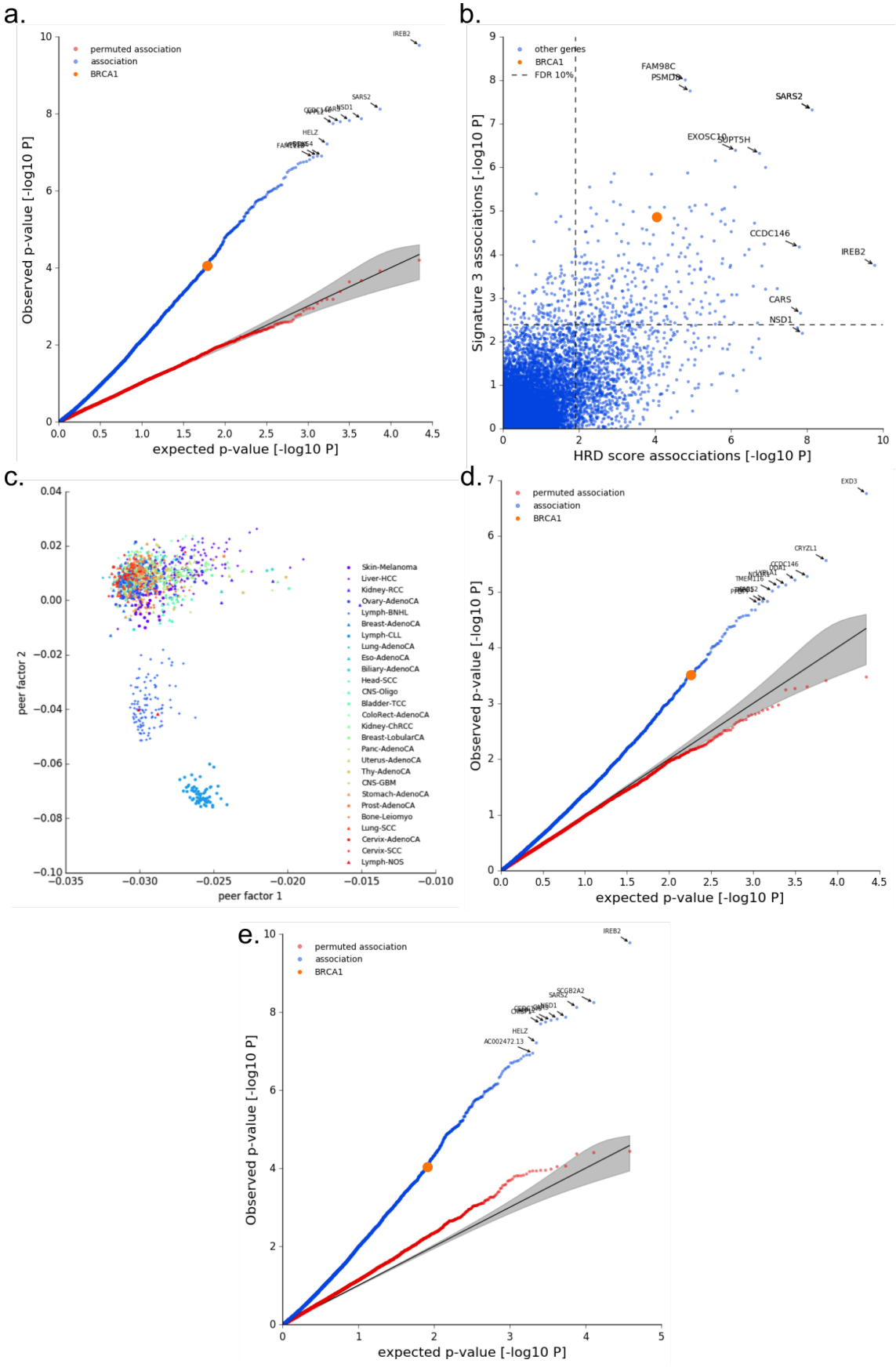

Figure 13. See next page for figure caption.

Figure 13. Quality control of HRD-gene expression screen. **a.** QQ plot of P-values ($-\log_{10}$ P) from a linear mixed model association test between the HRD score and gene expression levels of 10,959 genes across 1,159 patients. Nominal P-values are shown in blue, P-values from a permutation experiment are shown in red. Gene expression values are adjusted for 35 peer factors. The association between the HRD score and *BRCA1* is shown in orange. The ten most significant genes are annotated. **b.** Scatterplot of P-values ($-\log_{10}$ P) of the linear mixed model association study between Signature 3 (y-axis) and gene expression levels, against P-values ($-\log_{10}$ P) of the linear mixed model association study between HRD score (x-axis) and gene expression levels. *BRCA1* is highlighted in orange, the five most significant genes per association study are annotated, and the FDR threshold of 10% is indicated as dashed grey lines. **c.** Data decomposition by the first two of 35 peer factors. Peer factors 1 and 2 segregate the different cancer types. **d.** QQ plot of the P-values ($-\log_{10}$ P) of the linear mixed model associating HRD with gene expression of 10,959 genes across 1,159 patients. Nominal P-values are shown in blue, permuted P-values are shown in red. Gene expression values are adjusted for 50 peer factors. The association between the HRD score and *BRCA1* is highlighted in orange. The ten most significant genes are annotated. **e.** Since cancer type is known to be a strong driver of mutational signature variability, the permutation experiment from (a) was repeated, but patients were only permuted within each cancer type. Nominal P-values are shown in blue, P-values from a permutation experiment are shown in red. The permuted P-values still closely follow the diagonal of the QQ plot.

As expected, our analysis identified significant associations with canonical HR-linked genes, including *BRCA1* ($P=9\text{x}10^{-5}$, $\beta=-0.470$), *RAD51* ($P=0.002$, $\beta=0.381$) and *PTEN* ($P=0.001$, $\beta=-0.383$) (Section 4.1.2). Other expected positive controls, in particular *BRCA2*, had been excluded from the analysis due to low expression across patients. To systematically investigate the relationship of the identified gene set with known HRD-related genes, we compared our results to genes reported by Zimmermann et al. (2018) who undertook CRISPR (clustered regularly interspersed palindromic repeats) gene editing screens to identify genes that might be involved in PARPi resistance. In three cell lines, namely HeLa, RPE1-hTERT, and SUM149PT, they identified 64, 61, and 116 genes, respectively, for which knockdown resulted in increased PARPi sensitivity. As these genes are candidates for causing HRD, we overlaid our total gene set of 10,959 genes with their total gene set of 15,463 genes, resulting in 9,259 genes. We then studied the relationship between their PARPi sensitivity-related genes and our HRD-associated genes. Hereby, we either considered the genes they had identified per cell line, or the union of genes they had found across cell lines. This enrichment analysis identified a significant enrichment of our gene set in their SUM149PT-related gene set ($P=0.023$; hypergeometric test). We did not find any significant enrichment for the gene sets related to the other cell lines, or for the combined gene set ($\alpha=5\%$). In the case of SUM149PT, we overlapped the HRD-related genes common to both analyses,

namely 688 genes of our gene set with their 95 PARPi sensitivity-related genes. Twelve of these genes, including the genes *BRCA1*, *RAD51*, and *FANCI*, were contained in both gene sets. Interestingly, SUM149PT is a cell line derived from a *BRCA1*-hemizygous triple-negative breast cancer, a cancer subtype that is known to be particularly prone to HRD (Lord & Ashworth, 2016; Chartron et al., 2019; Telli et al., 2018).

To obtain more insight into the functionality our gene set of 834 genes, we performed gene enrichment with Gene Ontology (Ashburner et al., 2000). The 834 genes were indeed dominantly enriched for cell cycle pathways: Gene set enrichment analysis (Subramanian et al., 2005) found positive enrichments of cell cycle and checkpoint related process, DNA replication, P53-independent DNA damage, cyclin-associated events, and condensation of prophase chromosomes. One of the cell cycle checkpoint genes that we interestingly found to be downregulated in HRD-positive tumours was *RAD17* ($P=4x10^{-5}$, $\beta=-0.562$). The phosphorylation of this chromatin-binding protein by the ATR/ATM pathway is required for cell cycle arrest in G2 upon DNA damage. As such, it is believed to be an important checkpoint after DNA damage (Post et al., 2003). We hypothesized that if *RAD17* was downregulated in HRD-positive tumours and hence not available in sufficient quantity, molecular signalling leading to early cell cycle arrest might be disrupted and - instead of cell death or senescence - more DNA damage could accumulate. This involvement of *RAD17* in HRD might point towards indirect effects of cell cycle regulation in HR-deficient tumours. The relevance of cell cycle control during replication for HR has been known since Jackson & Bartek (2010) have discovered that fast replication leads to skipping or diminution of HR activity.

We additionally performed TFBS enrichment analyses to gain more insights into the deregulation of cell cycle processes. Like this, we hoped to identify master regulators of our deregulated cellular processes that might not be obvious on the gene expression level. We therefore applied the iRegulon tool (Janky et al., 2014), which searches for enriched transcriptional regulatory networks (so-called regulons).

Briefly, iRegulon uses *cis*-regulatory sequence analysis of a gene set to detect these underlying regulons. The predictions are based on more than 10k known TF binding motifs and 1,120 experimental ChIP-Seq tracks (provided in the form of position weight matrices), that are employed to search for putative and experimentally validated TF binding up to 20k base pairs around the TSS of each gene. Motif enrichment for any TFBS is measured using a *ranking-and-recovery* process: First, all background genes are *ranked* for each motif by scoring for homotypic clusters across ten vertebrate species. Then, during *recovery* of the set of input genes, enrichment of these genes in each of the motif-based rankings is determined by an AUC that is retrieved from the cumulative recovery curve for the input set, along the whole-genome ranking of each motif. The final AUC is computed based on the top 3% of the rankings for all motifs. The AUC is then normalised across the AUCs of all motif rankings to the so-called enrichment score (ES) of each motif. The ES indicates how large the proportion of the input genes is that can be recovered in the top 3% of a motif-specific ranking. Janky et al., (2014)

propose to use an ES cut-off of 3 that corresponds to a FDR threshold of between 3% and 9% (Janky et al., 2014).

In our HRD-associated gene set, the only strong enrichment (ES ≥ 3) we found was for ETS motifs (assessed by ES per TF, *e.g.*, ES(ETV6)=6.12). The ETS TF family is defined by its binding to the ETS domain, a structurally highly conserved winged helix-turn-helix motif. Although there are ETS subfamilies with specific functions, the inappropriate functioning of ETS TFs in general is assumed to be essential for the development of certain cancer types (Sharrocks, 2001).

In total, 563 out of our 834 (67%) genes were potentially regulated by ETS TFs. We hypothesised that perturbation of ETS TF binding and of its regulatory network might be a consequence of HRD or *vice versa*. Therefore, we tested for associations between HRD and gene expression of the altogether 18 ETS TFs which binding sites we had found to be enriched in our dataset, using linear mixed models as above; *ELF5* was the only gene that was differentially expressed between HRD-positive and -negative cases (P=0.001, β=0.481). ELF5 is an interesting tumour suppressor TF that has been associated with aggressiveness of tumours (Sizemore et al., 2017); it represses *Snail2* expression, which in return induces epithelial-mesenchymal transition and metastasis in breast cancer (Chakrabarti et al., 2012). However, ELF5 is generally overexpressed in luminal breast cancers and confers endocrine resistance (Kalyuga et al., 2012). Increased aggressiveness of tumours due to enhanced carcinogenesis and metastatic spread after ELF5 overexpression has also been observed in mouse models (Gallego-Ortega et al., 2015). Hence, the role of ELF5 seems to be ambiguous: At low levels it helps to repress epithelial-mesenchymal transition, at high levels it increases aggressiveness and metastisation of tumours. Considering this ambivalent role of ELF5 in cancer progression, the relatively weak statistical power with which we were able to associate *ELF5* expression with HRD, and the quite low expression of *ELF5* in our dataset, we concluded that this relationship would require an experimental validation. Such an approach was, however, outside the scope of this thesis.

In addition to the lack of association between expression of ETS TFs and HRD, we could also not detect any correlation between their expression and *BRCA1* expression. The enrichment for ETS TFs motifs might hence stem from increased activity of the proteins, and not from increased expression. To assess ETS activity, we used known high-confident interactions between 16 ETS TFs and genes based on TF motifs to calculate a mean ETS activity score with the help of the Dorothea software (Garcia-Alonso et al., 2018). We then used this activity score as phenotype in a linear mixed model to associate it with our HRD score. This only detected one candidate ETS TF whose activity was associated with HRD, namely ETV6 (out of 16 associated genes, FDR ≤ 5%). Interestingly, *ETV6* has previously been identified as recurrently rearranged genomic region in breast cancers by Nik-Zainal et al. (2016). However, no expression changes were discovered by this study, and the significance of this finding therefore remained elusive. *ETV6* was also identified as a short tandem duplicator hotspot by Menghi et al. (2018). As our

HRD score is partially based on the ratio of short tandem duplications, the observed ETV6 activity change could be attributable to an increase in tandem duplications, an effect cumulating in hotspots like the one of the *ETV6* gene.

In summary, we found some evidence for a relationship between HRD and increased ETS activity. To elucidate this relationship, we set out to replicate our HRD-ETS TF associations in a more homogenous subset, namely in just the breast cancers of our previously analysed dataset (n=87). Hereby, it has to be taken into account that the molecular state of breast cancers is modulated strongly by their hormone and growth factor receptors. *E.g.*, DNA damage repair is affected by estrogen concentrations (Caldon, 2014). Hereby, strong differences can be detected between breast cancers that test positive for estrogen, progesterone, or HER2 receptor. This necessitates stratifying the patients according to these molecular subtypes. We used the PAM50 classification system for stratification (Parker et al., 2009), which is based on known molecular differences between basal, normal, Luminal A, Luminal B and HER2-enriched breast cancers, and which has implications for prognosis and therapy sensitivity (Caan et al., 2014). We therefore analysed gene expression-HRD associations based on all PCAWG breast cancers while taking PAM50 classification into account. Altogether, 232 genes were significantly associated with HRD (FDR ≤ 10%), 151 of which were also associated with HRD in the full PCAWG analysis. The 232 genes were, however, only weakly enriched for ETS binding motifs (ES across ETS TFs < 2). Due to the small sample size of the analysed dataset, we were not able to determine if the reduced enrichment for ETS TFBSs was due to decreased power, or if the more homogenous dataset resulted in a true biological disappearance of binding site enrichment, pointing towards data heterogeneity as source of the ETS TFBS enrichment.

To distinguish between these two possibilities, we repeated the analyses in an independent cohort of 560 breast cancer patients from Nik-Zainal et al. (2016). Here, Signature 3 was the only available biomarker of HRD. We therefore associated the binary signature (signature count > 0 was set to 1, *i.e.*, presence of signature) with gene expression changes (gene expression pre-processing as in PCAWG dataset, see Section 2.2). We then employed a linear model based on 252 patients and 8,972 genes. We only considered the 252 patients for which mutational signatures, gene expression data, sample purity, age, gender and PAM50 classification were provided, and 15,064 genes that were observed in all 252 patients. The genes were then filtered to overlap with our previously used set of 10,959 genes (Section 4.2.2), resulting in the final number of 8,972 genes. Altogether, 45 genes were significantly associated with Signature 3 (FDR ≤ 10%), with *BRCA1* being the only HR-related gene. As expected, *BRCA1* was downregulated in HR-deficient samples ($P=2x10^{-4}$, $\beta=-0.690$). Of the 45 genes, 20 contained ETS binding motifs what led to a substantial enrichment of ETS TFBSs, with a maximum of ES(ETV6)=4.93. We therefore concluded that the enrichment of ETS TFBSs in HRD-related genes was a true biological finding that applied across cancer types. Without the inclusion of the PAM50 classification in our linear mixed model, no significant genes were detected at all, emphasising the importance of subtype stratification.

As Nik-Zainal et al. (2016) had identified *ETV6* as a recurrently rearranged genomic region in their dataset, we additionally specifically investigated the tumours with *ETV6* rearrangement in their dataset. Here, we focused on 553 samples with available mutational signatures. Three of these samples harboured fusions, and four deletions in the *ETV6* genomic region. Two of the three fusion and three of the four deletion samples showed evidence of Signature 3 (of altogether 153 samples with positive Signature 3; P=0.02 for a combination of fusion and deletions, hypergeometric test). This signal was even stronger when we considered quantitative Signature 3 counts (P=0.003, Mann-Whitney-U-Test), with the top 30%-quantile samples of Signature 3 (n=168) comprising all seven *ETV6* fusion and deletion tumours. As this pointed towards a contribution of *ETV6* genomic rearrangements to HRD, we investigated the same relationship in the PCAWG dataset. However, only two PCAWG samples showed *ETV6* fusions. While one of these samples was HRD-positive, and additionally a basal breast cancer with decreased *BRCA1* expression, this was quite weak evidence for an interplay between *ETV6* rearrangements and HRD due to the small sample size. However, we found 16 *ETV6* deletions, 5 of which were HRD-positive (P=0.010, hypergeometric test). The PCAWG study therefore provided further evidence of an association between *ETV6* genomic rearrangements and HRD. However, with such small numbers of rearrangement events and without longitudinal data of cancer progression, we could not infer if these rearrangement contributed to HRD, or if HRD led to the accumulation of these mutations.

In the cohort of 560 breast cancer patients from Nik-Zainal et al. (2016), neither gene expression nor activity of ETS TFs was associated with Signature 3. We therefore concluded that although the ETS regulatory network had an effect on HRD, individual ETS TFs only exerted a weak effect. Altogether, our analyses pointed towards a potentially interesting involvement of TF network deregulation in HRD progression, a signal that can be followed up in larger and more homogenous datasets that will probably be acquired in the future.

In summary, the functional assessment of our set of HRD-associated genes pointed towards deregulation of generic cell cycle processes and potential involvement of ETS TF activity.

4.2.3 Effects of data heterogeneity

Whereas we found interesting functional enrichment of the 834 HRD-associated genes (Section 4.2.2), we did not discover any enrichment for DNA repair or HR pathways. Furthermore, overall surprisingly many genes (7.6% of all tested genes) were significantly associated with our HRD score, while of the 16 genes that Knijnenburg et al. (2018) defined as core HR-associated, only *BRCA1* and *RAD51* were significantly associated. On the one hand, HRD might truly affect many genes and broad pathways, and known HRD-related genes might exacerbate their effect via different mechanisms than transcriptomic changes. On the other hand, the deviation of the detected gene expression deregulation from known molecular HRD signatures let us doubt whether our gene set truly described deregulation related to HRD. Therefore, we decided to

investigate in detail if the detected gene expression associations were truly related to HRD or if they could be explained by hidden sources of variation in our data. We did so by investigating how modelling variable numbers of hidden confounding factors in the gene expression data would affect our HRD-gene expression association analyses.

We firstly assessed which sources of variation were captured by the 35 peer factors we had initially adjusted for (Section 4.2.1). We therefore correlated all peer factors against all known patient covariates (assessed by Pearson correlation coefficients and P-values; we here define significant correlations as correlations with a nominal P-value < 10^{-5}). The most significant correlations were found between cancer types and altogether 23 peer factors. Separation of cancer types by the 1st and 2nd peer factor is shown in **Figure 13c**. A strong cancer type effect is hereby expected since gene expression profiles differ substantially between tumours of different tissues of origin. Multiple other significant correlations with age, sex and sample purity were found. For altogether five peer factors, we did not find significant correlations with any known covariate, suggesting that these peer factors modelled unknown hidden structure of the data.

This reveals an inherent problem of estimating confounders in highly heterogeneous data: How can the number of hidden factors an analysis should optimally adjust for be determined when some of the confounding covariates and the true signal are unknown? If the number of peer factors is too large, we correct for molecular signal that might be causal for HRD. If the number of peer factors is too small, we detect spurious associations with HRD that can be explained by sub-structure of the data, *e.g.*, by molecular cancer subtypes.

To systematically investigate the effect of different number of peer factors on our results, we repeated the HRD-gene expression association analyses with different numbers of peer factors (ranging from n_p=10 to n_p=300 with a step size of 10). In general, we observed a substantial effect of the number of peer factors on our analysis (*e.g.*, in terms of number of significant genes). As individual gene-HRD associations varied strongly in P-value and effect size, we compared functional enrichments of the significant gene sets in order to understand the effect of the number peer factors (**Figure 14**). We firstly assessed how enrichment of the significant genes (FDR ≤ 10%) for the Gene Ontology category 'DNA repair regulation' varied across the different number of peer factors n_p. We chose this functional category since we expected true HRD-related genes to be enriched for it, while preventing a too narrow definition of HRD-related processes solely based on previously discovered associations. The strongest enrichment for 'DNA repair regulation' was identified for n_p=50 peer factors (**Figure 14a**). The majority of the analyses with n_p>70, however, did not result in any significant genes (FDR ≤ 10%). This might introduce a bias into enrichment analysis with enrichment for a specific category depending strongly on the significance threshold arbitrarily chosen in the association analyses. We therefore conducted gene enrichment of the top 120 (~1%) of all genes per peer factors (genes sorted according to increasing P-value). The enrichment for the 'DNA repair regulation' category remained similar across numbers of peer factors (**Figure 14b**), confirming that the different functional enrichments were not only due to a lack of power in some of the analyses.

To ensure that the gene set of the optimal number of peer factors (n_p=50) did not maximise enrichments for other random pathways as well, we decided to repeat the peer factor-dependent enrichment analysis across a complete set of independent pathways. As the organisation of Gene Ontology as a directed acyclic graph inherently defines ontologies in dependence to each other (Ashburner et al., 2000), we used the Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation of proteins with altogether 530 pathways (Kanehisa et al., 2016). Again, we used the top 120 (~1%) of all genes per gene set to avoid bias due to arbitrarily chosen significance thresholds. As we were only interested in the relative enrichment of genes across pathways, we did not apply multiple testing correction across pathways what would have resulted in no significant enrichment at an FDR of 5%. Even based on nominal P-values, only few KEGG pathways were enriched for at least one gene set ($P \leq 0.1$; **Figure 14c**). The two most significant pathways were 'Homologous Recombination' and 'DNA replication'. Both were enriched at n_p=50 and n_p=60 (P=0.03 and P=0.04, respectively; **Figure 14c**). These enrichments point towards HRD-related processes: Homologous recombination constitutes a central mechanism of HR that allows for faithful repair of DNA double-strand breaks by sister chromatid invasion. HR further depends on DNA replication since it requires sister chromatids to be available during S and G2 phase of the cell cycle. Enrichment for cell cycle regulation was only detected to be significant at n_p=20 (**Figure 14c**). We note that this observation differs from our previous Gene Ontology enrichment that showed cell cycle regulation enrichment at n_p=35. This difference might be due to the inherent difference between Gene Ontology categories and KEGG pathways, or due to the different gene sets used: In the latest analysis we only used the top 1% (n=120), whereas the Gene Ontology enrichment was based on all significant genes (n=834 at FDR ≤ 10%).

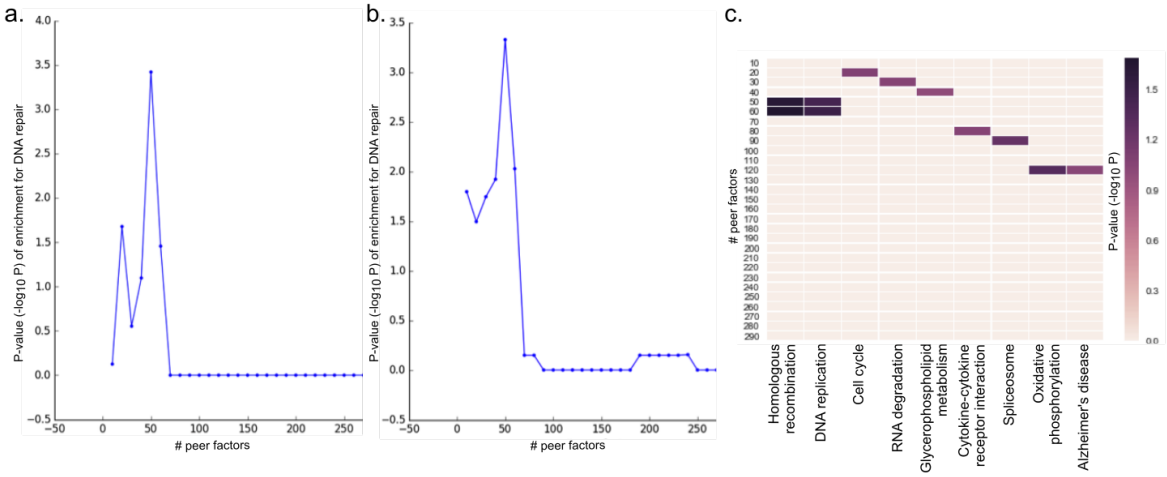

Figure 14. Functional enrichment of HRD-associated genes in dependence of the number of peer factors. **a.** Enrichment (-$\log_{10}$ P) of HRD-associated genes (FDR ≤ 10%) for the Gene Ontology category 'DNA repair regulation' across peer factors. **b.** Enrichment (-$\log_{10}$ P) of HRD-

associated genes (top 120 genes sorted according to increasing P-value) for the Gene Ontology 'DNA repair regulation' across peer factors. **c.** Heatmap of the enrichment ($-\log_{10}$ P) of HRD-associated genes (top 120 genes sorted according to increasing P-value) for all 530 KEGG pathways across peer factors; only KEGG pathways with P ≤ 0.1 for at least one gene set are shown. The heatmap visualises P-values > 0.1 as P=1.

Based on these observations, we considered 50 peer factors as optimal number of peer factors that maximised the HRD signal in the gene expression data (in the case of both, GO and KEGG functional categories). The number of significant genes (FDR ≤ 10%) dropped considerably: Instead of 834 genes in the analysis with 35 peer factors, we now detected 110 genes (**Figure 13d**). The most significant gene in the n_p=50 analysiswas *EXD3* (P=2x10^{-7}, β=-0.631), a gene that is involved in 3'-5' exonuclease activity (Ashburner et al., 2000). Within cancer research, its overexpression has been associated with lower risk in endometrial cancer (Uhlén et al., 2005), but it has not yet been linked to HRD.

As expected, the gene set was enriched for HRD-related pathways since the number of peer factors had been determined with the objective to maximise this enrichment: After multiple testing across pathways, the genes were enriched for the Gene Ontology categories 'Double-strand break repair' (P=2x10^{-4}) and 'Strand displacement' (P=4x10^{-4}), the KEGG pathways 'Homologous recombination' (P=5x10^{-4}) and 'DNA replication' (P=1x10^{-3}), and also for *BRCA1*-binding TFs (P=3x10^{-5}). We were also able to replicate our findings concerning enrichment of ETS TFBSs, although with weaker signal (*e.g.*, ES(ETV6)=3.16) than in the association analysis that was based on 35 peer factors (Section 4.2.1).

To directly compare our association analyses that adjusted for 35 and 50 peer factors, respectively, we correlated the resulting P-values across genes (**Figure 15a**). The correlation plot clearly shows an increased number of significant genes in the analysis with 35 peer factors, partially due to a shifted FDR threshold due to different P-value distributions. The correlation plot, however, also shows that the association results across genes are strongly correlated (Pearson correlation coefficient R=0.607; **Figure 15a**). We concluded that both analyses seemed to describe the same molecular process, with the analysis based on 35 peer factors tagging a larger number of genes and emphasising more general deregulation on the cell cycle level, and with the analysis based on 50 peer factors defining a more precise molecular basis of HRD.

When comparing significant HRD-related genes between the two analyses with n_p=35 and n_p=50, we found *BRCA1* (P=3x10^{-4}, β=-0.441) and *RAD51* (P=4x10^{-5}, β=-0.500) to be significant in both analyses, but *PTEN* to be only significant in the analysis with n_p=35. While *BRCA1* and *RAD51* are direct players of the HR pathway, *PTEN* has been shown to have an indirect effect in *BRCA1*-deficient tumours: It usually antagonises the PI3K/ALT3 pathway that itself leads to cell proliferation and differentiation and is known to be aberrantly active in tumours with *BRCA1*

mutations (Minami et al., 2014). As such, *PTEN* normally acts as a cell cycle checkpoint that allows HR to take place (Mansour et al., 2018). Therefore, the significant association with *PTEN* at n_p=35 that disappears at n_p=50 is another indicator that the former analysis describes cell cycle deregulation indirectly associated with HRD whereas the latter one tags more precise HRD-core processes. Accordingly, a known DNA damage repair gene that was not discovered with 35 peer factors, but with 50 peer factors, is the DNA topoisomerase *TOP3A* ($P=5x10^{-4}$, $\beta=0.424$). While all DNA topoisomerases are involved in DNA repair pathways due to their ability of breaking DNA and changing its topological state, *TOP3A* has been shown to be particularly important since it is able to process different DNA repair intermediates (Yang et al., 2010).

Taking all evidence together, we decided to rely our following analyses on 50 peer factors; this approach seemed to define HRD in a more precise way, what might facilitate our assessment of BRCAness-specific gene expression effects. However, we were not able to rule out that the 35 peer factors may describe the true HRD process more accurately, following the assumption of a broad effect of HRD on cell cycle regulation.

Our analyses emphasises the relevance of adjusting for an appropriate number of confounding factors before conducting association analyses on highly heterogeneous datasets. Here, we proposed a systematic approach to detect the optimal number of peer factors that takes into account prior knowledge of the biological processes in question. While the numbers of peer factors of the two analyses we investigated in detail were quite close to each other (n_p=35 and n_p=50, respectively) and the P-values of the associations across genes were strongly correlated (Pearson correlation coefficient R=0.607), the set of significant genes at the same FDR threshold differed considerably. We therefore showed that biological conclusions depend on the choice of both, the number of confounding factors and the significance threshold. This is an important topic that should be discussed in every statistical analysis of heterogeneous datasets.

4.3 Alternative pathways of homologous recombination repair deficiency

4.3.1 Molecular differences between BRCA-dependent and BRCA-independent cases

To elucidate the molecular basis of HRD in BRCA-independent, *i.e.*, so-called BRCAness cases, we considered the HRD-association analysis (with correction for 50 peer factors; Section 4.2.3) and modified it to search for BRCA-independent signal. Importantly, we only defined tumours as BRCA cases that (i) harboured known *BRCA1/2* alterations and (ii) showed evidence of HRD (*i.e.*, HRD=1). BRCAness cases, on the contrary, were defined as those HRD

cases that did not harbour any *BRCA1/2* alteration. We used a binary BRCA variable to define BRCA (BRCA=1) and BRCAness (BRCA=0) cases. Hereby, we used two different definitions of BRCA: Our first BRCA definition assigned all germline, somatic and epigenetic BRCA alterations (*i.e.*, germline and somatic *BRCA1/2* cases, and hypermethylated *BRCA1* promoter cases) to be BRCA cases. As somatic mutations, we considered missense, frameshift, stop gained, splice acceptor, splice donor, start lost, and inframe deletion mutations. We will refer to these BRCA and BRCAness cases as 'gBRCA' and 'gBRCAness', respectively ('g' for 'genetic'). Next, we defined 'eBRCA' and 'eBRCAness' cases by using the gBRCA/gBRCAness cases as a basis and additionally assigning tumours with strong *BRCA1* down-regulation (*BRCA1* FPKM ≤ 0.5) to be BRCA cases ('e' for 'expression'). We considered this additional BRCA variable since there has been strong evidence for a relationship between HRD and *BRCA1* downregulation (Zhou et al., 2003).

Firstly, we set out to directly identify molecular differences between BRCA and BRCAness cases, *i.e.*, between HRD cases with and without known *BRCA1/2* alterations. In our altogether 102 HRD cases, many cancer types were represented only a few times. To keep the analysis robust, we decided to restrict the analysis of HRD-positive tumours to breast, ovarian and prostate cancers, which (i) showed larger sample sizes (n > 5), and (ii) are also known to be the cancer types that are most commonly affected by HRD. This filtering resulted in 83 HRD cases. We then classified these cases as either BRCA or BRCAness cases (see above). Altogether, the 83 HRD cases contained 46 gBRCA and 58 eBRCA cases.

We then trained a linear mixed model to associate the BRCA variable with gene expression across the 102 HRD cases, taking age, sex, sample purity, cancer type and kinship into account. At an FDR threshold of 10%, the gBRCA approach did not result in any significant genes. Also no significant deregulation of *BRCA1* could be observed. The most significant gene was *RRAS2* ($P=3x10^{-5}$, β=-0.844). In the case of eBRCA, we found two genes to be significantly deregulated (FDR ≤ 10%), namely *BRCA1* ($P=7x10^{-7}$, β=-1.77) and *RRAS2* ($P=1x10^{-5}$, β=-0.80) (**Figure 15b**). As the downregulation of *BRCA1* was expected due to classification of samples with low *BRCA1* expression as eBRCA, *RRAS2* was the only interesting gene that was differentially regulated between BRCA and BRCAness cancers. While this analysis might suffer from a lack of power due to small sample size and data heterogeneity, the detection of the down-regulation of *RRAS2* in BRCA cases was of potential interest: *RRAS2* is downregulated in BRCA cases as *BRCA1* is, but unlike *BRCA1* it is known as an oncogene: *RRAS2* was identified as one of the 299 cancer driver genes detected by Bailey et al. (2018). It is a Ras-like small GTPase that is involved in signal transduction via association with the plasma membrane, and has been shown to be involved in controlling cell proliferation. Mutations and over-expression of *RRAS2* have been linked to the growth of certain tumours, *e.g.*, of breast and ovarian cancer (Movilla et al., 1999). Ras signalling has also been linked to HR, *BRCA1*-mediated DNA repair and hereditary breast cancer signalling networks (Kalimutho et al., 2017). We therefore hypothesised that transcriptome-based differences between BRCA- and BRCAness-related HRD might depend on the downregulation of *BRCA1* in BRCA cases, and

the upregulation of *RRAS2* in BRCAness cases, and that the combination of both ultimately results in a HRD phenotype. Our gBRCA analyses (see above), however, showed no significance of *RRAS2* after multiple testing, what may be due to either the lack of statistical power or the dependency of *RRAS2* expression on *BRCA1* downregulation.

To search for more genes potentially contributing to HRD independently of BRCA, we decided to increase our sample size, and set out to analyse all of our PCAWG samples, including HRD-negative cases.

4.3.2 Dependence of molecular deregulation on BRCA

We extended our search for BRCA-independent gene associations with HRD to the analysis of the entire dataset (n=1,159), assuming increased power to detect BRCAness-specific signal. We associated our HRD score with gene expression across all patients while accounting for the BRCA variable as a fixed covariate in the linear mixed models (50 gBRCA and 61 eBRCA cases, respectively). We hypothesised that model that accounts for any BRCA signal would allow detect genes that are associated with HRD beyond BRCA effects.

When accounting for the gBRCA variable, we found 9 genes to be significantly associated with HRD (FDR ≤ 10%); when accounting for the eBRCA covariate, we found 17 genes (FDR ≤ 10%). None of these gene sets were enriched for any known pathway. A comparison of the analyses that accounted for gBRCA or eBRCA (**Figure 15c**) showed that the results were very similar (Pearson correlation coefficient between P-values R=0.808), with no significant association between *BRCA1* expression and HRD in both analyses.

We then compared both analyses with our HRD association analysis that did not account for BRCA cases (Section 4.2.3). Again, we observed strong correlation of the P-values (Pearson correlation coefficient between P-values R=0.627 and R=0.508 for gBRCA and eBRCA, respectively). Accordingly, when plotting the P-values against each other, we observed a general reduction in statistical power in the analysis the accounted for BRCA, but no strong systematic effect of the BRCA factor on the associations between individual genes and HRD (**Figure 15d** for gBRCA). Only a few significant genes including *BRCA1*, *MAPK9*, *FGFR3*, *ZFAND1*, *TMEM106B*, *CHD7*, *NGRN*, *TSSC4*, *DCLRE1A*, and *ZNF512* were only significant when the BRCA variable was not accounted for (*i.e.*, $P < 0.001$ in the standard HRD analysis, and $P > 0.1$ in the HRD analysis that accounted for gBRCA; genes are highlighted in violet in **Figure 15d**). *BRCA1* and *DCLRE1A* are known players of DNA repair pathways, *TSSC4* acts as a tumour suppressor, and the remaining genes are involved in cell growth, regulation of cellular stress or metastasis. While regulation of these genes seemed to depend on the effect of BRCA, we did not detect any genes that were not significantly associated with HRD but gained significance when the BRCA covariate was accounted for. We therefore concluded that the cancer transcriptome did not show any evidence of alternative molecular processes unique to

BRCAness tumours, and that HR-deficient tumours in our dataset were dominantly driven by the deregulation of *BRCA1/2* processes.

We repeated the analysis of the effect of the BRCA variable in gene expression-HRD screens (i) per cancer type (in cancer types with a sample size n > 10), (ii) only in European samples, (iii) only in solid tumours, or (iv) after correcting the gene expression data for 35 instead of 50 peer factors. All of these approaches resulted in the same observations, *i.e.*, a decrease of power in gene expression-HRD signals when accounting for BRCA cases, but no strong systematic effect.

As directly assessing gene expression associations with BRCAness did not point towards a clear mechanism of BRCA-independent HRD, we studied the mutation states of the 110 HRD-associated genes and how they differed between BRCA and BRCAness cases. We hypothesised that if certain genes showed enrichments of somatic mutations in HRD-positive tumours, they might have a strong impact on HRD independently of BRCA mechanisms.

We firstly queried the somatic mutation state of the 110 genes across our 1,159 samples, and investigated if their somatic mutagenesis was increased in HR-deficient samples. We would then be able to check if these genes also showed different somatic mutational load between BRCA and BRCAness cases. As somatic mutations, we considered missense, frameshift, stop gained, splice acceptor, splice donor, start lost, and inframe deletion mutations. Altogether 107 genes were somatically mutated in 405 samples. The somatic mutations of these genes were enriched in HR-deficient samples (56 out of 102 HR-deficient patients and 349 out of 1,057 HR-sufficient patients were mutated in at least one of the genes; $P=1x10^{-5}$, hypergeometric test). To ascertain that the HR-deficient samples did not show enrichment of somatic mutations in any set of genes, we randomly sampled 110 genes from our gene set and studied their enrichments in HR-deficient samples. We repeated this random sampling 10k times and assessed how often enrichments of somatic mutations in HR-deficient samples were significant (**Figure 15e**). By comparing the empiric distribution of P-values with the P-value from our original gene set (green line in **Figure 15e**), we found a marginal significance of enrichment of somatic mutations of HRD-related genes in HR-deficient samples (P=0.062). We concluded that somatic mutagenesis of the HRD-related genes was to some extent associated with HRD development. When repeating the analysis per gene, however, only *BRCA1* and *SEPP1* showed significant enrichment of somatic mutations in HR-deficient tumours (at a significance threshold of $\alpha=0.1$). *BRCA1* again pointed towards importance of BRCA factors for HRD, even on a genomic level. *SEPP1* is known to be involved in oxidative stress, but has not been implicated with HRD yet. We still checked if these genes were differentially somatically mutated between BRCA and BRCAness cases (based on both definitions of BRCA cases, as gBRCA or eBRCA). However, such an approach did not make sense for the *BRCA1* gene since our definition of BRCA and BRCAness cases was inherently based on the somatic mutation status of *BRCA1/2*. In the case of *SEPP1*, we saw no significant difference in mutational load between BRCA and BRCAness cases, neither in the case of the gBRCA nor in the case of the eBRCA definition. In addition,

Welcsh et al., (2002) found that *SEPP1* expression is dependent on *BRCA1* induction, so *SEPP1* might just be another player of the complex BRCA-regulated HRD pathways. We therefore had to conclude that even on the somatic mutational level, BRCA alone seemed to be driving HRD and that we were not able to discover a gene that may contribute to BRCAness through somatic mutagenesis.

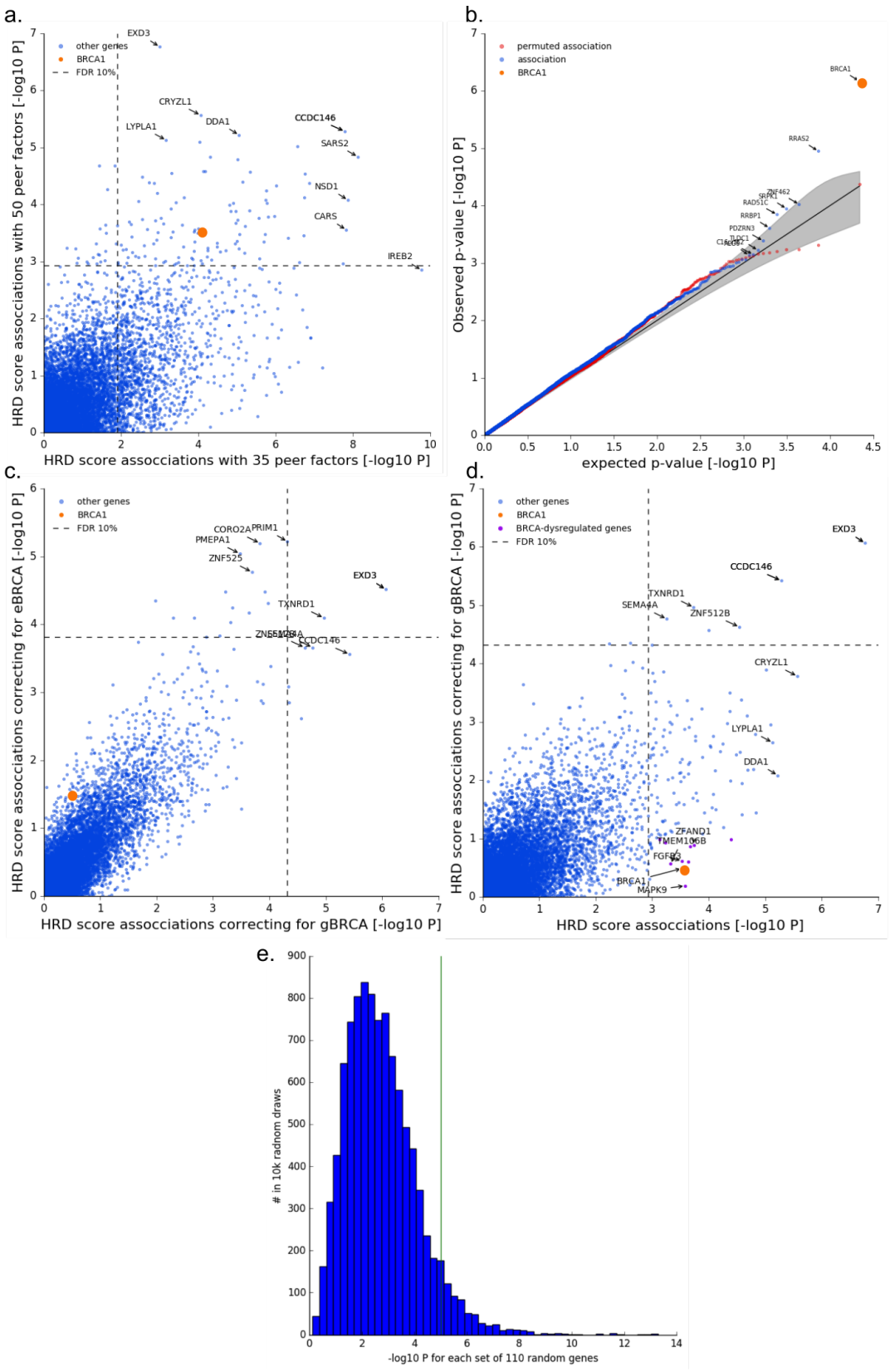

Figure 15. See next page for figure caption.

Figure 15. Results of linear mixed models to assess the association between gene expression and BRCAness. **a.** Correlation between P-values ($-\log_{10}$ P) of the linear mixed model associating HRD score with gene expression after correcting for 35 (x-axis) or 50 (y-axis) peer factors. *BRCA1* is highlighted in orange, the five most significant genes per association study are annotated, and the FDR threshold of 10% is indicated as dashed grey lines. **b.** QQ plots of the P-values ($-\log_{10}$ P) of the linear mixed model associating expression of 10,959 genes with BRCA state across 102 HR-deficient samples. Nominal P-values are shown in blue, permuted P-values are shown in red. Gene expression is corrected for 50 peer factors. The ten most significant genes are annotated. **c.** Correlation between P-values ($-\log_{10}$ P) of the linear mixed model associating HRD score with gene expression after correcting for gBRCA (x-axis) or eBRCA cases (y-axis). *BRCA1* is highlighted in orange, the five most significant genes per association study are annotated, and the FDR threshold of 10% is indicated as dashed grey lines. **d.** Correlation between P-values ($-\log_{10}$ P) of the linear mixed model associating HRD score with gene expression after not accounting for any BRCA cases (x-axis) or after accounting for gBRCA cases (y-axis). *BRCA1* is highlighted in orange, the five most significant genes per association study are annotated, and the FDR threshold of 10% is indicated dashed grey lines. Highlighted in violet are the ten genes that were significantly associated with HRD only if no BRCA cases were accounted for ($P < 0.001$ in HRD analysis, and $P > 0.1$ in HRD analysis accounting for gBRCA). **e.** Histogram showing the frequency of P-values of the enrichment of somatic mutations of 110 random genes in HR-deficient samples. The green horizontal line represents the enrichment P-value of the set of 110 genes that was associated with HRD in our association analysis.

4.4 Discussion and conclusion

To investigate the relevance of gene expression for the development of HRD and BRCAness in various cancer types, we set up a gene expression-mutagenesis screen to analyse molecular aberrations of HR-deficient tumours in the PCAWG dataset. This section of the thesis extends the analysis of the relationship between genetic variation and gene expression to the investigation of a clinically relevant genetic signature that is not completely understood yet, neither on the level of its genetic determinants nor on the level of its molecular causes and consequences. We investigated how a gene expression-mutagenesis screen might be leveraged to obtain biologically meaningful and potentially clinically useful information. Therefore, we first studied associations between gene expression and HRD in tumours while taking differences between cancer types into account. We then used the gene expression-HRD associations and corrected for mutational and epigenetic alterations and germline predisposition of the *BRCA1/2* genes to understand the molecular basis of BRCAness, a postulated BRCA-independent mechanism of HRD.

In the gene expression-HRD screen, we found a strong difference in expression levels of more than 800 genes between tumours that did and did not show evidence of HRD. While some of these genes were known to play a role in HR-related processes, other known HRD genes were not significantly deregulated. Some HRD-related genes might therefore exert their function via alternative pathways, including protein activity, translational efficacy, interactive effects with other transcripts, or positional regulation.

Overall, the strongest molecular signal of HRD on the transcriptomic level was the deregulation of cell cycle-related processes (when accounting for 35 peer factors in the association analysis). This relationship was biologically meaningful considering that cell cycle control during replication is an important component in HR. For example, fast replication leads to skipping or diminution of HR activity since the faithful repair of DNA double-strand breaks can only take place in the S/G2 cell cycle states. This finding also confirms the results by Peng et al. (2014) who showed that genes deregulated in breast cancer cell lines with inactivated HR functionality were enriched for cell cycle and DNA replication control. However, their results were challenged by Wang et al. (2016) who contested the study design by Peng et al. (2014) and proved that the gene expression changes that were attributed to HRD in Peng et al. (2014) might be a general consequence of the reduced proliferation rate of the cell lines after RNA-mediated inactivation. As our study was based on *in vivo* tumour tissues and not on perturbed artificial systems, we were, on the contrary, able to confirm the biological importance of cell cycle regulation for the development of HRD.

Moreover, we found *RAD17* to be downregulated in HR-deficient tumours, an interesting cell cycle-associated gene. This cell cycle checkpoint gene is required for eventually activating cell cycle arrest in G2 upon DNA damage. We therefore hypothesised that cell cycle arrest might be disrupted in the case of downregulation of *RAD17*, leading to the accumulation of more DNA damage instead of regulated cell death or senescence.

Another molecular signal we found to be associated with HRD was the enrichment of ETS-TFBSs in HR-deficient tumours, a finding that we replicated in two more homogeneous datasets. However, only ETS TF ELF5 was associated with HRD on the transcriptomic level. Due to the weak signal of this association and the generally complicated role of ELF5 in tumour progression, we decided to assess the estimated activity of ETS TFs instead of their gene expression. While we did not find an association between HRD and ELF5 activity, we discovered a significant relationship between HRD and activity of ETV6, a gene with a recurrently rearranged genomic region in breast cancers and a known tandem duplicator hotspot across various cancer types. As our analyses detected a differential activity of ETV6 binding between HRD-positive and HRD-negative tumours, we hypothesised that the activity change might be attributable to an increase in tandem duplications captured by our HRD score, an effect that would ultimately culminate in duplicator hotspots like the one located in the *ETV6* gene itself. In addition, both the breast cancer dataset from Nik-Zainal et al. (2016) and the PCAWG dataset pointed towards an association between genomic rearrangements of *ETV6*

and HRD. However, as we lacked a substantial number of rearrangement events in our datasets and longitudinal data of cancer progression, we could not infer if these rearrangement had contributed to HRD, or if HRD had led to the accumulation of these mutations.

While these ETS TFBS enrichments were of potential interest to understand HRD, we were not able to assess their functional significance. A possible experiment to elucidate their role in HRD would be to modulate ETS expression or activity via knockdowns in cell lines and investigate the molecular consequences. A candidate could be the ETS TF ERG, which has been shown to have vast effects on the molecular composition of the prostate cancer cell line VCaP (Tomlins et al., 2008).

In summary, this first gene expression-mutagenesis screen predicted that genes segregating HRD-positive and -negative patients were enriched in general cell cycle regulatory processes. Many canonical HRD-related genes were not deregulated, and we did also not discover an involvement of direct DNA repair or HR pathways. To ensure that we captured true HRD-associated gene expression changes, we extended our association analyses to assess the effect of the number of hidden factors we adjusted for. We showed that the estimated number of confounders had a substantial impact on the association studies, and proposed a systematic approach to detect the optimal number of hidden factors that takes into account prior knowledge of investigated biological processes. We concluded that biological inferences strongly depend on the choice of both, the number of confounding factors and the significance threshold. We chose to account for more hidden factors (increasing the number of peer factors in our analysis from 35 to 50) since we thus discovered a clear molecular signal of HRD and DNA repair deregulation in our associated genes. However, by correcting for more hidden factors, we might have accounted for biologically meaningful associations that were linked to cell cycle deregulation in HR-deficient tumours. Whereas many association studies like QTL mapping approaches rely on estimating the number of confounders by maximising the power of their analyses, finding a biologically meaningful way of determining the optimal number of hidden factors in highly heterogeneous data still poses a difficult problem: If the number of hidden factors is too large, important molecular signal might be adjusted for. If not enough hidden factors are accounted for, spurious associations explain sub-structure of the data and not the true biological association. We propose that this important topic should be discussed in every statistical analysis of heterogeneous datasets.

We then used our gene expression-mutagenesis screen (firstly based on 50 peer factors, and then replicated with the analysis based on 35 peer factors) to study the BRCA-independent association between HRD and gene expression changes. Our first analysis was based on only HRD-positive patients. This revealed *RRAS2* and *BRCA1* to be downregulated in BRCA cases as compared to BRCAness cases. *RRAS2* is a known cancer driver gene that is involved in Ras signalling which in turn has already been linked to HR. A hypothesis to describe transcriptome-based differences between BRCA- and BRCAness-dependent HRD might therefore postulate the downregulation of *BRCA1* in BRCA cases, and the upregulation of *RRAS2* in BRCAness

cases, both molecular processes that ultimately result in a HRD phenotype. Again, to obtain a deeper understanding of the role of *RRAS2*, experimental studies like functional knockdowns would be necessary to confirm or reject our findings.

In an analysis of all PCAWG tumours including HR-sufficient patients that accounted for BRCA-dependent effects on HRD-gene expression associations, we only detected few significant genes. These genes were neither enriched for any biological function nor - as far as we know - functionally important on their own (except for *BRCA1*). This pointed towards a strong and possibly prerequisite role of *BRCA1/2* alterations in HRD development. As we aimed to find BRCA-independent molecular signal, we conducted more analyses to ensure the correctness of our findings: We conducted the same analyses (i) per cancer type (in cancer types with a sample size n > 10), (ii) only in European samples, (iii) only in solid tumours, or (iv) after correcting the gene expression data for 35 instead of 50 peer factors. All of these approaches resulted in the same strong decrease of power in gene expression-HRD signals after accounting for BRCA cases. Altogether, we concluded that HR-deficient tumours in our dataset were dominantly driven by the deregulation of *BRCA1/2* processes. Other pathways which aberrations have previously been associated with HRD (*e.g.*, Polak et al., 2017) may ultimately act by impairing BRCA activity in ways that are not detectable from DNA sequence and gene expression data. Indeed, most previously detected genes act upstream of *BRCA1/2* in double-strand break recognition or HR-related pathways (e.g., *PALB2* and *RAD51C*). Also ETS TFs have previously been linked to *BRCA1* activity: The *BRCA1* promoter contains an ETS TFBS (Suen & Goss, 1999) and Atlas et al. (2000) showed that ETS TFs transcriptionally activate *BRCA1*.

We note that the presented analyses of BRCA-independent HRD are exploratory studies, involving repeated testing on the same fixed dataset. This means that the P-values of significant associations would have to be corrected for multiple testing across all analyses, and that - if we had found more than negative results - further validation of these associations would have to be performed to achieve high confidence in these results.

A repetition of our analyses in more homogenous and larger datasets might, however, lead to different results. At the time that this thesis is being written, the TCGA program (Cancer Genome Atlas Research Network et al., 2013) has assembled various large datasets for many cancer types. However, the sole focus on exonic regions of most of their studies did not allow us to query this data to replicate our results. We conducted preliminary pre-processing and analyses of the TCGA-BRCA breast cancer cohort (n=1,098 patients; Cancer Genome Atlas Network et al., 2012), which showed us that the quantity of large indels what we used to define HRD in the PCAWG dataset did not correlate with known HRD-related processes like *BRCA1* downregulation or promoter hypermethylation in this dataset. This might be due to an uneven distribution of large indels between exonic and non-exonic regions of the genome, demonstrating that we would not be able to repeat our analyses on exon-based datasets.

Finally, we would like to point towards a general problem of association studies between gene expression changes and genetic variation, namely the impact of time. In our analyses, genetic signatures that are predictive of HRD, including Signature 3, large indels and *BRCA1* promoter methylation status, might have been inherited from ancestor cells but might not be signs of on-going deregulation associated with HRD. Hence, in certain tumours, no gene expression alterations due to HRD might be expected despite the detection of genetic or epigenetic evidence. Notably, PARPi resistance has developed in originally HR-deficient tumours through demethylation of *BRCA1* or *RAD51C* promoters (Montoni et al., 2013). In such cases, we might erroneously define a tumour as HR-deficient which would bias our association studies.

5 Concluding remarks

In this thesis, I investigated how functional associations between genetic variation and gene expression alterations can be studied in highly heterogeneous molecular genetic datasets using statistical models (Section 1). In addition to assessing biologically meaningful associations in large-scale data, I showed how causality between genetic and molecular variation can be established with the help of germline variation, and how heterogeneity might lead to spurious associations. The correction for confounding factors that might be readily available as covariates of the dataset (*e.g.*, cancer type in our pan-cancer analysis, Section 2) or that are hidden in the data (*e.g.*, effects of tumour sample purity or cancer subtypes, Sections 2 and 4) is hereby a prerequisite for all statistical association studies. This includes accounting for inherent population structure in genomic datasets, what we achieved by modelling the kinship between individuals as a random effect in linear mixed models (Sections 2 and 4). Further statistical model regularisation is essential to not over-fit the data and to prevent the effect of multi-collinearity, a problem we had to deal with in our alternative splicing analyses on single-cell transcriptome data: Here, we integrated the genomic features that were highly correlated with each other per cassette exon by applying regularised regression, namely ridge regression (Section 3). We showed that alternative statistical approaches like CNNs can improve modelling non-linear dependencies within complex data (Section 3). These approaches might not only improve modelling performance, but also extend the catalogue of biological conclusions we can potentially draw from s dataset; in the case of predicting splicing from DNA sequence, we could query the model to detect local sequence motifs that are putatively responsible of altered splicing rates (Section 3). Finally, we emphasised throughout this thesis that replications in independent datasets are a crucial prerequisite for ensuring the credibility of association studies, and of scientific findings in general.

In the first section of this thesis, I introduced statistical methods that have been developed for and applied to genomics data to face the described challenges. I introduced concepts concerning genetic variation, and gene expression as a proximal readout for molecular variability, and presented human cancer as a biological system that can be used to assess the applicability and efficiency of these statistical methods.

In the second section of this thesis, I analysed the highly heterogeneous, cancer type specific regulatory landscape of human cancers. This study system allowed us to investigate multi-dimensional associations between gene expression and genetic variation from various sources, including local and genome-wide, germline and somatic, protein-coding and non-coding genetic variants. Using joint modelling, we explored the impact of these variants on gene expression.

Collectively, our analyses of the PCAWG genomic and transcriptomic data provided a comprehensive picture of the regulatory landscape of cancer tissue, and of how different germline and somatic variations alter gene expression levels in a pan-cancer setting. We

showed that co-regulation of the same genes by multiple different types of genetic variants is common in cancer, and we simultaneously reported the relative magnitudes of these different effects. Previous studies have been limited by the lack of WGS data, which is essential for identifying contributions of non-coding variants to gene expression variability. Indeed, our analyses of the currently largest cohort of matched tumour WGS and RNA-Seq data of 1,188 patients identified previously underappreciated associations between somatic regulation in distal regulatory elements and gene expression. Besides emphasizing the effect of local non-coding variation on gene expression, we established associations between genome-wide mutational processes and gene expression, and elucidated cause and consequence of this relationship by taking into account the effect of germline variation. Importantly, this allowed us to derive *de novo* functional annotations of mutational signatures with previously known aetiologies. Our results hence pointed towards non-coding somatic deregulation as a functional driver of carcinogenesis beyond mutations of the coding sequence.

While these results are a valuable resource to better understand cancer heterogeneity, it will be important to conduct similar analyses per cancer type as soon as respective dataset will be available. Although we corrected for cancer type specific effects, associations that only occur in few cancer types will inherently be confounded by tissue type effects and cannot be anticipated in other cancers (PCAWG Transcriptome Core Group et al., 2018).

In the third section of this thesis, I studied the relationship between genetic and epigenetic variation and alternative splicing, and extended the scope of this thesis from associations between genetic variation and gene expression (Section 2) by introducing (i) the functionally relevant phenotype of alternative splicing, (ii) the study of single-cell instead of bulk data, (iii) the impact of epigenetic variation on molecular functions, and (iv) the application of more complex (non-linear) machine learning models like deep neural networks. In this project, we conducted the first analyses of alternative splicing in single cells while considering the impact of genetic and epigenetic factors. Functionally relevant splicing events have been identified in bulk data, but variability in splicing between single cells from the same tissue and its relationship with epigenetic alterations remained poorly understood.

Instead of cancer, I studied differentiating iPS cells as a model, and replicated the results in mouse ES cells. This allowed us to not only assess the variability of both, alternative splicing and the underlying impact of genetic and epigenetic variation across cells, but also on a longitudinal scale, namely during the important process of stem cell differentiation. We were able to show for the first time that alternative splicing and splicing variability across cells can be predicted from genomic and DNA methylation variation in single cells. We studied the impact of DNA methylation and cellular features on cassette exon splicing, and characterised splicing variation and its determinants. We proved consistency of the detected relationships beyond cell type and species boundaries. Importantly, we finally investigated stability and variance of splicing during cellular differentiation and we found evidence that DNA methylation primes splicing switches during cell differentiation (Linker, Urban et al., 2019).

In the fourth section of this thesis, I showed how gene expression-mutagenesis screens could be used to understand complex mutational signatures, using the cancer hallmark of DNA repair deficiency as an example. This section of the thesis extended the analysis of the relationship between genetic variation and gene expression to the investigation of a clinically relevant genetic signature: HR-deficient tumours have been shown to be responsive to various inhibition therapies, but the molecular cause and consequence and genetic determinants of this repair deficiency are not yet fully understood. I investigated genome-wide molecular aberrations caused by HRD beyond the few key genes that are known by literature. I therefore firstly set up a gene expression-HRD screen and then used it to study the molecular basis of BRCAness, a postulated BRCA-independent mechanism of HRD, by correcting for known mutational and epigenetic alterations and germline predisposition of the *BRCA1/2* genes. My results pointed towards a dominant effect of *BRCA1* mutagenesis in both, a pan-cancer setting and per cancer type.

This analysis confronted me with two main limitations of gene expression association studies. First, estimating the number of confounding factors is not trivial. Whereas many association studies like QTL mapping approaches just maximise the power of their analyses across a range of confounding factors, this does not ensure finding an optimal number of hidden factors that is biologically meaningful in heterogeneous dataset. Hence, association analyses are at risk of correcting for important molecular signal if the number of hidden factors is too large, or of explaining non-targeted substructure of the data via spurious associations if the number of hidden factors is too small. Second, I detected many genes that were putatively associated with HRD, but which role in cancer progression has not been described yet. To be able to draw functional conclusions from such statistical findings, computational analyses like the ones presented in this thesis would have to be complemented and confirmed by functional downstream experiments like gene knockouts.

6 References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *J. R. Statist. Soc. B*, 44(2), 139–177.

Alberts, B., Bray, D., Hopkins, K., Johnson, A., Lewis, J., Raff, M., … Walter, P. (2009). *Garland Science: New York*.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V., … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. https://doi.org/10.1038/nature12477

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., & Stratton, M. R. (2013). Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports*, *3*(1), 246–259. https://doi.org/10.1016/j.celrep.2012.12.008

Alexandrov, L., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., Boot, A., … Stratton, M. R. (2018). The Repertoire of Mutational Signatures in Human Cancer. *BioRxiv*, 322859. https://doi.org/10.1101/322859

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, *33*(8), 831–838. https://doi.org/10.1038/nbt.3300

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angermueller, C., Clark, S. J., Lee, H. J., Macaulay, I. C., Teng, M. J., Hu, T. X., … Reik, W. (2016). Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nature Methods*, *13*(3), 229–232. https://doi.org/10.1038/nmeth.3728

Angermueller, C., Lee, H. J., Reik, W., & Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biology*, *18*(1), 67. https://doi.org/10.1186/s13059-017-1189-z

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*. https://doi.org/10.1038/75556

Atlas, E., Stramwasser, M., Whiskin, K., & Mueller, C. R. (2000). GA-binding protein α/β is a critical regulator of the BRCA1 promoter. *Oncogene*, *19*(15), 1933–1940. https://doi.org/10.1038/sj.onc.1203516

Avsec, Ž., Kreuzhuber, R., Israeli, J., Xu, N., Cheng, J., Shrikumar, A., … Gagneur, J. (2019). The Kipoi repository accelerates community exchange and reuse of predicitive models for genomics. Nature Biotchnology. https://doi.org/10.1038/s41587-019-0140-0

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., … Karchin, R. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, *173*(2), 371–385.e18. https://doi.org/10.1016/j.cell.2018.02.060

Bajrami, I., Frankum, J. R., Konde, A., Miller, R. E., Rehman, F. L., Brough, R., … Ashworth, A. (2014). Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Research*, *74*(1), 287–297. https://doi.org/10.1158/0008-5472.CAN-13-2541

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., … Blencowe, B. J. (2012). The evolutionary landscape of alternative splicing in vertebrate species. *Science*, *338*(6114), 1587–1593. https://doi.org/10.1126/science.1230612

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182.

Bass, J., Dabney, A., & Robinson, D. (2015). qvalue: Q-value estimation for false discovery rate control. R package (version 1.1).

Benjamini, Y., Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methological)*. https://doi.org/10.2307/2346101

Bernhart, S. H., Kretzmer, H., Holdt, L. M., Jühling, F., Ammerpohl, O., Bergmann, A. K., … Hoffmann, S. (2016). Changes of bivalent chromatin coincide with increased expression of developmental genes in cancer. *Scientific Reports*, *6*. https://doi.org/10.1038/srep37393

Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P., & Plemmons, R. J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics and Data Analysis*, *52*(1), 155–173. https://doi.org/10.1016/j.csda.2006.11.006

Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S., & Raghava, G. P. S. (2017). Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer. *Scientific Reports*, *7*. https://doi.org/10.1038/srep44997

Black, D. L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry*, *72*(1), 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720

Boot, A., Huang, M. N., Ng, A. W. T., Ho, S. C., Lim, J. Q., Kawakami, Y., … Rozen, S. G. (2018). In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Research*, *28*(5), 654–665. https://doi.org/10.1101/gr.230219.117

Boyer, L. A., Tong, I. L., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., … Young, R. A. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, *122*(6), 947–956. https://doi.org/10.1016/j.cell.2005.08.020

Brooks, A. N., Aspden, J. L., Podgornaia, A. I., Rio, D. C., & Brenner, S. E. (2011). Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans. *RNA*, *17*(10), 1884–1894. https://doi.org/10.1261/rna.2696311

Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences*, *101*(12), 4164–4169. https://doi.org/10.1073/pnas.0308531101

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Caan, B. J., Sweeney, C., Habel, L. A., Kwan, M. L., Kroenke, C. H., Weltzien, E. K., … Bernard, P. S. (2014). Intrinsic subtypes from the PAM50 gene expression assay in a population-based breast cancer survivor cohort: Prognostication of short- and long-term outcomes. *Cancer Epidemiology Biomarkers and Prevention*, *23*(5), 725–734. https://doi.org/10.1158/1055-9965.EPI-13-1017

Calabrese, C., Lehmann, K.-V., Urban, L., Liu, F., Erkek, S., Fonseca, N., … Stegle, O. (2017). Assessing the Gene Regulatory Landscape in 1,188 Human Tumors. *BioRxiv*, 225441. https://doi.org/10.1101/225441

Caldon, C. E. (2014). Estrogen Signaling and the DNA Damage Response in Hormone Dependent Breast Cancers. *Frontiers in Oncology*, *4*. https://doi.org/10.3389/fonc.2014.00106

Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. (2017). Pan-cancer analysis of whole genomes. *BioRxiv*, 162784. https://doi.org/10.1101/162784

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, *487*(7407), 61–70. https://doi.org/10.1038/nature11412

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A. D., Akbani, R., Liu, Y., … Levine, D. A. (2013). Integrated genomic characterization of endometrial carcinoma. *Nature*, *497*(7447), 67–73. https://doi.org/10.1038/nature12113

Cancer Genome Atlas Research Network, Weinstein, J., Collisson, E., Mills, G., Shaw, K., Ozenberger, B., & Ellrott, K. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for allelic expression analysis. *Genome Biology*, 016097. https://doi.org/10.1101/016097

Chakrabarti, R., Hwang, J., Andres Blanco, M., Wei, Y., Lukačišin, M., Romano, R. A., … Kang, Y. (2012). Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. *Nature Cell Biology*, *14*(11), 1212–1222. https://doi.org/10.1038/ncb2607

Charlesworth, B., & Charlesworth, D. (2010). Elements of evolutionary genetics. *Roberts and Company Publishers*, 432. https://doi.org/10.1525/bio.2011.61.5.12

Chartron, E., Theillet, C., Guiu, S., & Jacot, W. (2019). Targeting homologous repair deficiency in breast and ovarian cancers: Biological pathways, preclinical and clinical data. *Critical Reviews in Oncology/Hematology*. https://doi.org/10.1016/j.critrevonc.2018.10.012

Chen, K., Dai, X., & Wu, J. (2015). Alternative splicing: An important mechanism in stem cell biology. *World Journal of Stem Cells*, *7*(1), 1. https://doi.org/10.4252/wjsc.v7.i1.1

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., … Soranzo, N. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, *167*(5), 1398–1414.e24. https://doi.org/10.1016/j.cell.2016.10.026

Clark, S. J., Smallwood, S. A., Lee, H. J., Krueger, F., Reik, W., & Kelsey, G. (2017). Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nature Protocols*, *12*(3), 534–547. https://doi.org/10.1038/nprot.2016.187

Cordell, H. J. (2002). Epistasis: what it meains, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics, 11*(20), 2463-2468.

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., Zou, X., … Nik-Zainal, S. (2017). HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine*, *23*(4), 517–525. https://doi.org/10.1038/nm.4292

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J. B., Gaffney, D. J., Pickrell, J. K., ... Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, 482(7385), 390–394.

Denat, L., Kadekaro, A. L., Marrot, L., Leachman, S. A., & Abdel-Malek, Z. A. (2014). Melanocytes as instigators and victims of oxidative stress. *Journal of Investigative Dermatology*. https://doi.org/10.1038/jid.2014.65

Ding, C., Li, Y., Xia, Y., Wei, W., Zhang, L., & Zhang, Y. (2017). Convolutional neural networks based hyperspectral image classification method with adaptive kernels. *Remote Sensing*, *9*(6). https://doi.org/10.3390/rs9060618

Ding, J., McConechy, M. K., Horlings, H. M., Ha, G., Chun Chan, F., Funnell, T., … Shah, S. P. (2015). Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9554

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt. *Nature Protocols*, *4*(8), 1184–1191. https://doi.org/10.1038/nprot.2009.97

Evans, K. W., Yuca, E., Akcakanat, A., Scott, S. M., Arango, N. P., Zheng, X., … Meric-Bernstam, F. (2017). A population of heterogeneous breast cancer patient-derived xenografts demonstrate broad activity of PARP inhibitor in BRCA1/2 wild-type tumors. *Clinical Cancer Research*, *23*(21), 6468–6477. https://doi.org/10.1158/1078-0432.CCR-17-0615

Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., … D'Eustachio, P. (2018). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, *46*(D1), D649–D655. https://doi.org/10.1093/nar/gkx1132

Faigenbloom, L., Rubinstein, N. D., Kloog, Y., Mayrose, I., Pupko, T., & Stein, R. (2015). Regulation of alternative splicing at the single-cell level. *Molecular Systems Biology*, *11*(12), 845.

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., … Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, *343*(6175). https://doi.org/10.1126/science.1246949

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., … Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nature Genetics*, *44*(5), 502–510. https://doi.org/10.1038/ng.2205

Fan, Y., Xi, L., Hughes, D. S. T., Zhang, J., Zhang, J., Futreal, P. A., … Wang, W. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, *17*(1), 178. https://doi.org/10.1186/s13059-016-1029-6

Fica, S. M., Tuttle, N., Novak, T., Li, N. S., Lu, J., Koodathingal, P., … Piccirilli, J. A. (2013). RNA catalyses nuclear pre-mRNA splicing. *Nature*, *503*(7475), 229–234. https://doi.org/10.1038/nature12734

Fisher, R. A. (1918). The correlation between relatives on the supposition of genomic imprinting. *Transactions of the Royal Society of Edinburgh*, *52.02*, 399–433. https://doi.org/doi:10.1017/S0080456800012163

Fong, P. C., Boss, D. S., Yap, T. A., Tutt, A., Wu, P., Mergui-Roelvink, M., … de Bono, J. S. (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *The New England Journal of Medicine*, *361*(2), 123–134. https://doi.org/10.1056/NEJMoa0900212

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., … Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Forbes, S. A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., … Stratton, M. R. (2008). The catalogue of somatic mutations in cancer (COSMIC). *Current Protocols in Human Genetics*, (SUPPL. 57). https://doi.org/10.1002/0471142905.hg1011s57

Forbes, S., Clements, J., Dawson, E., Bamford, S., Webb, T., Dogan, A., … Stratton, M. R. (2006). COSMIC 2005. *British Journal of Cancer*. https://doi.org/10.1038/sj.bjc.6602928

Frazer, K. A., Murray, S. S., Schork, N. J., & Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2554

Fredriksson, N. J., Ny, L., Nilsson, J. A., & Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, *46*(12), 1258–1263. https://doi.org/10.1038/ng.3141

Gallego-Ortega, D., Ledger, A., Roden, D. L., Law, A. M. K., Magenau, A., Kikhtyak, Z., … Ormandy, C. J. (2015). ELF5 Drives Lung Metastasis in Luminal Breast Cancer through Recruitment of Gr1+ CD11b+ Myeloid-Derived Suppressor Cells. *PLoS Biology*, *13*(12). https://doi.org/10.1371/journal.pbio.1002330

Galton, F. (1871). Hereditary genius. *Notes and Queries*. https://doi.org/10.1093/nq/s4-VII.178.451-c

Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Peat, G., … Saez-Rodriguez, J. (2018). Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Research*, *78*(3), 769–780. https://doi.org/10.1158/0008-5472.CAN-17-1679

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., … Relton, C. L. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*, 61. https://doi.org/10.1186/s13059-016-0926-z

Goeman, J. J., & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statistics in Medicine*, *33*(11), 1946–1978. https://doi.org/10.1002/sim.6082

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., … Snyder, M. (2015). Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell*, *162*(5), 1051–1065. https://doi.org/10.1016/j.cell.2015.07.048

Ha, G., Roth, A., Lai, D., Bashashati, A., Ding, J., Goya, R., … Shah, S. P. (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Research*, *22*(10), 1995–2007. https://doi.org/10.1101/gr.137570.112

Hajirasouliha, I., Mahmoody, A., & Raphael, B. J. (2014). A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, *30*(12). https://doi.org/10.1093/bioinformatics/btu284

Hanafusa, T., Mohamed, A. E. A., Domae, S., Nakayama, E., & Ono, T. (2012). Serological identification of Tektin5 as a cancer/testis antigen and its immunogenicity. *BMC Cancer*, *12*. https://doi.org/10.1186/1471-2407-12-520

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hannan, N. R. F., Segeritz, C.-P., Touboul, T., & Vallier, L. (2013). Production of hepatocyte-like cells from human pluripotent stem cells. *Nature Protocols*, *8*(2), 430–437.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., … Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, *7 Suppl 1*, S4.1-9. https://doi.org/ARTN S4DOI 10.1186/gb-2006-7-s1-s4

Hartmann, S., Schuhmacher, B., Rausch, T., Fuller, L., Döring, C., Weniger, M., … Hansmann, M. L. (2016). Highly recurrent mutations of SGK1, DUSP2 and JUNB in nodular lymphocyte predominant Hodgkin lymphoma. *Leukemia*, *30*(4), 844–853. https://doi.org/10.1038/leu.2015.328

Hassold, T., Hall, H., & Hunt, P. (2007). The origin of human aneuploidy: Where we have been, where we are going. *Human Molecular Genetics*. https://doi.org/10.1093/hmg/ddm243

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. *Book*. https://doi.org/10.1007/978-0-387-84858-7

He, L., Bai, Q., & Tang, L. (2015). Alternative splicing regulates pluripotent state in pluripotent stem cells. *Current Stem Cell Research & Therapy*, *10*(2), 159–165.

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg3729

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955–959. https://doi.org/10.1038/ng.2354

Huang, Y., & Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, *18*(1), 123. https://doi.org/10.1186/s13059-017-1248-5

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., … Clamp, M. (2002). The Ensembl genome database project. *Nucleic Acids Research*, *30*(1), 38–41.

Hubel, D. H., & Wiesel, T. N. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of Physiology*, *165*(3), 559–568. https://doi.org/10.1113/jphysiol.1963.sp007079

Huber, K. R., Schmidt, W. F., Ettinger, R. S., & Neuberg, R. W. (1988). Antiproliferative Effect of Verapamil Alone on Brain Tumor Cells in Vitro. *Cancer Research*, *48*(13), 3617–3621.

Ilboudo, A., Nault, J. C., Dubois-Pot-Schneider, H., Corlu, A., Zucman-Rossi, J., Samson, M., & Le Seyec, J. (2014). Overexpression of phosphatidylinositol 4-kinase type IIIα is associated with undifferentiated status and poor prognosis of human hepatocellular carcinoma. *BMC Cancer*, *14*(1). https://doi.org/10.1186/1471-2407-14-7

Ingram, W. J., Crowther, L. M., Little, E. B., Freeman, R., Harliwong, I., Veleva, D., … Hallahan, A. R. (2013). ABC transporter activity linked to radiation resistance and molecular subtype in pediatric medulloblastoma. *Experimental Hematology & Oncology*, *2*(1), 26. https://doi.org/10.1186/2162-3619-2-26

Jackson, S. P. & Bartek, J. (2009). The DNA-damage response in human biology and disease. *Nature*, *461*(7267), 1071–1078. https://doi.org/10.1038/nature08467.The

James Kent, W., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102.

Janky, R., Verfaillie, A., Imrichová, H., van de Sande, B., Standaert, L., Christiaens, V., … Aerts, S. (2014). iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Computational Biology*, *10*(7). https://doi.org/10.1371/journal.pcbi.1003731

Jia, P., Pao, W., & Zhao, Z. (2014). Patterns and processes of somatic mutations in nine major cancers. *BMC Medical Genomics*, *7*(1). https://doi.org/10.1186/1755-8794-7-11

Jimbow, K., Chen, H., Park, J. S., & Thomas, P. D. (2001). Increased sensitivity of melanocytes to oxidative stress and abnormal expression of tyrosinase-related protein in vitiligo. *British Journal of Dermatology*, *144*(1), 55–65. https://doi.org/10.1046/j.1365-2133.2001.03952.x

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., … Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, *23*(22), 5866–5878. http://dx.doi.org/10.1093/hmg/ddu309

Kalimutho, M., Bain, A. L., Mukherjee, B., Nag, P., Nanayakkara, D. M., Harten, S. K., … Khanna, K. K. (2017). Enhanced dependency of KRAS-mutant colorectal cancer cells on RAD51-dependent homologous recombination repair identified from genetic interactions in Saccharomyces cerevisiae. *Molecular Oncology*, *11*(5), 470–490. https://doi.org/10.1002/1878-0261.12040

Kalyuga, M., Gallego-Ortega, D., Lee, H. J., Roden, D. L., Cowley, M. J., Caldon, C. E., … Ormandy, C. J. (2012). ELF5 Suppresses Estrogen Sensitivity and Underpins the Acquisition of Antiestrogen Resistance in Luminal Breast Cancer. *PLoS Biology*, *10*(12). https://doi.org/10.1371/journal.pbio.1001461

Kanchi, K. L., Johnson, K. J., Lu, C., McLellan, M. D., Leiserson, M. D. M., Wendl, M. C., … Ding, L. (2014). Integrated analysis of germline and somatic variants in ovarian cancer. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4156

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. https://doi.org/10.1093/nar/gkv1070

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics*, *178*(3), 1709–1723. https://doi.org/10.1534/genetics.107.080101

Karran, P. (1996). Microsatellite instability and DNA mismatch repair in human cancer. *Seminars in Cancer Biology*. https://doi.org/10.1006/scbi.1996.0003

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., & Stamm, S. (2013). Function of alternative splicing. *Gene*, *514*(1), 1–30. https://doi.org/10.1016/j.gene.2012.07.083

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: Learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990–999. https://doi.org/10.1101/gr.200535.115

Kennedy, S. R., Loeb, L. A., Herr, A. J. (2012). Somatic Mutations in Aging, Cancer and Neurodegeneration. Mech Ageing Dev, 133(4), 118–126.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, A. D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., … Gaffney, D. J. (2017). Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*, *546*(7658), 370–375. https://doi.org/10.1038/nature22403

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*, *35*(1), 125–131. https://doi.org/10.1093/nar/gkl924

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217, 624–626.

Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Knijnenburg, T. A., Wang, L., Zimmermann, M. T., Chambwe, N., Gao, G. F., Cherniack, A. D., … Mariamidze, A. (2018). Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports*, *23*(1), 239–254.e6. https://doi.org/10.1016/j.celrep.2018.03.076

Knudson, A. G. (2002). Cancer genetics. *American Journal of Medical Genetics, 111*(1), 96-102.

Korir, P. K., & Seoighe, C. (2014). Inference of allele-specific expression from RNA-seq data. *Methods in Molecular Biology*, *1112*, 49–69. https://doi.org/10.1007/978-1-62703-773-0_4

Krueger, F. (2011). Trim Galore! https://doi.org/https://github.com/FelixKrueger/TrimGalore

Krueger, F., & Andrews, S. R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics (Oxford, England)*, *27*(11), 1571–1572. https://doi.org/10.1093/bioinformatics/btr167

Kvam, E., & Tyrrell, R. M. (1999). The role of melanin in the induction of oxidative DNA base damage by ultraviolet A irradiation of DNA or melanoma cells. *Journal of Investigative Dermatology*, *113*(2), 209–213. https://doi.org/10.1046/j.1523-1747.1999.00653.x

Lamberti, A., Caraglia, M., Longo, O., Marra, M., Abbruzzese, A., & Arcari, P. (2004). The translation elongation factor 1A in tumorigenesis, signal transduction and apoptosis: Review article. *Amino Acids*, *26*(4), 443–448. https://doi.org/10.1007/s00726-004-0088-2

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., … Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Larsen, M. J., Kruse, T. A., Tan, Q., Lænkholm, A. V., Bak, M., Lykkesfeldt, A. E., … Thomassen, M. (2013). Classifications within Molecular Subtypes Enables Identification of BRCA1/BRCA2 Mutation Carriers by RNA Tumor Profiling. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064268

Lasorella, A., Benezra, R., & Iavarone, A. (2014). The ID proteins: Master regulators of cancer stem cells and tumour aggressiveness. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc3638

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Leung, M. K. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics*, *30*(12). https://doi.org/10.1093/bioinformatics/btu277

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., … Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS Biology*, *5*(10), 2113–2144. https://doi.org/10.1371/journal.pbio.0050254

Li, M., Sun, Q., & Wang, X. (2017). Transcriptional landscape of human cancers. *Oncotarget*, *8*(21). https://doi.org/10.18632/oncotarget.15837

Liang, P., & Pardee, A. B. (2003). Analysing differential gene expression in cancer. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc1214

Lindeboom, R. G. H., Supek, F., & Lehner, B. (2016). The rules and impact of nonsense-mediated mRNA decay in human cancers. *Nature Genetics*, *48*(10), 1112–1118. https://doi.org/10.1038/ng.3664

Linker, S., Urban, L., Clark, S., Chhatriwala, M., Mccarthy, D., Ebersberger, I., … Reik, W. (2019). Combined single cell profiling of expression and DNA methylation reveals splicing regulation and heterogeneity. *Genome Biology*, (20), 30.

Lippert, C., Casale, F. P., Rakitsch, B., & Stegle, O. (2014). LIMIX: genetic analysis of multiple traits. *bioRxivorg*. https://doi.org/10.1101/003905

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833–835. https://doi.org/10.1038/nmeth.1681

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., … Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*. https://doi.org/10.1038/ng.2653

Lord, C. J., & Ashworth, A. (2016). BRCAness revisited. *Nature Reviews Cancer*, *16*(2), 110–120. https://doi.org/10.1038/nrc.2015.21

Lu, C., Xie, M., Wendl, M. C., Wang, J., McLellan, M. D., Leiserson, M. D. M., … Ding, L. (2015). Patterns and functional implications of rare germline variants across 12 cancer types. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms10086

Macaulay, I. C., Teng, M. J., Haerty, W., Kumar, P., Ponting, C. P., & Voet, T. (2016). Separation and parallel sequencing of the genomes and transcriptomes of single cells using G&T-seq. *Nature Protocols*, *11*(11), 2081–2103. https://doi.org/10.1038/nprot.2016.138

Maciejewski, J. P. & Mufti, G. J. (2008). Whole genome scanning as a cytogenetic tool in hematologic malignancies. *Blood*, 112(4), 965–974.

Mansour, W. Y., Tennstedt, P., Volquardsen, J., Oing, C., Kluth, M., Hube-Magg, C., … Rothkamm, K. (2018). Loss of PTEN-assisted G2/M checkpoint impedes homologous recombination repair and enhances radio-curability and PARP inhibitor treatment response in prostate cancer. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-22289-7

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Masica, D. L., & Karchin, R. (2011). Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Research*, *71*(13), 4550–4561. https://doi.org/10.1158/0008-5472.CAN-11-0180

Matlin, A. J., Clark, F., & Smith, C. W. J. (2005). Understanding alternative splicing: Towards a cellular code. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/nrm1645

Maunakea, A. K., Chepelev, I., Cui, K., & Zhao, K. (2013). Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Research*, *23*(11), 1256–1269. https://doi.org/10.1038/cr.2013.110

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., & Wills, Q. F. (2017). Scater: Pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics*, *33*(8), 1179–1186. https://doi.org/10.1093/bioinformatics/btw777

McClellan, J., & King, M. C. (2010). Genetic heterogeneity in human disease. *Cell, 141*(2), 210–217. https://doi.org/10.1016/j.cell.2010.03.032

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0974-4

Medvedev, S. P., Shevchenko, A. I., & Zakian, S. M. (2010). Induced Pluripotent Stem Cells: Problems and Advantages when Applying them in Regenerative Medicine. *Acta Naturae*, *2*(2), 18–28.

Mendel, G. (1865). Versuche ueber Pflanzenhybriden. *Verhandlungen Des Naturforschenden Vereines in Brünn*, 3–47.

Menghi, F., Barthel, F. P., Yadav, V., Tang, M., Ji, B., Tang, Z., … Liu, E. T. (2018). The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. *Cancer Cell*, *34*(2), 197–210.e5. https://doi.org/10.1016/j.ccell.2018.06.008

Middlebrooks, C. D., Banday, A. R., Matsuda, K., Udquim, K. I., Onabajo, O. O., Paquin, A., … Prokunina-Olsson, L. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nature Genetics*, *48*(11), 1330–1338. https://doi.org/10.1038/ng.3670

Milacic, M., Haw, R., Rothfels, K., Wu, G., Croft, D., Hermjakob, H., … Stein, L. (2012). Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers*, *4*(4), 1180–1211. https://doi.org/10.3390/cancers4041180

Minami, A., Nakanishi, A., Ogura, Y., Kitagishi, Y., & Matsuda, S. (2014). Connection between Tumor Suppressor BRCA1 and PTEN in Damaged DNA Repair. *Frontiers in Oncology*, *4*. https://doi.org/10.3389/fonc.2014.00318

Montoni, A., Robu, M., Pouliot, É., & Shah, G. M. (2013). Resistance to PARP-inhibitors in cancer therapy. *Frontiers in Pharmacology*. https://doi.org/10.3389/fphar.2013.00018

Morison, I. M., Ramsay, J. P., & Spencer, H. G. (2005). A census of mammalian imprinting. *Trends in Genetics*. https://doi.org/10.1016/j.tig.2005.06.008

Movilla, N., Crespo, P., & Bustelo, X. R. (1999). Signal transduction elements of TC21, an oncogenic member of the R-Ras subfamily of GTP-binding proteins. *Oncogene*, *18*(43), 5860–5869. https://doi.org/10.1038/sj.onc.1202968

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., … Stratton, M. R. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, *534*(7605), 47–54. https://doi.org/10.1038/nature17676

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., … Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*, *149*(5), 994–1007. https://doi.org/10.1016/j.cell.2012.04.023

Nik-Zainal, S., Wedge, D. C., Alexandrov, L. B., Petljak, M., Butler, A. P., Bolli, N., … Stratton, M. R. (2014). Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nature Genetics*, *46*(5), 487–491. https://doi.org/10.1038/ng.2955

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., … Bustamante, C. D. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98–101. https://doi.org/10.1038/nature07331

Ongen, H., Andersen, C. L., Bramsen, J. B., Oster, B., Rasmussen, M. H., Ferreira, P. G., … Dermitzakis, E. T. (2014). Putative cis-regulatory drivers in colorectal cancer. *Nature*, *512*(1), 87–90. https://doi.org/10.1038/nature13602

Othman, R. T., Kimishi, I., Bradshaw, T. D., Storer, L. C. D., Korshunov, A., Pfister, S. M., … Coyle, B. (2014). Overcoming multiple drug resistance mechanisms in medulloblastoma. *Acta Neuropathologica Communications*, *2*(1). https://doi.org/10.1186/2051-5960-2-57

Palazzo, A. F. & Lee, E. S. (2015). Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics (6)*. https://doi.org/10.3389/fgene.2015.00002

Parker, J. S., Mullins, M., Cheang, M. C. U., Leung, S., Voduc, D., Vickery, T., … Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. https://doi.org/10.1200/JCO.2008.18.1370

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

PCAWG Transcriptome Core Group, Calabrese, C., Davidson, N. R., Fonseca, N. A., He, Y., Kahles, A., … Zhang, Z. (2018). Genomic basis for RNA alterations revealed by whole-genome analyses of 27 cancer types. *BioRxiv*, 183889. https://doi.org/10.1101/183889

Peng, G., Chun-Jen Lin, C., Mo, W., Dai, H., Park, Y.-Y., Kim, S. M., … Lin, S.-Y. (2014). Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4361

Petljak, M., & Alexandrov, L. B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, *37*(6), 531–540. https://doi.org/10.1093/carcin/bgw055

Petljak, M., Alexandrov, L. B., Brammeld, J. S., Price, S., Wedge, D. C., Grossmann, S., … Stratton, M. R. (2019). Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell*, *176*(6), 1282–1294.e20. https://doi.org/10.1016/j.cell.2019.02.012

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., … Stratton, M. R. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, *463*(7278), 191–196. https://doi.org/10.1038/nature08658

Polak, P., Kim, J., Braunstein, L. Z., Karlic, R., Haradhavala, N. J., Tiao, G., … Getz, G. (2017). A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nature Genetics*, *49*(10), 1476–1486. https://doi.org/10.1038/ng.3934

Post, S. M., Tomkinson, A. E., & Lee, E. Y.-H. P. (2003). The human checkpoint Rad protein Rad17 is chromatin-associated throughout the cell cycle, localizes to DNA replication sites, and interacts with DNA polymerase E. *Nucleic Acids Research*, *31*(19), 5568–5575. https://doi.org/10.1093/nar/gkg765

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, and Computers*, *36*(4), 717–731. https://doi.org/10.3758/BF03206553

Premi, S., Wallisch, S., Mano, C. M., Weiner, A. B., Wakamatsu, K., Bechara, E. J. H., … Brash, D. E. (2015). Photoproducts Long after UV Exposure, *347*(6224), 842–847. https://doi.org/10.1126/science.1256022.Chemiexcitation

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rausch, T., Jones, D. T. W., Zapatka, M., Stütz, A. M., Zichner, T., Weischenfeldt, J., … Korbel, J. O. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, *148*(1–2), 59–71. https://doi.org/10.1016/j.cell.2011.12.013

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., … Yosef, N. (2017). The Human Cell Atlas. *ELife*, *6*. https://doi.org/10.7554/elife.27041

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, *44*(W1), W83–W89. https://doi.org/10.1093/nar/gkw199

Revil, T., Gaffney, D., Dias, C., Majewski, J., & Jerome-Majewska, L. A. (2010). Alternative splicing is frequent during early embryonic development in mouse. *BMC Genomics*, *11*, 399. https://doi.org/10.1186/1471-2164-11-399

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *2012*, *48*(2), 36. https://doi.org/10.18637/jss.v048.i02

Roundtable on Translating Genomic-Based Research for Health; Board on Health Sciences Policy; Health and Medicine Division; National Academies of Sciences, Engineering, and M. (2016). Large Genetic Cohort Studies: A Background. *Applying an Implementation Science Approach to Genomic Medicine: Workshop Summary. Washington (DC): National Academies Press (US)*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK373714/

Saini, N., Roberts, S. A., Sterling, J. F., Malc, E. P., Mieczkowski, P. A., & Gordenin, D. A. (2017). APOBEC3B cytidine deaminase targets the non-transcribed strand of tRNA genes in yeast. *DNA Repair*, *53*, 4–14. https://doi.org/10.1016/j.dnarep.2017.03.003

Saito, H., Fushida, S., Miyashita, T., Oyama, K., Yamaguchi, T., Tsukada, T., … Ohta, T. (2017). Potential of extravasated platelet aggregation as a surrogate marker for overall survival in patients with advanced gastric cancer treated with preoperative docetaxel, cisplatin and S-1: A retrospective observational study. *BMC Cancer*, *17*(1). https://doi.org/10.1186/s12885-017-3279-4

Sakharkar, M. K., Perumal, B. S., Lim, Y. P., Chern, L. P., Yu, Y., & Kangueane, P. (2005). Alternatively spliced human genes by exon skipping--a database (ASHESdb). *In Silico Biol 5*(3):221-225.

Sammeth, M., Foissac, S., & Guigó, R. (2008). A general definition and nomenclature for alternative splicing events. *PLoS Computational Biology*, *4*(8), e1000147. https://doi.org/10.1371/journal.pcbi.1000147

Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., & Chen, Y. T. (2002). Cancer/testis antigens: An expanding family of targets for cancer immunotherapy. *Immunological Reviews*, *188*, 22–32. https://doi.org/10.1034/j.1600-065X.2002.18803.x

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., … Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biology*, *6*(5), 1020–1032. https://doi.org/10.1371/journal.pbio.0060107

Schaetzlein, S., Chahwan, R., Avdievich, E., Roa, S., Wei, K., Eoff, R. L., … Edelmann, W. (2013). Mammalian Exo1 encodes both structural and catalytic functions that play distinct roles in essential biological processes. *Proceedings of the National Academy of Sciences*, *110*(27), E2470–E2479. https://doi.org/10.1073/pnas.1308512110

Shalek, A. K., Satija, R., Adiconis, X., Gertner, R. S., Gaublomme, J. T., Raychowdhury, R., … Regev, A. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, *498*(7453), 236–240. https://doi.org/10.1038/nature12172

Sharma, P., Jha, A. B., Dubey, R. S., & Pessarakli, M. (2012). Reactive Oxygen Species, Oxidative Damage, and Antioxidative Defense Mechanism in Plants under Stressful Conditions. *Journal of Botany*, *2012*, 1–26. https://doi.org/10.1155/2012/217037

Sharp, P. A. (1994). Split genes and RNA splicing. *Science*, *77*(4390), 805–815. https://doi.org/10.1126/science.373120

Sharrocks, A. D. (2001). The ETS-domain transcription factor family. *Nature Reviews Molecular Cell Biology*, *2*(11), 827–837. https://doi.org/10.1038/35099076

Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., … Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, *479*(7371), 74–79. https://doi.org/10.1038/nature10442

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., … Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, *15*(8), 1034–1050. https://doi.org/10.1101/gr.3715005

Simpson, A. J. G., Caballero, O. L., Jungbluth, A., Chen, Y. T., & Old, L. J. (2005). Cancer/testis antigens, gametogenesis and cancer. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc1669

Sizemore, G. M., Pitarresi, J. R., Balakrishnan, S., & Ostrowski, M. C. (2017). The ETS family of oncogenic transcription factors in solid tumours. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc.2017.20

Slatkin, M. (2008). Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* *9*(6), 477-485. https://doi.org/10.1038/nrg2361

Smith, K. S., Yadav, V. K., Pedersen, B. S., Shaknovich, R., Geraci, M. W., Pollard, K. S., & De, S. (2015). Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Research*, *43*(11), 5307–5317. https://doi.org/10.1093/nar/gkv419

Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., & Yeo, G. W. (2017). Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation. *Molecular Cell*, *67*(1), 148–161.e5. https://doi.org/10.1016/j.molcel.2017.06.003

Stark, A. L., Hause, R. J., Gorsic, L. K., Antao, N. N., Wong, S. S., Chung, S. H., … Dolan, M. E. (2014). Protein Quantitative Trait Loci Identify Novel Candidates Modulating Cellular Response to Chemotherapy. *PLoS Genetics*, *10*(4). https://doi.org/10.1371/journal.pgen.1004192

Stegle, O., Parts, L., Piipari, M., Winn, J., & Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500–507. https://doi.org/10.1038/nprot.2011.457

Storey, J. D. (2010). False Discovery Rates. *Princeton University, Princeton, USA*, 1–7. https://doi.org/10.1198/016214507000000941

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the “perceptron” algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research*, *10*(9), 2997–3011. https://doi.org/10.1093/nar/10.9.2997

Streeter, I., Harrison, P. W., Faulconbridge, A., The HipSci Consortium, Flicek, P., Parkinson, H., & Clarke, L. (2017). The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Research*, *45*(D1), D691–D697. https://doi.org/10.1093/nar/gkw928

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Suen, T. C., & Goss, P. E. (1999). Transcription of BRCA1 is dependent on the formation of a specific protein-DNA complex on the minimal BRCA1 bi-directional promoter. *Journal of Biological Chemistry*, *274*(44), 31297–31304. https://doi.org/10.1074/jbc.274.44.31297

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*, *45*(W1), W98–W102. https://doi.org/10.1093/nar/gkx247

Telli, M. L., Stover, D. G., Loi, S., Aparicio, S., Carey, L. A., Domchek, S. M., … Winer, E. P. (2018). Homologous recombination deficiency and host anti-tumor immunity in triple-negative breast cancer. *Breast Cancer Research and Treatment*. https://doi.org/10.1007/s10549-018-4807-x

Temko, D., Tomlinson, I. P. M., Severini, S., Schuster-Boeckler, B., Graham, T. A. (2018). The effects of mutational processes and selection on driver mutations across cancer types. *Nature Communications*, 9(1857). https://doi.org/10.1038/s41467-018-04208-6

The GTExArd Consortium, Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., … Bunney, W. E. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, *348*(6235), 648–660. https://doi.org/10.1126/science.1262110

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for Causal Mediation Analysis. *Journal of Statistical Software*, *59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Tomlins, S. A., Laxman, B., Varambally, S., Cao, X., Yu, J., Helgeson, B. E., … Chinnaiyan, A. M. (2008). Role of the TMPRSS2-ERG Gene Fusion in Prostate Cancer. *Neoplasia*, *10*(2), 177-IN9. https://doi.org/10.1593/neo.07822

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K., Amini, B., Andersen, E., … Pontén, F. (2005). A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics. *Molecular & Cellular Proteomics*, *4*(12), 1920–1932. https://doi.org/10.1074/mcp.M500279-MCP200

Valavanidis, A., Vlachogianni, T., & Fiotakis, C. (2009). 8-Hydroxy-2′ -deoxyguanosine (8-OHdG): A critical biomarker of oxidative stress and carcinogenesis. *Journal of Environmental Science and Health - Part C Environmental Carcinogenesis and Ecotoxicology Reviews*, *27*(2), 120–139. https://doi.org/10.1080/10590500902885684

van ’t Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536. https://doi.org/10.1038/415530a

Viel, A., Bruselles, A., Meccia, E., Fornasarig, M., Quaia, M., Canzonieri, V., … Bignami, M. (2017). A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine*, *20*, 39–49. https://doi.org/10.1016/j.ebiom.2017.04.022

Waddell, N., Pajic, M., Patch, A. M., Chang, D. K., Kassahn, K. S., Bailey, P., … Grimmond, S. M. (2015). Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature*, *518*(7540), 495–501. https://doi.org/10.1038/nature14169

Wahl, M. C., Will, C. L., & Lührmann, R. (2009). The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*. https://doi.org/10.1016/j.cell.2009.02.009

Wainberg, M., Alipanahi, B., & Frey, B. (2016). Does conservation account for splicing patterns? *BMC Genomics*, *17*(1), 787. https://doi.org/10.1186/s12864-016-3121-4

Wallace, C. (2013). Statistical testing of shared genetic control for potentially related traits. *Genetic Epidemiology*, *37*(8), 802–813. https://doi.org/10.1002/gepi.21765

Wang, Y., Mark, K. M. K., Ung, M. H., Kettenbach, A., Miller, T., Xu, W., … Cheng, C. (2016). Application of RNAi-induced gene expression profiles for prognostic prediction in breast cancer. *Genome Medicine*, *8*(1). https://doi.org/10.1186/s13073-016-0363-3

Wang, Y., Ung, M. H., Cantor, S., & Cheng, C. (2017). Computational Investigation of Homologous Recombination DNA Repair Deficiency in Sporadic Breast Cancer. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-16138-2

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wang, Z., Liu, P., Inuzuka, H., & Wei, W. (2014). Roles of F-box proteins in cancer. *Nature Reviews Cancer*. https://doi.org/10.1038/nrc3700

Waszak, S. M., Delaneau, O., Gschwind, A. R., Kilpinen, H., Raghav, S. K., Witwicki, R. M., … Dermitzakis, E. T. (2015). Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. *Cell*, *162*(5), 1039–1050. https://doi.org/10.1016/j.cell.2015.08.001

Waszak, S. M., Tiao, G., Zhu, B., Rausch, T., Muyas, F., Rodriguez-Martin, B., … Net, I. P.-C. A. of W. G. (2017). Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. *BioRxiv*, 208330. https://doi.org/10.1101/208330

Weinberg, R. A. (2006) The Biology of Cancer. Garland Science, ISBN 0815340788.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., & Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature Genetics*, *46*(11), 1160–1165. https://doi.org/10.1038/ng.3101

Weir, B., Zhao, X., & Meyerson, M. (2004). Somatic alterations in the human cancer genome. *Cancer Cell*. https://doi.org/10.1016/j.ccr.2004.11.004

Weiss, S. (1999). Fluorescence spectroscopy of single biomolecules. *Science*. https://doi.org/10.1126/science.283.5408.1676

Welch, J. D., Hu, Y., & Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, *44*(8), e73. https://doi.org/10.1093/nar/gkv1525

Welcsh, P. L., Lee, M. K., Gonzalez-Hernandez, R. M., Black, D. J., Mahadevappa, M., Swisher, E. M., … King, M.-C. (2002). BRCA1 transcriptionally regulates genes involved in breast tumorigenesis. *Proceedings of the National Academy of Sciences*, *99*(11), 7560–7565. https://doi.org/10.1073/pnas.062181799

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Woenckhaus, M., Klein-Hitpass, L., Grepmeier, U., Merk, J., Pfeifer, M., Wild, P. J., … Dietmaier, W. (2006). Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *Journal of Pathology*, *210*(2), 192–204. https://doi.org/10.1002/path.2039

Wu, H. T., Hajirasouliha, I., & Raphael, B. J. (2014). Detecting independent and recurrent copy number aberrations using interval graphs. *Bioinformatics*, *30*(12). https://doi.org/10.1093/bioinformatics/btu276

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., … Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, *347*(6218), 1254806. https://doi.org/10.1126/science.1254806

Yamanaka, S. (2012). Induced pluripotent stem cells: Past, present, and future. *Cell Stem Cell*. https://doi.org/10.1016/j.stem.2012.05.005

Yang, J., Bachrati, C. Z., Ou, J., Hickson, I. D., & Brown, G. W. (2010). Human topoisomerase IIIα is a single-stranded DNA decatenase that is stimulated by BLM and RMI1. *Journal of Biological Chemistry*, *285*(28), 21426–21436. https://doi.org/10.1074/jbc.M110.123216

Yao, C., Joehanes, R., Johnson, A. D., Huan, T., Liu, C., Freedman, J. E., … Levy, D. (2017). Dynamic Role of trans Regulation of Gene Expression in Relation to Complex Traits. *American Journal of Human Genetics*, *100*(4), 571–580. https://doi.org/10.1016/j.ajhg.2017.02.003

Yearim, A., Gelfman, S., Shayevitch, R., Melcer, S., Glaich, O., Mallm, J.-P., … Ast, G. (2015). HP1 is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Reports*, *10*(7), 1122–1134. https://doi.org/10.1016/j.celrep.2015.01.038

Yu, G., & He, Q.-Y. (2016). ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.*, *12*(2), 477–479. https://doi.org/10.1039/C5MB00663E

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., … Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., … Kasprzyk, A. (2011). International cancer genome consortium data portal-a one-stop shop for cancer genomics data. *Database*, *2011*. https://doi.org/10.1093/database/bar026

Zhou, C., Smith, J. L. & Liu, J. (2003). Role of BRCA1 in cellular resistance to paclitaxel and ionizing radiation in an ovarian cancer cell line carrying a defective BRCA1. *Oncogene* *22*(16), 2396–404.

Zimmermann, M., Murina, O., Reijns, M. A. M., Agathanggelou, A., Challis, R., Tarnauskaite, Ž., … Durocher, D. (2018). CRISPR screens identify genomic ribonucleotides as a source of PARP-trapping lesions. *Nature*, *559*(7713), 285–289. https://doi.org/10.1038/s41586-018-0291-z