André A. R. B. Duarte

Alternative splicing-mediated cis-regulation in Breast Cancer Risk

Department of Biomedical Sciences and Medicine

2020

André A. R. B. Duarte

Alternative splicing-mediated cis-regulation in Breast Cancer Risk

Master in Oncobiology – Cancer Molecular Mechanisms

Work under the supervision of:

Ana-Teresa Maia, Ph.D.

Department of Biomedical Sciences and Medicine

2020

Alternative splicing-mediated cis-regulation in Breast Cancer Risk

Authorship Statement

I hereby declare to be the author of this work, which is original and unpublished. Authors and papers consulted are duly cited in the text and are listed in the included references.

Acknowledgments

First and foremost, I'd like to express my most sincere thank you to Ana-Teresa Maia for providing scientific and personal mentorship throughout this work. Without her supervision and support the present work would not be accomplished.

Additionally, I would like to acknowledge the contributions made by all members of the research group, both individually and as a collective for the discussions, knowledge and tips they provided during past 15 months. In no particular order, Joana Xavier, Ramiro Magno, Marinella Ghezzo, Isabel Duarte and Filipa Esteves, without whom my scientific and technical progress would be tapered.

To Paulo Martel, for technical expertise, troubleshooting advice and availability when I required it the most. Furthermore, all the software developers who were available through GitHub to resolve small issues, Nuno Morais with psichomics, Alex Dobin with STAR and François Agnet with tensorQTL.

To my girlfriend, Catarina Garcia, for putting up with all the late-night work, busy weekends and overall unavailability. Her support and hot meals provided the energy to push forward.

For all the laughs, company and friendly competition stimuli provided, all the friends and colleagues who I've met in this master's degree. Without you it would get rather tedious.

A special thank you to all my family for concern for my wellbeing and comprehension and support displayed towards my decision to return to academia, even when it meant moving across the country.

This would not be complete without mentioning all the contributions made by all specimen donors who allow databases to be created and used to for good of medical care and science. Furthermore, all the curious people who came before and worked towards improving science and knowledge. We are working on the shoulders of giants that preceded us.

Abstract

Genome-wide association studies (GWAS) were pivotal in identifying genomic variants associated with susceptibility to breast cancer (BC). Given most identified loci are on non-coding regions, unidentified causal variants are predicted to act through cis-regulatory mechanisms. Post-GWAS era relies on functional analysis to identify causal and characterize how risk-associated variants modulate gene expression. To this end several in silico techniques are used associating molecular phenotypes with genotype but most of these focus on transcriptional processes.

With this work, I analyse the potential contribution of alternative splicing (AS) altering variants in breast tissue to BC susceptibility. To this end I used RNA-seq data from healthy breast tissue with matching sample genotype. Afterwards, I employed two different packages to independently quantify AS. Splicing Quantitative Trait Loci (sQTL) analysis was performed in order to identify sQTLs, variants associated with changes in splicing patterns. BC GWAS associated single nucleotide polymorphisms (hit-SNPs) were retrieved and co-localization analysis based on ancestry-specific linkage disequilibrium was performed to assess if any of the identified sQTLs is associated with GWAS hit-SNPs.

Three loci were identified where sQTLs are co-localized with BC GWAS hit-SNPs. In locus 1p36 variants rs4908724 and rs17229081 are associated with changes in splicing of PARK7, a protein deglycase previously identified as an oncogene. Regarding locus 11q13, 4 sQTLs – rs56984820, rs6591195 and rs9735063 – were identified, reporting changes in splicing patterns of BANF1, responsible for activating genome repair pathways interacting with PARP; with publications proposing as multi-cancer biomarker. Changes in ULK3 splice pattern were associated with variants rs12591513 and rs12898397 on locus 15q24. This gene is a regulator of Hedgehog pathway, whose dysregulation is associated with carcinogenesis. These changes in AS impact isoform ratios resulting in different protein and/or on UTRs, impacting other gene expression regulatory mechanisms. Further functional studies are required to identify causal variants as well as impacted cis-regulatory elements. Thus, variants may increase risk for BC modulating CRE of AS.

Keywords: Breast Cancer, Genome-wide Association Study, splicing Quantitative trait Loci, Breast Cancer Risk, Alternative Splicing.

Resumo

Diferentes metodologias têm sido implementadas com o objetivo de identificar variantes genómicas associadas ao aumento do risco de cancro da mama. A mais recente são os estudos de associação genómica (do inglês *genome-wide association studies*, GWAS), onde polimorfismos de nucleótido único (*single nucleotide polymorphism*, SNP) ao longo de todo o genoma são testados em um só estudo. No entanto, dado a forma como estes são conduzidos, não identificam uma variante causal, mas sim um locus genómico associado ao risco onde várias variantes se encontram em desequilíbrio de ligação (*linkage desiquilibrium*, LD). Adicionalmente, a maior parte dos *loci* identificados são em regiões não-codificantes do genoma, colocando-se como hipótese que estes afetam elementos de *cis*-regulação de mecanismos de expressão génica. Por forma a identificar as variantes causais, bem como caracterizar os mecanismos através dos quais estas modulam o risco, diferentes métodos como estudos de associação de genótipo ou alelo e fenótipo molecular (*quantitative trait loci*, QTL e *differential allelic studies*, DAS) são empregues. No entanto, estes têm se focado maioritariamente na modulação da expressão génica através de alterações na ligação de fatores de transcrição em elementos *cis*-reguladores, ignorando outros mecanismos.

Com este trabalho examino a potencial contribuição para o risco do cancro da mama por parte de variantes genómicas que alteram o *splicing* alternativo. Para tal uso informação de sequenciação de ácido ribonucleico (ARN-seq) de tecido mamário saudável de dadores para os quais estão disponíveis genótipos obtidos através do projecto *Genotype-Tissue Expression* (GTEx). Estes dados são processados de modo a remover enviesamentos técnicos e contaminantes conhecidos de sequenciação de ácido ribonucleico. Em seguida utilizei duas ferramentas informáticas que me permitiram quantificar *splicing* alternativo, LeafCutter e psichomics, usando a métrica de percentagem de inclusão de *splicing* (PSI). De seguida procurei associações entre alterações de PSI e genótipo de variantes próximas de cada evento de *splicing* utilizando o tensorQTL. De forma a reduzir o efeito de outras variáveis, utilizei os cinco componentes principais obtidos a partir da análise dos componentes principais nas contagens dos intrões. Os *loci* significativamente associados com alteração em eventos de *splicing* são denominadas de sQTL (*splicing quantitative trait loci*). Utilizando o pacote de R gwasrapidd, acedi aos registos do GWAS Catalog, um repositório online de estudos GWAS, de forma a obter todas as variantes genómicas que foram previamente associadas com risco de cancro de mama na população europeia. Utilizando

padrões de LD da população europeia, procurei co-localização entre os *loci* associados a cancro de mama e sQTLs, utilizando um limiar mínimo de LD ($r^2 \geq 0.4$).

Três *loci* diferente foram identificados, cada um com pelo menos um sQTL obtido por cada método de quantificação de *splicing* alternativo.

No locus 1p36, o gene *PARK7* foi identificado sendo o seu padrão de *splicing* dependente das variantes rs4908724 e rs17229081. Este gene produz uma deglicase de proteínas que foi previamente identificada como oncogene, inibindo a proteína PTEN, um gene supressor de tumores extensamente estudado. Alelos associados com risco parecem alterar a expressão das isoformas reduzindo o transcrito ENST00000338639 e aumentando os transcritos ENST00000493678 e ou ENST00000377493. Uma destas isoformas em particular, ENST00000377493, resulta numa alteração da dimensão da sequência codificante da proteína, com cerca de menos 20 aminoácidos. Adicionalmente as regiões transcritas e não traduzidas (UTR) em ambos os extremos dos transcritos também são modificados.

No locus 11q13, quatro sQTLs foram detetados alterando os rácios de *splicing* no gene *BANF1*. Este gene é responsável pela ativação de vias de reparação do genoma, interagindo com a PARP, para além de desempenhar funções na organização do genoma dentro do núcleo. Mutações no *BANF1* são comuns em diferentes cancros, sendo particularmente associados com cancros da mama triplo negativos, no quais não há sobreexpressão de recetores hormonais nem de receptores de fatores de crescimento epidérmico humanos (HER2). A expressão deste gene foi proposta como biomarcador de diversas doenças oncológica. O alelo de risco de cancro de mama é associado à redução das isoformas ENST00000533166, ENST00000312175 e ENST00000445560 sem identificar aumento de qualquer isoforma. Apesar de todas as isoformas terem sequências codificantes idênticas, as porções não-codificantes não o são, podendo o risco estar associado a diferentes níveis de estabilidade do ácido ribonucleico mensageiro de cada isoforma.

No locus 15q24, variação de *splicing* nos transcritos do gene *ULK3* foram associados com duas variantes, rs12591513 e rs12898397. ULK3 interage com diversas proteínas participando na regulação da atividade de PTEN, um gene supressor de tumores bem caracterizado. Alelos associados a risco parece reduzir as isoformas ENST00000440863, ENST00000566479 e ENST00000567472 e aumentar o rácio dos transcrito ENST00000569437, ENST00000561725 ou ENST00000568718. Comparando a sequência codificante dos transcritos que são traduzidos em

proteína, ENST00000440863 e ENST00000569437, estes distinguem-se na sua dimensão, reduzida em seis nucleótidos, que resulta numa proteína com menos dois aminoácidos.

Apesar de identificação de diversas alterações nos rácios de *splicing* dos genes, a caracterização quanto à função das isoformas alternativas é dificultada uma vez que os estudos de função de genes focam-se no fenótipo em resposta a alteração quantitativa da isoforma principal e não de variantes estruturais originadas posteriormente.

Variantes associadas ao aumento de risco de cancro da mama são associadas a alteração dos rácios de *splicing*. Estas podem exercer o seu efeito modulando elementos *cis*-reguladores de *splicing* e modificando a expressão das diferentes isoformas. Desta forma modulação de *splicing* alternativo parece ser um dos mecanismos modificados por variantes associados ao risco de cancro sendo, no entanto, um mecanismo com menor contributo que outros anteriormente estudados.

É importante salientar que os dados de ARN-seq que foram analisados provêm de um conjunto heterogéneo de tipos celulares, com programas de expressão genética diferentes sendo, portanto, a expressão medida uma média do conjunto. Identificação e análise de células individualmente tem potencial para identificar com maior precisão mudanças na expressão genética devido à presença de variantes genómicas.

Para além de mudanças na sequência codificante, alteração nas regiões não traduzidas tem potencial para modificar regulação pós-processamento como a localização sub-celular, estabilidade do ARN mensageiro e eficiência de tradução em proteína, como é sugerido pela dupla nomeação das variantes identificadas como associadas não só a alterações de *splicing* como também de expressão total. Estudos funcionais futuros serão importantes para identificar a(s) variante(s) responsável(eis) pela variação nos padrões de *splicing* identificadas. Adicionalmente, métodos de co-localização mais robustos aliados a estudos *in vitro* e *in vivo* irão clarificar se é por alteração da regulação dos mecanismos de *splicing* que ocorre o aumento de risco ou se a co-localização de variantes associadas a risco de cancro de mama e alteração de *splicing* é fortuita.

Palavras chave: Cancro de mama, Estudos de associação genómicos, Splicing Alternativo, Risco de Cancro da Mama, percentagem de inclusão de splicing,

Table of contents

List of Figures

List of Tables

Abbreviations

(fill the table and in the end just sort from A to Z)

3'SS	3' Splice Site
5'SS	5' Splice Site
A3SS	Alternative 3' Splice Site
A5SS	Alternative 5' Splice Site
AA	Aminoacid
AFE	Alternative First Exon
ALE	Alternative Last Exon
AMD	Acute Macular Degeneration
AS	Alternative Splicing
BED	Browser Extensible Data
BED	PLINK Binary Biallelic Genotype Table
BP	Base Pair
BPS	Branching Point Site
BPP	Branching Point Protein
CI	Confidence Interval
CpG	Cytosine-Phosphate-Guanine
CRAN	The Comprehensive R Archive Network
CRE	Cis-Regulatory Element
DAE	Differential Allelic Expression
DAS	Differential Allelic Studies
dbGaP	Database of Genotypes And Phenotypes
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
EJC	Exon Joining Complex
eQTL	Expression Quantitative Trait Loci
ESE	Exonic Splicing Enhancer
ESS	Exonic Splice Silencer
FDR	False Discovery Rate

GPU	Graphical Processing Unit
GRCh38	Genome Reference Consortium Human Reference 38
GTEx	Genotype-Tissue Expression
HDAC	Histone Deacetylase
Hh	Hedgehog pathway
HMT	Histone Methyltransferase
HRT	Hormone Replacement Therapy
IDE	Integrated Development Environment
Indel	Insertion/Deletion
IR	Intron Retention
ISE	Intronic Splicing Enhancer
ISS	Intronic Splicing Silencer
kb	Thousands Of Base Pairs
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MBP	Methyl Binding Proteins
MIT	Microtubule interacting and transport domain
mRNA	Messenger Ribonucleic Acid
MXE	Mutually Exclusive Exons
NA	Not Available
NES	Normalized Effect Size
NMD	Nonsense Mediated Decay
NT	Nucleotide
OR	Odds Ratio
PD	Parkinson Disease
Pre-mRNA	Precursor Messenger Ribonucleic Acid
pY	Polypyrimidine Tract
QTL	Quantitative Trait Loci
RBP	RNA Binding Proteins
RNA	Ribonucleic Acid
RNAPII	RNA Polymerase II

RNA-seq	RNA sequencing
ROS	Reactive Oxygen Species
SE	Skipped Exon
SNP	Single Nucleotide Polymorphism
snRNA	Small Nuclear Ribonucleic Acid
snRNP	Small Nuclear Ribonuclear Protein
sQTL	Splicing Quantitative Trait Loci
SV	Structural Variant
TF	Transcription Factor
TNBC	Triple Negative Breast Cancer
TNM	Tumour, lymph Nodes, Metastasis
TPM	Transcripts Per Million reads
TSS	Transcription Start Site
VCF	Variant Call Format
WT	Wild Type

1 Introduction

1.1 The first Genome projects

Providing the first complete reference for the human DNA sequence, The Human Genome Project also brought to light the magnitude of variation of the human genome within the population (Collins, Morgan and Patrinos, 2003). This was the seed that originated other projects such as the HapMap project and 1000 Genomes Project, with the aim to catalogue human polymorphisms to explain much of the inherited population diversity and disease risk. To date, the known number of variants has expanded as well as the knowledge of their frequency and global distribution, revealing patterns of divergence among different populations (Frazer *et al.*, 2007; Altshuler *et al.*, 2012; Auton *et al.*, 2015; Nguyen *et al.*, 2016). These differences occur naturally within populations and provide resilience to adversity, either disease or other environmental pressure. But not all variants provide advantageous characteristics, some increase the risk for disease, mainly for complex diseases. How exactly does the human genome diverge?

1.1.1 Population Genetics

The latest release of the human reference genome, compiled by the Genome Reference Consortium in 2019, is composed of 3.1 billion base pairs (Genome Reference Consortium, 2019) and the average human diverges at around 4.3 to 5 million sites (Auton *et al.*, 2015; Yates *et al.*, 2020). Apart from homozygotic twins, generated from the same zygote, all humans have a unique genetic footprint originated from the combination of millions of genetic polymorphisms inherited from their progenitors (Goldman and Domschke, 2014; Auton *et al.*, 2015).

An allele is the designation for one of at least two versions of the same variant. The most frequent type of variant is termed single nucleotide polymorphism (SNP), a substitution of a nucleotide for another, followed by single nucleotide insertions or deletions (indels). Insertions and deletions are particularly common in repetitive sequences, such as satellites, tandem repeats whose number of repetitions varies (Pearson, Edamura and Cleary, 2005). Together SNPs and indels account for 99.9% of variants found in the human genome. Larger sequence variants are labelled

structural variations (SV), an umbrella term that encompasses alterations that affect stretches of more than 50 base pairs. SVs can be insertions, deletions, as well as inversions, duplications, translocations or copy number variants. Although relatively rare when compared to SNPs (median number 4.3 million SNPs vs 2 100 SVs per genome), due to their size spanning up to 5 million bp, they affect a larger proportion of the genome (estimated ~ 20 million bp) (Altshuler *et al.*, 2012; Auton *et al.*, 2015; Sudmant *et al.*, 2015).

Genomic variants arise from mutations during embryogenesis or gametogenesis that are passed on to descendants. On average, each individual display 40 *de novo* mutations. Although most of these new mutations are lost in a few generations, due to selective pressure and chance, some are still successfully passed through generations increasing its frequency in the population. Hence, it is easy to understand that under normal circumstances common variants are significantly older than less frequent ones, and that the number of variants increases dramatically as the frequencies lower (Raychaudhuri, 2011; Auton *et al.*, 2015; Hernandez *et al.*, 2019).

There are differences in the frequencies of polymorphic variants across the globe. It is estimated that the typical genome of an individual of African ancestry deviates from the reference genome in approximately 5 million variant sites, representing the greatest genetic diversity, compared to around 4.1 million in individuals of European ancestry. This is partially explained by the out-of-Africa model of human origins, where small groups with a restricted pool of genomic variants expanded across and populated other continents (Altshuler *et al.*, 2010; Auton *et al.*, 2015).

1.1.2 Variant vs mutation

A point of discussion is the conceptual difference between variant and mutation with some authors making a distinction based on frequency of the minor allele frequency (MAF, < 1% indicates mutation) or consequence/pathogenicity (if it is pathogenic it is considered a mutation), with the previous being the generally accepted definition (Karki *et al.*, 2015). Nowadays, an arbitrary cut-off is applied where common variants are those whose MAF in a population is higher than 0.05, rare variants those between 0.01 and 0.05 and mutations as those below 0.01 (Altshuler *et al.*, 2010; Raychaudhuri, 2011). It is important to point out that alternative allele is not a synonym

of minor allele. While minor allele(s) describes the allele(s) of lower frequency, which is a population dependent metric, alternative allele is any allele that is not the reference for that genome.

The Hardy-Weinberg law describes the relationship between allele frequency and genotype frequency within a population under normal circumstances. If we consider $\boldsymbol{a}$ as the alternative allele with frequency $\boldsymbol{q}$, and $\boldsymbol{A}$ the reference allele with corresponding frequency $(1 - \boldsymbol{q})$. According to Hardy-Weinberg, for diploid individuals, the frequency of reference homozygotes (AA) should be $(1 - \boldsymbol{q})^2$, for heterozygotes ($\boldsymbol{Aa}$) it is $2 \times (1 - \boldsymbol{q}) \times \boldsymbol{q}$, and for alternative homozygotes ($\boldsymbol{aa}$) $\boldsymbol{q^2}$. Applying it to discover the expected frequency of genotypes derived from a common variant whose MAF, or $\boldsymbol{q}$, is 0.05 it returns a population frequency of 0.90 for the common homozygote, 0.095 for the heterozygote genotype and the alternative homozygote genotype of just 0.0025 of the whole population. In other words, this means that in a population of 400 individuals only one individual will carry a homozygotic alternative genotype. This illustrates the difficulty to find individuals of all genotypes in sufficient number to carry reach statistical power, even more for rarer variants (1 in 10 000 chance of a random person being homozygous alternative for a population MAF or $\boldsymbol{q}$ = 0.01).

Some populations are derived from a small group of individuals, a founder population, and present increased rates of consanguinity diverging from Hardy-Weinberg equilibrium. This allows that otherwise rare alleles that were present on the initial group to have an increased frequency when compared to other ancestries. Examples are the Finnish and Islandic populations. This can be leveraged to study the effect of rarer alleles on specific traits and overcome the lack of samples for rare homozygotes (Claussnitzer *et al.*, 2020).

1.1.3 Haplotype and Linkage Disequilibrium

Due to how variants emerge in populations, human migration patterns and demographics around the globe, early variants are generally more widespread across all populations and have a higher MAF than more recent variants, with some of these newer variants being ancestry specific with little or no representation in other populations.

Nevertheless, all these variants can be physically close in a chromosome, and subsequently inherited together, which generates a combination pattern of alleles at different SNPs. Linkage disequilibrium (LD) is the measure of how likely it is to have an allele given the status of the another known one (Belmont *et al.*, 2005). It is represented by the factor r^2, which is based on the

Pearson correlation coefficient between two alleles. It scales from zero to one ,with higher values indicating that the alleles are more likely to be inherited together, and zero meaning that there is no relationship between them (Wall and Pritchard, 2003). When, for example, LD $r^2 > 0.8$, many alleles are so closely correlated that they can define a haplotype block. The knowledge of population specific patterns of LD allowed techniques like *imputation*, assessing unknown alleles from the neighbouring measured ones, and *phasing*, knowing which alleles are in the same physical chromosome. Genotyping microarrays were developed using tag-SNPs, SNPs that due to being within a high LD block are able to inform on the sample's haplotype. In combination with powerful bioinformatic techniques, the use of tag-SNP in genetic association testing allowed a significantly decrease in the cost of genotyping (Belmont *et al.*, 2005; Frazer *et al.*, 2007; Altshuler, Daly and Lander, 2008; Altshuler *et al.*, 2010; Auton *et al.*, 2015). These haplotype blocks can be broken by recombination events during meiosis originating an LD decay. The decay rate is increased by the number of generations, population size and decreases with the number of individuals in the founder population. This helps explain why haplotype blocks are smaller in populations of African ancestry than in any other population (Bush and Moore, 2012).

1.2 Genetic Variants and Disease

It is believed that the variation on the human genome is responsible for phenotypic differences contributing to variation in human traits, from eye colour or height to disease risk. Over the years, several different projects sought to better understand the relationship between phenotype and genotype – with special interest in complex diseases (Ardlie *et al.*, 2015; Romanoski *et al.*, 2015; Schinzel, 2015; Moore *et al.*, 2020). Understanding how each genomic variant contributes to the phenotype may reveal underlying mechanisms allowing to create strategies for prevention, new tools for improved diagnosis, better treatment and overall improve life quality (Altshuler, Daly and Lander, 2008; Lappalainen *et al.*, 2013).

For some time it was common belief that these genetic variants with phenotypic effects were non-synonymous and acted by altering protein coding sequence with consequences on protein function (Cookson *et al.*, 2009). However, most of the variants associated with variation on traits are in non-coding regions of the genome (Ma *et al.*, 2015), and are particularly enriched in cis-regulatory elements (CREs), generating changes in gene expression (Maurano *et al.*, 2012; Schaub

et al., 2012). Thus, variants can modify CREs by increasing or decreasing the affinity of trans elements or, in extreme cases, creating or destroying CRE (Cookson *et al.*, 2009; Ma *et al.*, 2015).

1.2.1 Family Linkage Studies

The first studies to assess the relationship between genotype and phenotype were performed in rare diseases with mendelian pattern of transmission in the 1980's. These are monogenic and syndromic diseases that result from a rare allele exerting its effect over a single gene and follow Mendel's expected ratios in families and communities. Using available technology – Sanger sequencing and restriction fragment length polymorphism analysis – genetic polymorphisms were used as markers to identify which gene(s) was(were) linked to the disease, hence narrowing the list of possible causal variants responsible for the observed phenotype in affected family members. The first disease successfully characterized was Huntington's disease, soon followed by others such as familial cancer syndromes, hypertension and diabetes (Altshuler, Daly and Lander, 2008; Claussnitzer *et al.*, 2020). The identified mutations were associated with a strong predisposition to these diseases, but only explained a few cases of the affected population. Other mutations with lower penetrance or locus heterogeneity were more difficult to identify. Furthermore, many diseases had secondary and tertiary genetic modifiers, besides environmental factors, which altered disease severity resulting in confounding phenotypes (Scriver and Waters, 1999; Weatherall, 2001). Of the estimated 7,000 single-gene inherited diseases, this approach identified ~ 1,000 genes (Claussnitzer *et al.*, 2020).

Although this type of studies was revolutionary in shedding a light into specific disease-allele combinations, the most common phenotypes are complex and multifactorial, with each individual variant contributing with a modest effect to the total variability of the characteristic of interest (Altshuler, Daly and Lander, 2008).

1.2.2 Candidate gene studies

With better understanding of cellular basic biology and disease mechanism, some genes appeared to be logic candidates to have an impact in diseases. The candidate genes approach relies on researchers predicting which genes might have an impact on disease aetiology. Variants in or near those genes are mapped and those with hypothetical functional role are tested. To determine an association between the presence of a variant and the phenotype, the test population is divided

in cases and controls and correlation is determined. Examples of successful studies identified genes encoding glucose uptake proteins in diabetes and DNA repair in cancer. However, the results have been criticized due to the limitations on replication of results, as variant mapping is performed in a specific population as well as due to the inability to consider all genes and variants for study (Tabor, Risch and Myers, 2002).

1.2.3 Genome-Wide Association Studies

With increasing knowledge of the catalogue of genetic variation and cheaper genotyping technology available, genome-wide association studies (**GWAS**) studies were made possible. Taking advantage of population LD patterns, population specific microarrays were designed with tag-SNPs representing several SNPs and SVs in high correlation, which allowed tagging (most of) the genome with comparative cheaper cost than other methods. Initially using only approximately 500k SNPs, these microarrays accurately genotyped individuals across > 90% of the most common SNPs in non-African populations. The genotypes of the SNPs that were not directly measured on the microarray could be imputed from the tag-SNP information (Altshuler, Daly and Lander, 2008). Later, at each genetic marker, genotype or allele content is tested for association with the phenotype, where the presence of each variant gives information on the probability of phenotype (Lewis and Knight, 2012; Altman and Krzywinski, 2015). The GWAS main aim is to perform unbiased genome-wide identification of variants that have a role in complex phenotypes. When the phenotype in study is a disease, this research would allow the prediction on who are the more susceptible individuals as well as discover what genes and biological networks are affected, opening routes of investigation onto new strategies for prevention and treatment (Bush and Moore, 2012). The first successful study was on acute macular degeneration (**AMD**) which identified variants impacting the *CFH* gene, responsible for the production of Complement Factor H protein. Further analysis onto haplotypes allowed the identification of rs1061170, a variant in exon 9 of *CFH* as the best candidate.

The effect size of the variants associated with risk if often reported as the odds ratio (**OR**), a measure of the effect of an exposure to an outcome. It is calculated as $OR = \frac{(exposed\ cases)/(unexposed\ cases)}{(exposed\ non{-}cases)/(unexposed\ non{-}cases)}$. An OR above one implies that exposure is a risk factor for the outcome, while below one suggests exposure is protective for the outcome (Szumilas, 2010). In the previously cited studies on AMD, OR for the risk allele C of rs1061170 is between

2.45 and 7.4, and it was estimated that this variant is responsible for 43% of all cases of AMD (Edwards *et al.*, 2005; Haines, 2005; Klein *et al.*, 2005; Claussnitzer *et al.*, 2020). Another common metric used is the Confidence Interval (**CI**), employed to assess precision of the measured OR and treated as a proxy for significance. If CI includes the value one it suggests a lack of significance, while a CI with a wide interval reveals low confidence on the measured OR (Szumilas, 2010). In the case of previously reported AMD, risk allele showed 95% CI range from 1.41 – 4.25 to 3.0-19 depending on publication (Edwards *et al.*, 2005; Haines, 2005; Klein *et al.*, 2005).

Contrary to findings on AMD, most of the variants identified by GWAS have an effect size smaller than two per allele, most commonly in the 1.2 to 2 range (Bush and Moore, 2012). Considering most complex phenotypes show high degree of inheritability, small effect size per allele implies that a great number of variants are involved in these phenotypes (Hansen, 2002; Zdravkovic *et al.*, 2002; Ridge *et al.*, 2013; Mucci *et al.*, 2016). As such, phenotypes – specially diseases – should not be seen as binary, either present or absent, but as a continuum with higher severity and early age of onset being, in part, the result of the combined effect of risk alleles (McClellan and King, 2010).

Given the number of statistical tests performed when thousands and millions of variants are tested in one study, a strong multiple test correction as Bonferroni is required. The average GWAS performs 10^6 tests changing the threshold of significance from the generally accepted $\alpha = 0.05$ to a stricter 5×10^{-8}. Although the number of tested variants can be greater, this correction assumes that there is no relationship between variants, which we know is untrue as measured by LD. In 2012 researchers collected GWAS data on borderline associations, defined as those whose p-value of association between phenotype and genotype were $10^{-7} \geq P - value > 5 \times 10^{-8}$. Using data available from subsequent studies, they evaluated 26 variants, of which 19 reached statistical significance on replication data (Panagiotou *et al.*, 2012). As an alternative, false discovery rate (**FDR**) is a less conservative method that can be applied. It stems from the knowledge that, by chance, a fraction of all tests performed equal to our significance threshold α, typically 0.05, will be false positives. FDR procedure adjusts the p-value of significant findings, a q-value, considering the expected number of false positives by chance. Permutation tests can also be employed but due to requiring more computational resources, it is often disregarded. It retrieves an empirical distribution of tests statistics randomly assigning samples to case and control and compares with the value obtained to see how significant that association is (Bush and Moore, 2012).

A typical representation of GWAS results is a Manhattan plot, with the X axis representing the position of each SNP along the genome and the Y axis representing the p-value of association between genotype and trait.

However, the reliability of GWAS studies depends greatly on their statistical power, defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true given the design of the study. In the context of GWAS, it means the probability of correctly identifying true associations between phenotype and genotype. Statistical power in GWAS is dependent on different factors as phenotype prevalence, minimum MAF threshold to be tested, number of variants and LD between genotyped and imputed variant, number of cases and controls and their ratio, among others. Furthermore, bias on population stratification, ancestry and unclear phenotype further reduce study power, possibly inducing false positives. From these, only the number of samples and their ratio can be controlled during study design stage. In this context, 80% statistical power is regarded as the best compromise to reduce false negatives and costs. In order to achieve such a power, a sample size of 9,962 is required in a 1:1 ratio between cases and controls, when considering a MAF of 0.05, complete LD ($r^2 = 1$), 1 million variants tested, an $\alpha = 0.05$ and a per allele OR of 1.3 (Altshuler, Daly and Lander, 2008; Hong and Park, 2012).

Nowadays, with the (relatively) reduced cost and increased throughput of whole exome and genome sequencing, as well as availability of genomic data consortiums (e.g.: UK Biobank (Bycroft *et al.*, 2018)), the number of samples per study has increased with GWAS surpassing 1M samples (Harati *et al.*, 2019; Mills and Rahal, 2019). As a result, association testing can be extended to lower frequency alleles, previously unsurveyed in earlier GWAS studies (Claussnitzer *et al.*, 2020).

When considering biological pathways, GWAS have confirmed many of the expected genes with phenotype associations, but have also revealed novel connections. The integration of genotype-phenotype databases in addition to multi-study meta-analysis allowed the retrieval of many risk SNPs for a wide range of diseases. Databases like the GWAS Catalog, an online repository of GWAS studies for multiple phenotypes, are nowadays available to researchers worldwide. In August 2020 GWAS Catalog reported 4,681 studies with more than 197k statistically significant associations. In order to facilitate data retrieval, API level access to curated data is also now possible (Buniello *et al.*, 2019; Magno and Maia, 2019).

1.2.4 The challenges with GWAS

Although GWAS provide more insight into complex diseases they are still presented with several challenges: 1) they do not directly identify the causal variant(s), as the tag-SNPs used are proxy variants for all other SNPs in strong LD with them, and so all associated variants for any given result can be located on loci spanning up to 1M bps; 2) many GWAS loci are in non-coding regions of the genome with unknown function nor affected gene; 3) most of loci found have small effect on the disease with unknown interaction amongst them 4) the combined effect with environmental variables is unknown; 5) as is the underlying disease mechanism (Weatherall, 2001; McClellan and King, 2010). This unbiased strategy facilitated the identification of variants on candidate genes for several diseases but also unforeseen connections between genome deserts and different phenotypes. What function could these regions perform to be associated with changes in phenotype? Moreover, gene expression and regulation are tissue/cell-type and development stage specific and may yield small effects. Knowing which cells are being affected by the presence of the variant and how it contributes to the trait in study is not always linear making disease relevant tissues hard to study (Gallagher and Chen-Plotkin, 2018; Walker *et al.*, 2019).

Increasing the complexity of this task is the fact that the causal SNP may be in incomplete LD with the tag-SNP. To this end, further resequencing and fine mapping has been carried in many loci to disentangle independent risk variants (Gallagher and Chen-Plotkin, 2018).

As previously mentioned, lower frequency variants require a higher number of individuals to be tested in order to reach a significant power to detect association. Due to natural selection, it is expected that lower frequency variants and mutations have the most deleterious effect when the phenotype in question is a disease, but gathering enough samples for cases and controls of each genotype may prove to be a difficult task (Altshuler, Daly and Lander, 2008). However, increasing the number of samples in a GWAS comes with a caveat. Although variants with a small effect on the trait at study will be unearthed, these may exert their effect through complex biochemical regulatory networks, making them non-actionable targets as they only explain a minor percentage of the phenotype individually (Popejoy and Fullerton, 2016; Ewen Callaway, 2017).

Populational representativity is a subject of discussion amongst researchers as most samples is reported studies, as much as 0.86, are from European ancestry. This is in part for technical reasons as, by design, GWAS tend to leverage ancestry specific LD patterns, with US, UK and Iceland providing 71.8% of all samples used in GWAS. A bias in recruitment can result in a loss

of information for lower frequency alleles more represented in other populations. The inclusion of other ancestries may provide new meaningful knowledge onto phenotypes and disease aetiology (Mills and Rahal, 2019; Claussnitzer *et al.*, 2020).

1.2.5 Post-Gwas analysis - from association to function

Interestingly, the majority of risk loci are in non-coding regions of the genome, pointing towards a regulatory role of gene expression (Maurano *et al.*, 2012). Different projects were set to map all the polymorphisms of human population (Yates *et al.*, 2020), help decode what these elements are (Moore *et al.*, 2020), how they modulate gene expression, including tissue specificity, (McLaren *et al.*, 2016), and which are the target genes (Ardlie *et al.*, 2015; Gallagher and Chen-Plotkin, 2018). The availability of multi-tissue gene expression data with matching genotype information funded by public institutions allowed further research into genomic regulatory sites and how these are modulated (Ardlie *et al.*, 2015).

1.2.6 Molecular Phenotype-Genotype Association Studies

To further understand the genotype-phenotype association and better comprehend the molecular mechanisms driven by functional genetic variation, including those possibly implicated in disease loci from GWAS, different methods may be used. These are based on finding statistically significant associations between quantifiable molecular features and genotype, or allele, in a tissue specific manner (Mostafavi *et al.*, 2013; Ongen *et al.*, 2016), and are commonly either quantitative trait locus (QTL) mapping or allele-specific studies .

In QTL analysis individuals are binned according to genotype at a given locus and a linear regression is performed trying to correlate minor allele content (either 0, 1 or 2 minor alleles) with the measured molecular phenotype, which can be overall gene expression (eQTL, expression quantitative trait loci) or splicing isoforms (sQTL, splicing quantitative trait loci), among others (Park *et al.*, 2018). Although the model used may have a variable degree of complexity accounting for multiple variables, the effect size – impact of the genotype on the phenotype – is given by the slope of the linear regression, with an error component and the base level of measured phenotype being given by the residual value.

Allelic-specific studies are performed using only heterozygotic individuals for a given variant. The molecular phenotype is measured in an allele-specific manner and differences are

measured comparing the molecular phenotype that is the product from one allele vs the alternative allele. A limitation in allele-specific expression studies is that the heterozygous SNP must be expressed or its impact can be significantly harder to evaluate.

Albeit complementary, both approaches still have a unique set of advantages and disadvantages. While in QTL studies all individuals are informative, hence a lower number of total individuals might be sufficient, they are susceptible to the effect of trans factors, molecules that are coded elsewhere in the genome and influence the measured molecular phenotype, as well as other unknown biological/environmental bias. On the other hand, allele-specific studies are performed comparing molecular phenotypes from the same individual, i.e. with the same genetic and environmental context, which provides an internal control (Park *et al.*, 2018). Nevertheless, depending on the MAF of the variant, a larger population might be necessary to identify a sufficient number of informative heterozygotes.

Given LD patterns amongst variants within a haplotype block, all variants in high LD with the variant associated with a QTL or allele-specific study are themselves associated with the measured alternative splicing events.

To understand how variants can impact molecular phenotypes we must first understand the mechanisms at play that modulate gene expression: how can genomic variants impact gene expression?

1.3 Gene Expression Regulation

The human genome is encodes for approximately 20k protein coding genes and many more non-coding genes, whose expression is tightly regulated by a series of processes and mechanisms (Hunt *et al.*, 2018; Audano *et al.*, 2019; Mudge *et al.*, 2019). The number of transcripts increases dramatically when we take into consideration the number of isoforms and alternative transcripts each gene can present (Pan *et al.*, 2008; Wang *et al.*, 2008; Barash *et al.*, 2010). Although cells in the body present virtually the same genome – with the exception for gametes – they do not produce the same proteins at all times. Throughout differentiation, cells communicate with each other and the environment allowing, through a series of synchronized signals, gene expression shifts so that cells can specialize. While some genes are essential for cellular function, others are only expressed in particular cell types. Furthermore, regulation complexity is increased when we consider that genes are not linearly distributed in the genome, can be overlapped, on opposing strands, or even

generate multiple transcripts from the same gene (Atkinson and Halfon, 2014). This regulation grants space and time specific activation and deactivation of RNA and proteins production, in a concerted manner among cells that constitute the organism generating a myriad of phenotypes by modifying gene's expression patterns allowing cellular differentiation into tissues and organs (Chen and Dent, 2014).

In the past decades, gene expression regulation has been studied generating a wealth of information that decoded the signals and effectors that coordinate these regulatory processes (Gamazon *et al.*, 2018; Yee *et al.*, 2019; Moore *et al.*, 2020). All of these gene expression regulatory processes have two basic input categories: cis-regulatory elements (CRE) that are part of the DNA/mRNA sequence in (relative) proximity or part of the regulated gene which will only affect local allelic expression; and trans elements, such as proteins and other molecules, that are encoded elsewhere in the genome and, by interaction with consensus motifs, can interact with CRE of multiple genes producing a wider effect and affecting both alleles (Xiao and Scott, 2011). In a response to stimuli, either external and internal, trans elements act like a switch: activity is regulated by inducing conformation changes, post translational modifications or others which will change the way these interacts with proteins or with nucleic acids (as CRE). Considering each cell type has a distinct gene expression program, the same signal or even protein modification can result in entirely unique consequences in different cell types, creating the necessity for the study of gene expression and regulation in a tissue specific manner.

1.3.1 Epigenetic and Transcription Regulation - from genome to RNA

1.3.1.1 How DNA accessibility conditions gene expression

The first layer of control is epigenetic regulation. It provides a broader control of gene expression without altering the genetic sequence. It includes chromatin structure, DNA methylation and non-coding RNAs which modulate proteins and other molecules that access the DNA sequence by tightening or loosening the compression of the DNA molecule around nucleosomes (Non and Thayer, 2019). One example of such control is cytosine methylation by DNA methyltransferases (DNMTs) at CpG islands. CG dimers are less frequent then expected when observing overall proportion of guanine and cytosine but are enriched in known transcriptionally active sites

(Romanoski *et al.*, 2015; Andersson and Sandelin, 2020). In a simplified manner, methylated cytosines provide an anchor that is recognized by methyl binding proteins (MBPs), whose function is to summon histone methyltransferase proteins (HMT) and histone deacetylase (HDAC). In turn, these histone modifying proteins will perform post-translational modifications of histone amino acids, locally compressing the chromatin thereby blocking DNA accessibility and preventing gene expression (Jaenisch and Bird, 2003; Chen *et al.*, 2017). These and other chromatin modifications create three distinct states, transcriptionally active, inactive and poised genes with the last having both marks from an open and condensed state (Kolovos *et al.*, 2012; Atkinson and Halfon, 2014). Major changes in chromatin structure occur mainly at early stages of cellular differentiation, shifting from a genome-wide open chromatin status to a more restricted state creating two chromatin states in the nucleus, heterochromatin which is more compact and physically closer to the nuclear periphery and a more transcriptionally active euchromatin closer to nuclear centre, which due to its loose conformation allows interactions between DNA sequences and nuclear molecules (Young, 2011; Chen and Dent, 2014; Solovei, Thanisch and Feodorova, 2016). It is important to state that epigenetic regulation is a dynamic process played by all its actors and deregulations of its "normal" state can induce changes in gene expression. One example is aging in which a genome wide hypomethylation occurs, with increased frequency at the edges of CpG islands. Other factors such as food regimens poor on folate or other vitamins were also associated with decrease methylation across the genome. Hypomethylation allows the expression of silenced genes as *c-ras* and *c-myc* that are relevant in cancer (Jaenisch and Bird, 2003).

From the brief previous introduction one can understand how variants present in CpG islands have an impact on epigenetic marks. Differential methylation patterns among groups of different genetic background with the same general environment were found, in particular in enhancer and promoter regions owing to its regulatory role (Bell *et al.*, 2011; Non and Thayer, 2019). An example of association between genotype, epigenetic marks and gene expression is present in *SPATC1L*. A single SNP (rs8133082, G > T) was associated with increased methylation marks in four different promoter-associated CpG islands which in turn were inversely correlated with gene expression. This association was confirmed in different tissues crossing tissue associated methylation patterns (Bell *et al.*, 2011; Heyn *et al.*, 2013).

1.1.1.1 Stop/Go signals on gene transcription

Coupled with epigenetic regulation, transcription is one of the better studied mechanisms. It is mediated by the promoter, short motifs on the DNA sequence – TATA box, BRE, etc. near the gene transcription start site (TSS) – which are recognized by RNA Polymerase II (RNAPII) and general transcription factors (TF) inducing assembly of transcription machinery. Although these elements are enough to start transcription, it only occurs at a basal level, requiring other elements to modulate transcription regulation (Atkinson and Halfon, 2014).

Enhancers and silencers are elements in the DNA sequence distal to promoters which, like the name suggests, assist enhancing or silencing the transcription of the gene. These are recognized by TF, co-activators and co-repressors based on a molecule specific consensus motif. Even though enhancers/silencers and promoters can be far apart (up to a million base pairs up- or downstream of TSS), due to how the chromatin folds inside the nucleus centre, it creates topological associated domains where genome-wise distant elements come into contact bridging the gap between enhancers/silencers and promoters. These loops allow interaction of sequence bound proteins and recruitment, assembly of transcription complex or transcription start is the case of enhancers. Notice that some authors consider as a proximal promoter a region up to 350bp from TSS where TF bind (Andersson and Sandelin, 2020). While historically the distinction between enhancer and promoter is established on the basis of RNAPII binding and gene TSS, their sequence and chromatin architecture are similar, granting promoters enhancer activity and enhancers that can induce transcription at their boundaries mudding the distinction between both (Lenhard, Sandelin and Carninci, 2012; Haberle and Stark, 2018; Andersson and Sandelin, 2020).

One other DNA element that is equal as important are insulators. These act as a barrier creating boundaries for heterochromatin expansion and preventing promiscuous interactions between enhancer/silencer and promotor (Atkinson and Halfon, 2014).

From all of these regulatory elements present in DNA sequence, one can reason how polymorphisms on the genetic code can hinder recognition of CREs of transcription by trans molecules. One example is rs1421085, a SNP that disrupts a repressor motif that is recognized by ARID5B in adipocytes. In the presence of the alternative allele there is an increased expression of *IRX3* and *IRX5*that generates an increased pre-adipocyte differentiation while lowered mitochondrial thermogenesis leading to increased bodyweight and obesity (Claussnitzer *et al.*, 2015).

1.3.2 Pre-mRNA processing and Alternative Splicing

During transcription, while the new RNA molecule is being produced, pre-processing takes place adding features to the recently formed molecule. Transcription and processing are highly correlated suggesting a strong co-regulation across tissues (Wang *et al.*, 2008). The RNAPII presents a C-terminal domain that interacts with a multitude of proteins allowing co-transcriptional processing of the emerging pre-mRNA strand (Zaborowska, Egloff and Murphy, 2016).

The first step is the addition of a methyl-guanosine to the 5' end, a process called capping. It allows the interaction with the Cap Binding Complex that will promote further processing, guide towards nuclear exportation and translation of the m-RNA. The next process is splicing, through which introns are removed and adjacent exons are joined forming a smaller mRNA strand. The final stage is RNA cleavage and tailing, where the 3' end of the pre-mRNA molecule is removed and a chain of terminal nucleotides are added. Although RNA tailing is canonically performed by poly(A)polymerases, which adds ~200 adenines, other terminal nucleotidyltransferases, as uridyltransferases, are known to exert its functions adding alternative nucleotides to previously cleaved termini (Herzel *et al.*, 2017; Liudkovska and Dziembowski, 2020; Yu and Kim, 2020). In all of these mechanisms, CRE exert their influence providing guidance on where trans molecules should perform their actions. Genomic variants present in any of these cis-regulatory elements can alter processing machinery recognition inducing changes in these processes (Manning and Cooper, 2017).

1.3.2.1 Splicing mechanism and regulation

Splicing mechanism allows post-transcriptional modification of the pre-mRNA strand and is mostly performed by the spliceosome. It combines the exons, 80% of which are smaller than 200 nts, while removing introns, averaging 3k nts in length, with more than 10% spanning past 11k nts (Webb *et al.*, 2014). This process removes on average > 90% of the primary transcript (Wang and Burge, 2008).

The spliceosome is a large and dynamic complex composed of up to ~ 250 molecules. It is formed by up to five small nuclear RNA (snRNA) – U1, U2, U4, U5 and U6 – each interacting with, at least, seven different proteins constituting small nuclear Ribonuclear Proteins (snRNP). Besides snRNPs the spliceosome complex also contains up to 170 spliceosome-associated factors,

with some playing a role in multiple gene expression mechanisms. Webb et al. describes in great detail each of the molecules as well as how they interact with each other (Webb *et al.*, 2014).

snRNPs are trans-elements that recognize and bind to cis-elements, splicing signals on the immature mRNA strand via consensus between the pre-mRNA and the snRNA. The basic cis-elements are the 5' splice site (5'SS or donor site), the 3' splice site (3'SS or acceptor site) that constitute the boundaries between exon and intron. By comparing the known exon/intron borders of proteins mRNA sequence, a consensus sequence was derived with an almost always present GT (GU in mRNA) at the 5'SS and AG at 3'SS. But these elements don't define themselves just by it, its consensus sequence extends beyond the dinucleotides present at the exon/intron boundaries as seen in the image below (Figure 1.1 (Levine, 2001)). Two other elements are as essential in spliceosome assembly and splicing reaction itself. They are the Branching Point site (BPA), an adenosine whose chemical structure is a pivotal part of the splicing reaction, and the polypyrimidine tract (pY), between the 3'SS and BPS, composed of cytosines and thymidines that help guiding spliceosome assembly. As can be seen in the consensus sequence for 3'SS (Figure 1.1), BPS and pY are frequetly within 50 nts from 3'SS and are as important for the splicing mechanism as the 5'SS and 3'SS (Webb *et al.*, 2014).

The splicing reaction itself consists of two sequential transesterifications promoted by the spliceome. It begins with (**1**) recognition and assembly of U1 onto 5'SS and BPP (Branching Point Protein/SF1) and U2AF (U2 Auxiliary Factor) to bind to BPS and pY respectively. Sequently (**2**) BPP recruits U2 snRNP and is displaced along with U2AF allowing interaction between U2 and BPS. At this point the (**3**) tri-snRNP sub-complex enters the reaction as a heterotrimer. U6 substitutes U1 at the 5'SS and the catalitic complex is now assembled. (**4**) Using the 2'-OH of the adenosine in the BPS, a nucleophilic attack occurs between BPS and 5'SS phosphate cleaving the phosphodiester bond and creating a loop within the inton. This structure is called a lairat and chains together the 5' guanosine at the phosphate residue and the BPS adenosine. The transesterification reaction creates (**5**) a 3'-OH on the end of the upstream exon that, with the help of the Exon Joining Complex (EJC), attacks the 3'SS stitching the exonic sequencing together while removing the intron as a lariat (Levine, 2001; Will *et al.*, 2001; Wongpalee and Sharma, 2014).

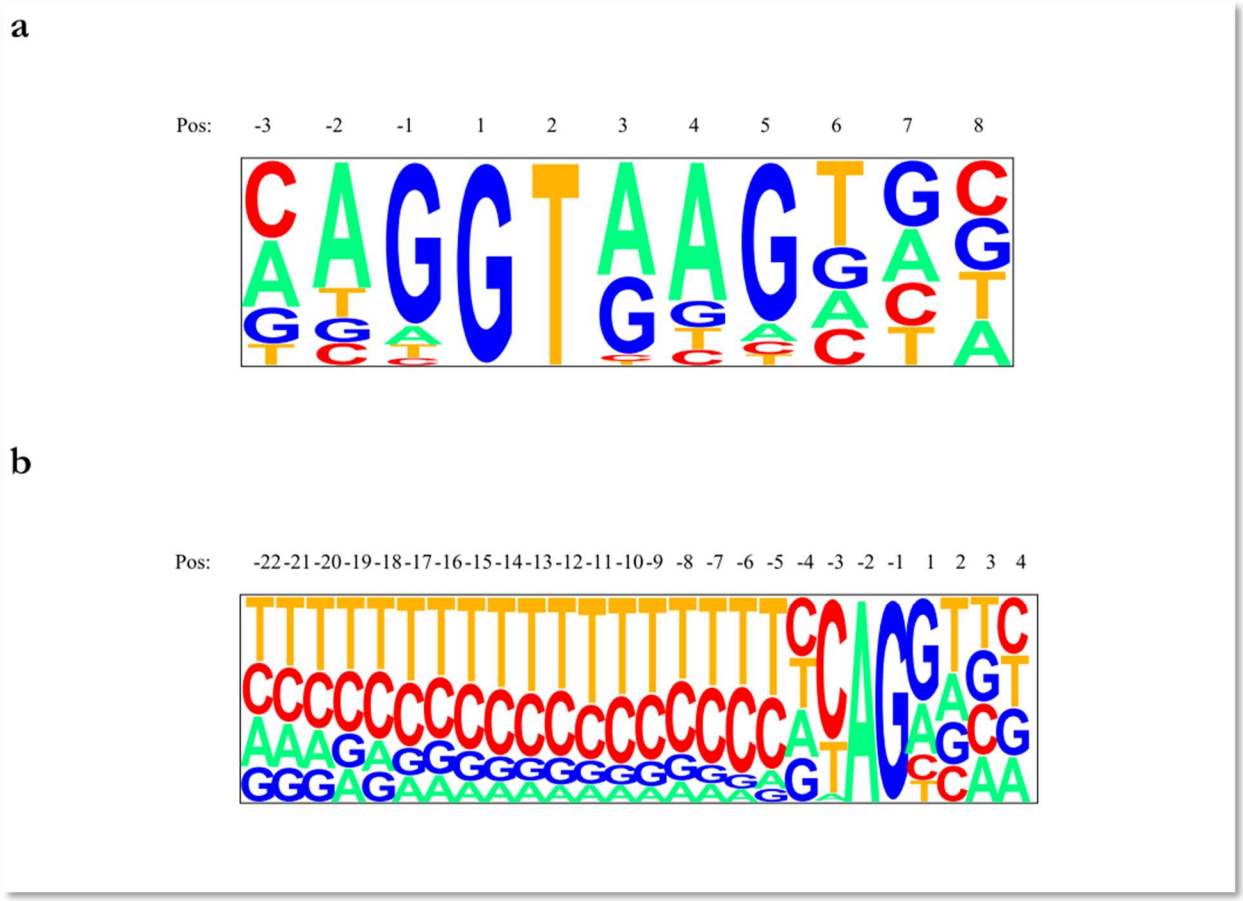


Figure 1.1 - Consensus sequence for splice sites. Donor site (a) and acceptor site (b) are represented by their consensus motifs by screening of known exon/intron bounderies. The size of each letter is proportinal to the frequency each base appeared. Although GT and AG are of upmost importante, the surrounding bases improve on snRNP recognition of the splice site.

While the splicing reaction *per se* does not require ATP hydrolysis, the rearrangement of RNA-RNA bonds between the pre-mRNA and the spliceosome complex needs energy. Additional ATP is necessary for protein rearrangement within the spliceosome complex in order to create the catalytic sites where the reaction will take place and to return the snRNPs to its initial conformation (Webb *et al.*, 2014).

While the described spliceosome complex is the major contributor for the splicing mechanism, a minor form is also present, the U12 spliceosome, responsible for less than one percent of all splice sites. It is composed of U11, U12, $U4_{atac}$, $U5_{atac}$ and $U6_{atac}$. Although the reaction and mechanism is similar, the consensus sequence is different with the 5'SS characterized by AT (or AU in mRNA) and the 3'SS as AC. Using comparative genomics, U12 spliceosome is highly conserved between species and allows little variation (Will *et al.*, 2001; Webb *et al.*, 2014).

It is worth mentioning that there are self-splicing introns, mainly in mitochondrial and chloroplast genes that do not involve the spliceosome complex. In these particular cases the pre-mRNA molecule folds into a specific 3D conformation and the 2 transesterification reactions occur spontaneously (Kruger *et al.*, 1982; Levine, 2001).

A problem in systematic identification of exons is the presence decoy sequences within introns. Athough the consensus scores are similar to authentic splice sites, they don't create splicing patterns detectable (Wang and Burge, 2008).

Similar to how transcription is affected by enhancers and silencers, splicing is also regulated by other elements, the splicing enhancers and silencers. They are 6 – 8 nucleotides long and are present in the pre-mRNA strand, usually within 300 nucleotides upstream or downstream of intron-exon boundary (Barash *et al.*, 2010; Webb *et al.*, 2014). Thus, these cis-elements can be divided into Exonic Splicing Enhancer (ESE), Intronic Splicing Enhancer (ISE), Exonic Splice Silencer (ESS) and Intronic Splicing Silencer (ISS) accordingly to the overall effect they have in splicing events and place in the strand. They are recognized by RNA Binding Proteins (RBP) and it depends on consensus-based interaction. These can be divided in two big groups, those that diminish splicing activity, mostly heteronuclear Ribonucleoproteins (hnRNP), and those who increase it, mostly SR proteins. SR proteins are named as such for their high content in serine and arginine. They increase splicing by recruiting spliceosome components through protein-protein interaction. On the contrary, hnRNP prevents through various processes. Some hnRNPs block U1 and U2

interaction, others change pre-mRNA strand conformation producing loops or even by directly displacing snRNPs assembly (Wang and Burge, 2008).

Based on different techniques, CRE as well as trans elements who influence splicing have been identified and their sequence and recognized consensus listed as part of available *in silico* tools (Desmet *et al.*, 2009; Jian, Boerwinkle and Liu, 2014; Paz *et al.*, 2014; Mao *et al.*, 2016; Tang, Prosser and Love, 2016).

1.3.2.2 Alternative Splicing

Alternative splicing (AS) is a process where, by combining different exons, multiple isoforms of the same gene can be produced. This mechanism is widespread among eukaryotes and AS variation is better correlated with organism complexity then genome size (Webb *et al.*, 2014).

In humans, more than 90% of all protein coding genes present alternative splicing patterns and contributes greatly to protein variety, with each gene presenting 5-10 isoforms. This results in increased transcriptome diversity producing proteins with different stability, translational efficiency, location and, sometimes, antagonist functions (Pajares *et al.*, 2007; Scotti and Swanson, 2016; Park *et al.*, 2018). 86% of all genes have a minor isoforms whose frequency is higher than 15% and AS complexity is correlated with phenotype complexity (Wang *et al.*, 2008). Considering that on average each gene has 8 introns and 9 exons, there are many possible combinations from each gene. Moreover, when we consider the size difference of the typical exon and intron, ~150 nts vs ~3k nts, there is ample opportunity for AS patterns within a gene with multiple alternative exons possible per intron (Webb *et al.*, 2014).

An example of antagonist isoforms is BCL-x, a gene that encodes a mitochondrial transmembrane protein that can be spliced into BCL-xL, a protein which is anti-apoptotic and promotes cell survival, or BCL-xS, which is pro-apoptotic. Changes in splicing of this gene have been associated with cancer (Stevens and Oltean, 2019).

But AS does not always produce new isoforms. These are designated non-productive AS patterns, produced by protein coding genes which harbour AS patterns that are not be translated or even be quickly degraded. Some forms of non-productive AS are evolutionary conserved across species suggesting some functional regulatory role as a post-transcriptional form of gene expression regulation. Others are promptly degraded via nonsense mediated decay (NMD), a mechanism activated by the presence of a stop codon before the last EJC. Such an example of

protein production via AS is present in ribosomal protein-coding genes. Due to the effect of environmental factors on splicing machinery, ribosomal protein-coding genes splicing pattern changes producing either the known coding transcript or a non-productive alternative (Webb *et al.*, 2014).

Given there is some heterogeneity within intron/exon borders – as can be seen in the consensus sequence in Figure 1.1 – the base pairing between snRNAs and pre-mRNA sequence is not always perfect. These differences result in a spectrum of splice site affinity. This translates into the presence of constitutive exons, those who are always present in the final mRNA which have high affinity with spliceosome, and alternatively spliced exons, whose presence in the final strand depends on other factors.

The choice of the final splicing pattern is dependent on molecular factors that interact with splicing CRE, to promote or prevent spliceosome assembly onto the strand. Given protein expression is cell-type and development stage specific, by modifying cellular protein content or its activity, different transcripts of the same gene can be produced. While there are ubiquitous RBPs, expressed universally on all protein producing cells, there are some who are characteristic of a cellular lineage with over 22k tissue-specific AS events (Wang *et al.*, 2008). Examples of cell type specific AS regulators are ESRP, CELF, MBNL, RBFOX and PTB protein family which control AS patterns in epithelial, muscle and neuronal cells (Park *et al.*, 2018). Furthermore, splicing is most often controlled by multiple splicing signals. While most of these factors are the previously mentioned RBPs, there is contribution from other factors. One variable that can alter pre-mRNA splicing is the chromatin state. Nucleosomes tend to be about the length of an exon, and it may act as a speed bump on transcription. This allows the formation of complex between the pre-mRNA and splicing proteins. When there are changes in chromatin structure a different splicing pattern may be chosen generating a different isoform. This can be explained by the rate of RNA formation as in less condense states favours high rate transcription and exon skipping. On the other hand, it is known that RNAPII has a pausing pattern when transcribing DNA to pre-mRNA due to a more condensed chromatin state. This allows SR proteins to attach themselves onto freshly transcribed sequences and recruit U2 and U1 leading to more splicing events (Ferreira *et al.*, 2007; Alberts *et al.*, 2017). Similarly, histone modifications can also induce changes in AS by interacting locally with RBP. Due to the close association of transcription complex and these components of the spliceosome histone alterations will affect splicing of specific genes.

The change of a single factor has the possibility of offsetting a splice pattern, especially relevant in weaker splice sites (Webb *et al.*, 2014). As each individual splicing decision within a gene extends beyond binary decision it is the sum of all contributions that determines whether splicing occurs but also under which pattern. Each of these decisions are called an event and each isoform results from a difference in at least one AS event. Although a high degree of variability is possible, the spliceosome assembly is very precise, identifying true splice sites over cryptic sites, which are very similar in sequence. This ensures high fidelity removal of intronic sequences while allowing for AS patterns to arise (Webb *et al.*, 2014).

The most common type of AS events in animals is skipped exon (SE), also termed cassette exon, corresponding to 42% of all AS events. It occurs due to failure to recognition of 3SS, an exon, or group of exons, is skipped and discarded. There is also intron retention (IR), where the otherwise removed sequence remains in the mRNA. It is the rarest event only accounting for 9% of all events and occurs when spliceosome fails to assemble. Mutually exclusive exon (MXE) where the splicing pattern restricts the presence of one of the exons. To further expand AS events, the specific splice site can also shift resulting in an alternative 5'SS (A5SS) or alternative 3' splice site (A3SS), when there are multiple 5' exon-intron or 3' intron-exon boundaries accounting for 23% and 26% of AS events respectively. It occurs when snRNP attach and slice pre-mRNA molecules at an alternative, cryptic site.

Furthermore, in combination with other gene expression regulatory mechanisms, there can also occur alternative first exon (AFE), whereby choosing a different TSS a different set of exons is present on the mRNA and alternative last exon (ALE) where by using a different poly-adenylation site there is variation on the size of the last exon. While some authors consider AFE and ALE as part of AS, others regard them as multi-mechanism products and outside of the scope of AS (Ferreira *et al.*, 2007; Li *et al.*, 2017; Saraiva-Agostinho and Barbosa-Morais, 2019).

1.3.2.3 Genomic Variants Impact in Alternative Splicing

Splicing machinery has a certain degree of flexibility allowing for a genomic variant on a gene to favour an alternative splicing pattern. This is due to the splicing complex choosing the optimal – strongest – available solution for splicing. Presence of SNV in splicing CRE provides some AS variability between individuals. Such CREs can be any of the previously discussed 3'SS, 5'SS, BPS or any auxiliary regulatory features such as exonic or intronic splicing enhancers or

silencers. The presence of variants on 5'SS or 3'SS has an expected bigger impact on splicing pattern then on any other CRE affecting all tissues as they are essential to the splicing mechanism. Nonetheless, variants on splicing enhancer or silencers can also have organism wide effect if the interacting trans-element ubiquitous or a narrower cell specific impact due to the tissue/time specific nature of some RBPs expression. It is important to point out that SNV not only disrupt CRE but can also create new ones leading to the usage of an alternative cryptic splice site (Park *et al.*, 2018). A recent analysis reclassified 22% of disease associated alleles adding splice altering as a novel consequence (Manning and Cooper, 2017).

1.3.2.4 Measuring Alternative Splicing

To assess differences in splicing we first need to be able to measure it reliably across samples and experimental data. Different metrics are available to measure splicing based on available data. One can assess splicing by 1) measuring the proportion of isoform production from each gene; 2) quantifying the usage of each exon and perform exon-wise eQTL; or 3) by determining usage of exon-exon junction boundaries. Each method has its own advantages and disadvantages: while whole isoform proportion seems to be the best method it depends on the availability of long reads, which is still not common, or pivots on *de novo* transcriptome assembly from short-read sequencing, whose performance is lacklustre on repetitive sequences and highly expressed genes; exon-wise eQTL appears to be a good alternative solution as it easily performed on short length reads but suffer from gene expression bias – which can be accounted for – and is unable to detect alterative splice site usage; lastly exon boundaries usage is dependent on good genome annotation and alignment tools to identify the excised introns. While proportion of isoforms is on itself a ratio, quantity of a specific isoform divided by the total number of isoforms from that gene, the exon-wise and exon-boundary quantification are not and are subject to bias from gene expression. To overcome this problem Percentage of Splice In (PSI) is used to be able to compare splicing measurements. It is given by the ratio of number of reads that include that exon/splice site over all the reads that span that segment. As such it is scaled from 0, the minimum where no event was detected, to 1, all the detected reads detected support that event. This way the gene expression bias is bypassed (Park *et al.*, 2018).

1.3.2.5 sQTL and ASAS

sQTLs and ASAS are specific cases of QTL and allele-specific studies where the phenotype to be mapped is splicing. It allows the assessment whether a variant is associated with changes in ratios of alternative splicing. Given how many single-gene and complex diseases are linked to changes in splicing, several tools are available. sQTL is usually a two-step method where first the splicing phenotype is quantified and later it is mapped against the sample's genome. Examples of such tools are psichomics, leafcutter, rMATS, cufflinks2, MAJIK which all measure splicing but using different definitions for alternative splicing event (Zhao *et al.*, 2013; Shen *et al.*, 2014; Y. I. Li *et al.*, 2018; Saraiva-Agostinho and Barbosa-Morais, 2019). Alternatively, one can choose to use the relative abundance of isoforms as measurement using sQTLseekeR or ASARP (Monlong *et al.*, 2014). To perform QTL mapping FastQTL is available or its implementation using GPUs, TensorQTL (Ongen *et al.*, 2016; Delaneau *et al.*, 2017; Taylor-Weiner *et al.*, 2019).

ASAS usage is not as widespread as sQTL, leading to fewer tools available for this method. ASARP and Pairadise are examples but examples of their application in literature is limited (Li *et al.*, 2012; Demirdjian and Xing, 2020).

1.3.3 Other mechanisms

Other post-processing mechanisms can also alter gene expression influencing phenotypic traits. The presence of distinct 5' and 3' untranslated regions (UTR) on a mature mRNA from an alternative transcription start site, splicing pattern or 3' cleavage result in different stability, translation efficiency and intracellular location. These processes are regulated by CRE present on the UTRs that can be modified by genomic variants. CRE in UTRs, in particular 3' UTRs, regulate miRNA recognition and binding modulating miRNA mediated decay. Another regulating mechanism of regulation is RNA structure of the RNA molecule. Variants that modify RNA structure are called riboSNitches and can prevent trans-molecules interaction with CRE creating pin-like structures or allow them by disrupting structural features. Furthermore, translation of mRNA into protein can also be affected by modifying the Kozak sequence or any other element necessary for ribosomal assembly or translation elongation. Subcellular localization is also susceptible to variants as it depends on RNA zipcode motifs to be transported to proper site (Manning and Cooper, 2017).

1.4 Splicing and disease

Mutations on CRE splicing mechanism are responsible for a wide variety of disease, accounting for 10 to 60% of genetic disease (Webb *et al.*, 2014). Furthermore, mutations on trans splicing elements are established as the cause of several disease by compromising splicing on a large-scale (Park *et al.*, 2018). An example of such mutation impacting splicing CRE is spinal muscular atrophy (SMA) caused by mutations on Survival of Motor Neuron 1 (SMN1) and as consequence loss of motor neurons(Li *et al.*, 2019).

Common variants have also been implicated in disease through changes in splicing. Changes in splicing promoted by variants on oxidizer low-density lipoprotein receptor 1 (*ORL1*) and low-density lipoprotein receptor (*LDLR*) have been identified increasing risk for atherosclerosis and coronary disease. These loci have been previously associated with risk for these diseases by GWAS providing a molecular basis for increased risk (Gretarsdottir *et al.*, 2015; Tejedor *et al.*, 2015).

1.4.1 Breast Cancer and alternative splicing

One of the diseases where changes in splicing have also been characterized is cancer, both as a result of mutations CRE an also in trans elements as RBP or spliceosome components (Calabrese *et al.*, 2020). In particular breast cancer, where changes in alternative splicing patterns of specific genes – *BRCA1*, *ESR2* and *HER2* as well as cell cycle progression and DNA response damage – have been identified either as a consequence of somatic or germline mutations (Venables, 2004; Venables *et al.*, 2008, 2009). Furthermore, changes on RBPs, transforming gene expression as a whole, have been detected in more than 50% of analysed samples contributing to loss of tissue morphology, enhanced cell proliferation and invasion (Dvinge *et al.*, 2016; Park *et al.*, 2019).

BRCA1 is a particularly well studied gene as germline mutations are the cause for hereditary breast and ovarian cancer. The initially annotated six isoforms have expanded to more than 100 splicing patterns that have been identified in both healthy and disease individuals (Li *et al.*, 2019). Full length *BRCA1* gene has 24 exons is responsible for activating DNA repair machinery, regulating cell cycle and gene transcription. BRCA-IRIS, a natural occurring isoform derived from alternative splicing of pre-mRNA, is primarily found on foetal skeletal muscle, adult leukocytes, spleen and highly expressed in some cancer types. Alternative splice sites on exons 1

and 11 results in a shorter protein that lacks RING domain on the N-terminal and BRCT on carboxyl-terminal disabling its interaction with other proteins. This BRCA-IRIS mRNA is more stable than BRCA1 full length transcript leading to an accumulation. The presence of this isoform is a marker of more aggressive forms of cancer, in particular for triple negative breast cancer where it pomotes the initial formation, invastion and migration(Clark *et al.*, 2012; Li *et al.*, 2019).

Another less drastic example is a common variant on *ESR1*, the gene that codes for alpha oestrogen receptor (ERα). The presence of the alternative allele rs3020314-C was associated with an increase of estrogen receptor positive breast cancer risk of 1.05 compared to reference homozygotic individuals. This allele was associated with a slight increase on a delta-5 isoform, which lacks hormone binding domain and constitutively activates gene expression (Chaidarun and Alexander, 1998; Dunning *et al.*, 2009; Maney, 2017).

1.5 Breast Cancer

Breast cancer is a heterogenous disease characterized as an abnormal growth of breast tissue. It can be derived from different cell types, but most often it is from ductal glands that proliferate uncontrollably.

1.5.1 Worldwide incidence

Breast cancer is the most common malignancy and the leading cause of death by cancer in women in the world, affecting 8–10% of all females. 2.1 million women were diagnosed with breast cancer in 2018, an increasing trend, accounting for one in every four cancers detected in women. In the same year, 626 thousand women died of breast cancer. Although there is a higher incidence in developed countries, BC is deadlier in developing nations reflecting the inequality of availability of screening, treatment and diagnostic tools (Bray, F; Ferlay, J; Soerjomataram, I; Siegel, RL; Torre, LA; Jemal, 2018; Bellanger *et al.*, 2020).

1.5.2 Clinical features and classification

Since early this millennia, the molecular classification for BC was adopted as it is best suited to group patients according to disease characteristics. It divides breast cancer into four different types: Luminal A and B, Basal like, and Human epidermal growth factor receptor 2 (HER2) enriched. In clinical practice this translates into the evaluation of molecular characteristics by profiling estrogen receptor (ER), progesterone receptor (PR) and HER2 expression in the tumor, with the caveat that it is not a completely direct relationship. The overexpression of ER and/or PR allows the classification as Luminal type A. Lower abundance of hormone receptors combined with mutations on other tumor drivers – as *PIK3CA* – are classified as Luminal type B. Luminal type cancers represent 70% of all BC cases. HER2 overexpressing tumors are classified as HER2 enriched, accounting for 20%. Basal type includes the triple negative breast cancers (TNBC) as neither of the previous referred receptors are overexpressed. It is associated with a worse prognosis as it derives from higher genomic instability associated with higher rates of *BRCA* and *TP53* mutations. It is interesting to notice that subtype incidence varies according to ancestry, with women of African ancestry displaying a higher rate of TNBC (Harbeck *et al.*, 2019). More recent classification based on expression of 50 genes (PAM50) is available with limited clinical application (Sørlie *et al.*, 2001; Cheang *et al.*, 2015).

1.5.3 Treatment and Prognosis

The best predictor for treatment success is not only cancer subtype but also stage. BC is curable in 70 - 80% of all early-stage, non-metastatic disease. On the other side of the spectrum, late cancer stages with distant metastases is practically incurable, as there are no therapies available with curative intent. In these cases, the best possible outcome is to achieve the maximum progress free survival while providing best quality of life possible.

Surgery is most frequent intervention, particularly on initial stages. Recently, refinement of surgical techniques allowed to move from mastectomy, with the complete removal of the breast to more exact extraction of lumpectomy where only the tumoral tissue and a safety margin are removed. Auxiliary lymph nodes are usually removed in order to reduce metastasis occurrence (Provencio *et al.*, 2018).

Radiotherapy is another option that can be use in locoregional disease both as neoadjuvant and/or as adjuvant in the case of lymph node involvement or as adjuvant for lumpectomy in order to eliminate residual cells. It benefits from abscopal effect where destruction of tumoral cells in the breast releases antigens stimulating immune cells to engage distant cells allowing the reduction of metastasis (Borrego-Soto, Ortiz-López and Rojas-Martínez, 2015).

Systemic therapy is dependent on clinical subtype. It can be used as neoadjuvant to test tumour response and reduce its size. Endocrine inhibitors are available that target ER^+/PR^+ with high degree of success, whose most representative drug is tamoxifen (Pompei and Fernandes, 2020). Anti-HER2 antibodies, as transtuzumab, are used to bind and inhibit HER2 activation. These also tag HER2 overexpressing cells for destruction by immune system or carry targeted drugs (Mignot *et al.*, 2017). Traditional chemotherapy, in particular taxanes, anthracyclines, antimetabolites, alkylating agents and platinum-based drugs are used when no specific target is available or in addition to targeted therapy in high metastatic risk patients (Moo *et al.*, 2018; Kalimutho *et al.*, 2019). Polymerase inhibitors target cells displaying genomic instability, characteristic of BRCA syndromic cancers(McCann and Hurvitz, 2018). Immunotherapy, a current trend in cancer therapy, is under evaluation on breast cancer, with particular interest on immune checkpoint blockers in TNBC (Vonderheide, Domchek and Clark, 2017; Schmid *et al.*, 2018; Adams *et al.*, 2019).

In general, TNBC is regarded as the most mortal subtype due to lack of actionable targets older anti-cancer drugs are still used (10-13 months of median overall survival for metastatic disease). The best prognosis is for luminal A and B, cancers that respond well to therapy (4 to 5 years median overall survival for metastatic disease).

1.5.4 Risk for breast Cancer

Breast cancer risk can be divided into two different components, environmental and genetic factors. Although this distinction appears to provide a clear division, some environmental/genetic elements only exert its effect on the presence of other enabling genetic factors (Nickels *et al.*, 2013).

These factors exert their effect increasing genomic instability, one of the hallmarks of cancer (Hanahan and Weinberg, 2011). Different mechanisms are proposed, from DNA damage

repair pathways failure, hyper-replicative induced stress, defective mitosis and transcription-associated stress (Lee *et al.*, 2016). These processes result in a stepwise acquisition of somatic mutations, leading to a heterogeneous mass of cells, each with a set of common and private mutations. This mass is under selective pressure where the best adapted cells thrive constituting the bulk of the tumoral mass. This heterogeneity also promotes resistance to therapeutical interventions as previous neutral or negative mutations may provide endurance to changing conditions (Stanta and Bonin, 2018).

1.5.4.1 Environmental factors

Environmental factors are responsible for up to 90% of all breast cancer cases. The effect of these environmental factors is documented as families migrating from low to high BC incidence countries found that breast cancer risk rose over generations, mimicking incidence among resident population (Ziegler *et al.*, 1993). Follow up studies showed that this increase in BC risk was due to the adoption of several risk behaviours (Key *et al.*, 2003, 2011)

One of the biggest hazards for breast cancer is hormonal exposure either from endogenous sources, i.e. ovaries and to a lower extent adipose tissue, or from external sources, such as birth control pills or hormone replacement therapy in menopausal women. Having a large window of time between menarche and menopause combined with nulliparity and exogenous hormone control is associated with increased risk, as they augment life-time exposure to oestrogen. Other known multi-cancer modifiable risk factors are alcohol, tabaco, diet, exercise and high body mass index, which play a role in disease aetiology (Strumylaite, Mechonošina and Tamašauskas, 2010; Parkin, Boyd and Walker, 2011).

Ionizing radiation, as X-ray and gamma radiation from medical imaging or background, promotes DNA damage both directly as double strand DNA breaks and indirectly by increasing reactive oxygen species (ROS)(Borrego-Soto, Ortiz-López and Rojas-Martínez, 2015). This results in contradicting screening policies as mammograms are a source for radiation, with most health organizations recommending regular screening at later in life (Bellanger *et al.*, 2020). Furthermore, in the presence of some syndromic diseases, as Li-Fraumeni syndrome and Human Breast Ovarian Cancer disease, radiological techniques are not advised as these individuals present higher sensitivity to radiation (Borrego-Soto, Ortiz-López and Rojas-Martínez, 2015).

Environmental pollutants, as heavy metals and polycyclic organic compounds, are also regarded as increasing risk for breast cancer. Cadmium forms a high affinity complex with hormone binding domain of ER. Moreover, pesticides as dichlorodiphenyltrichloroethane (DDT) and poly-chlorinated biphenyls (PCBs) bioaccumulate and persist in human tissues functioning as endocrine disruptors, mimicking or interfering with hormonal functions (Strumylaite, Mechonošina and Tamašauskas, 2010; White *et al.*, 2016; Morgan *et al.*, 2017; National Institutes of Health; and U.S. Department of Health and Human Services, 2018)

1.5.4.2 Genetic factors

While most cases of breast cancer seem to be sporadic, there are some well characterized families where BC has an increased incidence compared to the population. On women with affected relatives the risk of BC increases to 15% (Ripperger *et al.*, 2009), and the risk of women with a first-degree relative with BC has double the risk of the general population.

Currently, more than 50% of familial predisposition to BC is unexplained. The risk of BC is not determined by a single variant but by the collaboration of many variants along with the contribution from the environment. So far, no risk prediction model can integrate all available genetic variants and, as such, individual estimation of risk is yet difficult to assess.

Variants with greater risk and few carriers were rapidly identified through familiar clustering. Examples are those of the mutations in the genes *BRCA1* and *BRCA2*. It is estimated that germline *BRCA1* and *BRCA2* mutations alone are responsible for 20-40% of familial cases, amounting to less than 5% of overall BC cases. Women with these mutations have a lifetime risk of BC up to 85%, besides increased risk for ovarian cancer. Rare syndromes characterised by mutations in the gene *TP53*, *STK11*, *PTEN*, *CDH1*, *NF1* and *NBN* also have a sharp increase in risk for several types of cancer amongst them BC (Ripperger *et al.*, 2009; Ban and Godellas, 2014).

Candidate gene approaches identified rare mutations in genes involved on DNA repair that moderately increased risk for BC – 2 to 4.3 times higher than the general population. Such genes as *ATM, CHEK2, BRIP1, PALB2* and *RAD50*, when mutated confer a higher baseline mutation rate and, consequently, higher accumulation of mutations throughout life. These mutations are in higher frequency than familial syndromes but still below the 1% threshold to be considered SNPs (Ripperger *et al.*, 2009). Together, the genes identified by this approach account for 2.3% of the overall familial risk.

More recently, genome-wide association studies went further and identified low-penetrance and high frequency BC susceptibility variants. Some risk loci are within genes and the biological target is thus easily established – *FGFR2, TOX3, LSP1, MAP3K1, TGFB1*. However, other loci have no known gene in the vicinity and fall in large LD blocks, with many possible causal variants– as is the case of the loci 2q25 and 8q24 (Ripperger *et al.*, 2009; Kachuri *et al.*, 2020).

Functional characterization of GWAS hit-SNPs is an ongoing effort with most researchers relying on studies focusing on the expression changes to gather insight into causal variants and target genes in each loci. Contribution of each individual variant to tumorigenesis is still not well characterized as most of identified SNPs have no known biological meaning, but their full understanding is crucial to the advancement of prevention and treatment of BC.

2 Aim

GWAS have identified hundreds of loci that modify breast cancer susceptibility, with most locating to non-coding regions of the genome. The specific mechanisms by which they operate are not well understood but can help identify the causal variants of risk and lead to the development of new preventive measures.

Many studies have looked into the functional impact of variants located in cis-regulatory elements (CREs), mostly evaluating the modification of transcription factor binding as a possible mechanism, overlooking other possible mechanisms.

The main goal of this thesis is to investigate another plausible mechanism, specifically, whether cis-regulatory variants may contribute to breast cancer risk by modulating alternative splicing. To address this goal, the specific aims are:

1) Quantify alternative splicing patterns across samples in healthy female breast tissue;
2) Identify common cis-regulatory variants that impact alternative splicing in breast tissue, by performing sQTL analysis;
3) Identify sQTLs that are potentially associated with risk to breast cancer, by carrying co-localization analysis between sQTL and GWAS risk-associated variants.

3 Materials and Methods

3.1 Data sources

RNA-seq and genotype data used in the production of the current project were retrieved from GTEx Portal on 27/02/2020 and database for Phenotype and Genotype (dbGaP) accession number phs000424.v7.p2 and phs000424.v8.p2 (Ardlie *et al.*, 2015).

RNA-seq data was produced with the Illumina TruSeq sequencing protocol, resulting in a non-strand specific, polyA+ selected library. Oligo dT microbeads were used for selection to exclude rRNA, which accounts for > 90% of the cellular RNA content. Next, heat fragmentation was chosen to avoid transposase selection bias, and cDNA synthesis was performed using random hexamers. The library was prepared by generating blunt-end fragments, adding of 3' A (A-tailing) tails, and adapter ligation (taking advantage of the hanging A using a T-based overhanging). Fragments were sequenced producing 76 bp length paired-end reads with a median coverage of 76M (Head *et al.*, 2014).

The choice for paired-end, high depth, non-strand specific protocol improves accuracy of lowly expressed genes and transcripts making this dataset ideal for our objective (Conesa *et al.*, 2016).

Data is available in sequence read archive (SRA) format, a standard that allows easier data sharing while allowing access restriction and compression by design (*SRA-Tools - NCBI*, 2015; *The Sequence Read Archive (SRA): Getting Started*, no date). The SRA Toolkit is a collection of tools that facilitates extraction and decryption of SRA files into raw-reads as a fastq file, or into reference genome aligned BAM or SAM, among other formats. Using SRA Toolkit, I performed a fastq-dump to retrieve fastq files, which carry both sequence and quality score information. Given that the original data is paired-end I obtained two files per run with each line from the first file being paired with a line from the second file. As our aim is to study breast cancer, a mainly female disease, data was only selected from female breast samples.

3.2 Informatics languages and tools

In order to handle the enormous datasets and perform its analysis using the required tools for this work, three different programming languages were necessary. This took a considerable amount of time to achieve the required skill level to take on this thesis.

The most extensively used was R, a free programming language developed for statistical and graphics computing (R Core Team, 2020). Besides basic functions, publicly available packages can be imported that expand the capacity of this language, many of them at no cost.

To develop and run the scripts in R (version 3.6.2 from 2019-12-12) I used R Studio®, an Integrated Development Environment (IDE). All the scripts/tools were ran on a machine running Ubuntu (release 18.04.3), in a 64 bit Linux/GNU platform.

Python (version 3.6.9) and bash were also used to a minor extension. Python is also a free programming language maintained by Python Software Foundation (Python Software Foundation, 2020). It is a general-purpose language whose feature set is expanded by importing used published libraries. Bash is a command interpreter for GNU (GNU is Not Unix) systems that receives commands from the user via prompt and performs actions based on locally stored function libraries. Python and Bash scripts and code were ran on RStudio® command line terminal.

It is also important to mention GitHub and CRAN (The Comprehensive R Archive Network), online repositories that allows easier software sharing and usage. GitHub is a Git repository hosting service with a web-based graphical user interface. It can store code, register any changes and allows users to interact with repository creator by collaborating on a project, raising issues or even branching off previous projects. CRAN

3.3 RNA-seq analysis

3.3.1 Quality control, Pre-processing and Alignment.

FastQC v0.11.9 was used to perform quality control per file (*Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*, no date). This software parses through each fastq file originating an individual quality report with summary graphs and tables that facilitate interpretation (Dow, 2003). It has 12 quality indexes, which can be selectively applied:

1) Basic statistics – Describes the file in terms of name, file type and encoding used, total number of reads, sequence length and % of guanine and cytosine (GC) content. Total number of

reads allows a quick assessment of whether read depth is enough for the purpose at mind, while GC content is dependent on the species of origin and method used for RNA extraction.

2) Per base sequence quality – A "box and whisker" representation of the per base position quality score distribution. The vertical axis shows the Phred score index, a measurement of how likely is it that a base call is wrong. Considering P as probability of being wrong and Q the quality score, $Q = -10 \log_{10} P$. Q ranges from 1 to 40, where 1 is the lowest quality with probability of wrong call equal to 1, and 40 where chance of being incorrect is 1 x 10^{-4}. The authors provide a subjective evaluation as a coloured graphic background, where green represents very good quality (Phred score > 28), reasonable quality is represented in yellow (28 > Phred score > 20) and red denotes low quality (20 > Phred score). The extremes of the sequence reads are usually of lower quality then the remaining read, due to known issues, such presence of impurities in the sequencing run, that affect the first few bases and decrease signal to noise, and template damage over the sequencing run, which affects the latter positions. Importantly, there is no fixed threshold for quality, with the decision of threshold being left to the researchers' decision. Corrective measures available are trimming of low-quality base readings and filtering out low quality reads.

3) Per sequence quality score – Histogram of average Phred score per read. This graph shows whether a portion of reads have a lower average skewing per base sequencing quality. The presence of two or more peaks implies that a part of the flow cell may be faulty, in which case per lane/tile quality parameter should retrieve more information. Problems occur more frequently at the border of the cell. The appropriate correction is to introduce a per read quality threshold filtering low quality reads.

4) Per base sequence content – Percentage of each base called at each position. In a random library it is expected parallel lines without any peaks for nucleotides. A warning or failure is issued when the difference between complementary bases is higher than 10% or 20% at any position. Pre-sequencing amplification using random hexamers can induce some bias at the initial positions which should not affect downstream analysis. The presence of adapters or other tags used during sequencing should be visible here and be trimmed downstream.

5) Per sequence GC content – Percentage of G + C from overall reads, which should mimic the underlying data. Displayed as a histogram to improve visualization of multiple peaks, indicative of sample contamination. While this is specimen and genomic region specific, departures from expected values should be analysed. A theoretical normal distribution is computed from

inputted data and overlayed on histogram. FastQC gives a warning at 15% deviation from normal distribution and failure at 30%. The most common reason is a problem with library preparation, where sharp peaks are attributed to contamination with adapter dimers, while broader peaks are due to contamination with a different species. Corrective procedure involves redoing RNA-seq procedure.

6) Per base N content – Evaluation of introduction of N across the length of the read, which is done when the sequencing platform is unable to assign a base to a position. An increase in N content is frequent in the latter bases, especially in longer reads. Corrective measure is trimming when a user defined threshold is passed, with FastQC showing a warning or failure when N content crosses 5% or 20% at any position.

7) Sequence length distribution – Histogram of the length of each read. While most sequencers produce evenly sized sequences, if a quality trimming is pre-applied it will be shown here. In the presence of different sized reads a warning is raised and if there are any zero-length sequences a failure is applied.

8) Sequence duplication score – Indication of whether a particular sequence is over-represented in the input file. Given how a typical RNA library is prepared, it is expected that most sequences will occur once. The presence of duplication peaks may be due to an enrichment bias, scarce biological diversity on the analysed sample or of biological significance. Most frequently, technical bias are introduced during PCR pre-amplification, but biological duplicates are also possible in small genomes where increasing sequencing depth is not an advantage, for example. In order to manage resources, FastQC performs analysis of the first 10k reads and only up to the 50th base. The plot shows the proportion of the library made of duplicated sequences and potential loss in percentage of reads if only unique sequences where allowed. Warning and failure thresholds are set at 20% and 50% of total read counts.

9) Overrepresented sequences – Results from the previous analysis where the duplicated sequences are shown. It further explores the motive for why a certain sequence is so frequent. A threshold of 0.001 is applied for a warning listing the repeated sequence. A library of known contaminants is checked, such as adapters, primers, among others frequently used in RNA-seq experiments. It identifies similar sequences based on matching 20 bps while allowing for 1 mismatch.

10) Adapter content – Evaluation of the presence of adapters based on defined kmers, sequences of length k, usually 5 to 7 bp. It checks for the presence of adapter dimers. In the presence of duplicated sequences, this module will be triggered showing most frequent k-mers. The plot shows the cumulative proportion of the adapter at each position. Trimming is advised when a 5% threshold is passed, and failure is set to 10%.

11) Kmer content – Overrepresented sequence analysis is hindered by low sequence quality assigning wrong base identification and partial sequenced fragments. It is expected that small fragments are evenly distributed across all positions on the sequence. By analysing 7-mer content this module detects when a distribution is not random. It employs a binomial test reporting any kmer with positional biased content, showing the top 6 kmers and their distribution. A warning is given is p-value of the binomial test is > 0.01 and failure when > 10^{-5}.

12) Per tile/lane sequence quality – This option is only available when using certain sequencing platforms that retain flow cell/lane information. It checks if there are differences in quality depending on where the read is generated. It is able to identify problematic areas that should be discarded afterwards. As the data I had access to did not have flowcell/lane information this module was not used.

An independent quality report was generated for each file as FastQC does not support paired-end data input, hence we obtained two files per sample.

MultiQC v1.9 is a tool that aggregates previously generated multiple quality indicators outputs into an easy to interpret single file. It is particular useful when dealing with different QC tools from large sample sets (Ewels *et al.*, 2016). Some granularity is typically lost when aggregating many results into a single output. MultiQC tries to overcome this problem by generating dynamic html reports that allow selection, filtering and isolate data display on graphs. It is important to state that when dealing with larger number of samples it reverts to static output for ease of visualization and browser load restrictions. MultiQC was used to aggregate all the reports generated from FastQC into a single file forcing a dynamic html output.

Based on quality assessment, pre-processing was performed using CutAdapt (Martin, 2011). The primary aim of this software is to search for known adapter sequences and remove them from the read, discarding the remaining sequence if it was smaller than a defined threshold. It also removes primer sequences, poly-A tails and trims under user defined length or sequences. Quality filtering of reads was also added, which tolerates incomplete correspondence, allowing

mismatches, insertions and deletions up to a user defined value. CutAdapt v2.10 was used to trim the 76th base from all reads.

Due to quality concerns some samples were removed from further analysis. After pre-processing and before alignment a new QC round was performed with FastQC and MultiQC using the previously described procedure.

3.4 RNA-seq alignment

Alignment of RNA-seq to reference genome was performed using the tool STAR (Spliced Transcripts Alignment to a Reference) (Dobin and Gingeras, 2015). STAR can align non-contiguous segments of small-sized, either single or paired end, to full length RNA transcripts. It does so while detecting canonical splice junctions, as well as discovering new splicing sites and chimeric transcripts. It provides a fast, precise and reliable alignment. STAR accepts user defined settings for maximum number of mismatches and indels, providing flexibility. The mapping process is divided into two steps. The first step searches for the best matches with largest portion of the sequence included in the reference genome. Each hit is called a seed and is stored. The unmatched portion of each sequence is mapped again and stored as a second seed. In the second step STAR clusters all seeds from each read by position and scores the best alignment based on the number of mismatches, distance between seeds, etc. In practice it first generates a genome index from reference genome and annotation provided. Next it maps each read of the fastq file and outputs them as a SAM (sequence alignment/map) or BAM (binary alignment/map) file, which can use a significant amount of RAM, which in our case was up to 42Gb. Alignment quality index includes uniquely mapped reads (total number and percentage), average mapped length in nt, number of splices detected (total and annotated), type of splice sites (canonical and non-canonical), mismatch rate, deletion and insertion rate, multi-mapped reads and why, unmapped reads and chimeric reads.

Alignment was performed using v2.7.5a against the reference genome and annotation retrieved from ensemble (Homo_sapiens.GRCh38.dna.primary_assembly.fa for the genome and Homo_sapiens.GRCh38.100.gtf.gz for annotation). For genome index generation I specified an overhang (--sjdbOverhang) of 100nt, as there was no performance loss compared to the recommended 74nt. For alignment, general operation options were used (files location, input and output format) in addition to the following specific flags:

--twopassMode Basic – retrieval of novel splicing junctions in the first pass and inclusion on the second to provide higher sensitivity to this mechanism.

--outSAMstrandField intronMotif – reduces number of mapping errors by requiring that each splice site is either previously annotated or generated from well characterized intron motifs (GT/AG, GC/AG or AT/AC).

The quality control output generated from the alignment was aggregated and visualized using MultiQC v1.9, retrieving general alignment scores from each run.

3.5 Quantifying Alternative Splicing

To obtain independent alternative splicing (AS) measurements, usually given by various ways to determine a PSI (percentage of spliced in), I employed two different softwares: LeafCutter (Y. I. Li *et al.*, 2018) and psichomics (Saraiva-Agostinho and Barbosa-Morais, 2019).

LeafCutter is a suit of methods that detects splicing events from RNA-seq data. To extract and count introns it leverages CIGAR (Compact Idiosyncratic Gapped Alignment Report) strings produced during alignment present on BAM/SAM files. These strings report how the alignment from the sequence was to the reference genome. It is composed by pairs of one number and a character indicating how many bases it refers to in the reference genome and type of alignment. To detect introns, this script looks for N, representative of a gap in the sequence when compared to the reference genome. To reduce misalignments, a minimum of 6 aligned nts (M) are required at the edges of each intron. This generated an exon-exon junctions file per sample with position and count of all splicing events detected in that sample. From the sample-wise exon-exon junctions file, it detects common splice sites and clusters them across all samples provided (**Figure 3.1**). Flags were applied to specify the minimum and maximum size of 50 bp to 500kb for each intron. As previous, a filtering step was performed requiring at least 50 reads per cluster to ensure reliable exon detection and clustering. Cluster wise PSI was performed using the sum of read counts from the cluster boundaries as the dividend. Introns present in less than 40% of all samples or with no variation were filtered out. A BED file was prepared using leafcutter's recommended procedure.

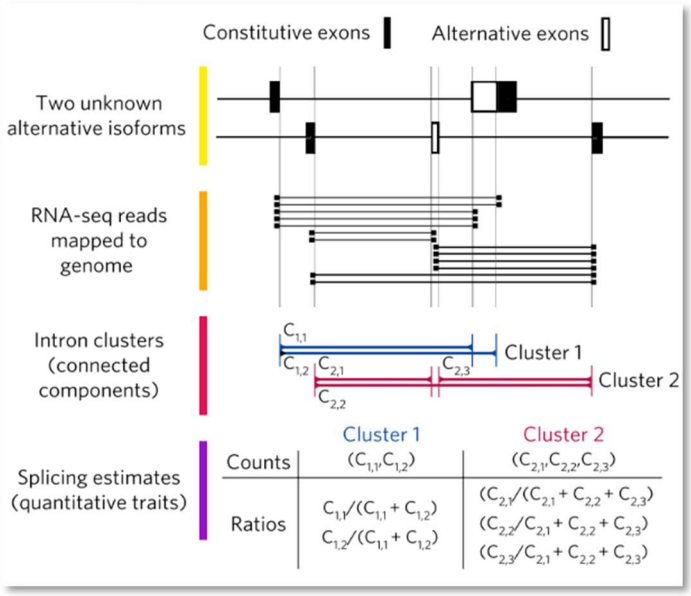


Figure 3.1 – Leafcutter clustering process. Leafcutter searches for splicing events from an intron perspective. Clustering is performed based on the exact splice site of each, if two different introns share a splice site they will be clustered together. Therefore, a cluster may span the whole gene if there are several different introns sharing at least a splice site(image from manuscript (Y. I. Li *et al.*, 2018)).

Psichomics is a modular R package capable of differential gene expression and splicing analysis, as well as survival analysis. It can receive input as SRA archive or pre-processed data. Here we used GTEx pre-processed data of exon-exon junction quantification (GTEx_Analysis_2017-06-05_v8_STARv2.5.3a_junctions.gct.gz, available at gtexportal.org/home/datasets). Exon-exon junction read counts were filtered by removing junctions with less than 10 reads in at least two samples, to ensure precise and unequivocal quantification removing unique events. AS event specific calculation of PSI was performed using psichomics provided annotation for hg38 (Agostinho, 2018). It does so by calculating the proportion of reads that support exon inclusion from known AS events: Skipped Exons (SE), mutually exclusive exons (MXE), alternative 5' splice site (A5SS), alternative first exon (AFE), alternative 3' splice site (A3SS) and alternative last exon (ALE) (Figure 3.2). Notice that there is no reference to intron retention as it is reclassified.

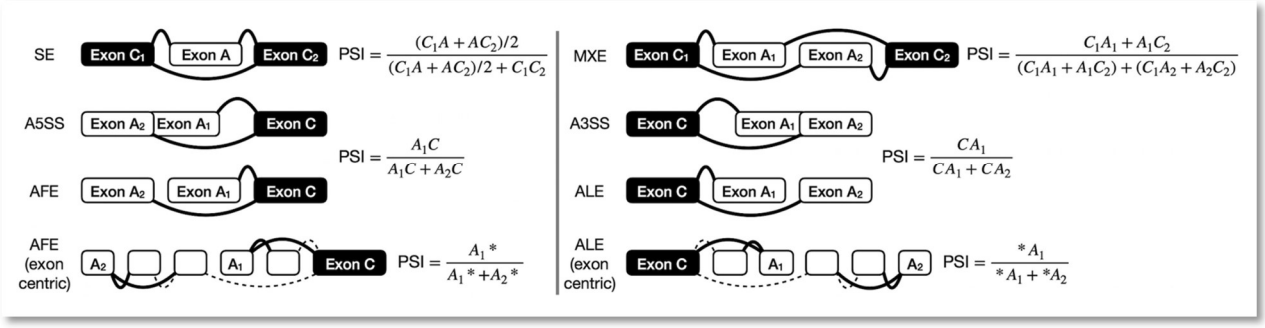


Figure 3.2 – Psichomics computed alternative splicing events. Psichomics is able to detect 6 different types of alternative splice patterns, Skipped Exons (SE), alternative 5' splice site (A5SS) , alternative first exon (AFE), mutually exclusive exons (MXE), alternative 3' splice site (A3SS) and alternative last exon (ALE). For each of the splice pattern, PSI is computed as the number of reads that support the alternative splice pattern divided by the total number of reads from all patterns that span those exons. Notice that AFE and ALE can be assess based on exon-exon junction detection or on exon read counts. (image from manuscript (Saraiva-Agostinho and Barbosa-Morais, 2019))

A bed file like structure was prepared using the first alternatively spliced site as the start position for each event.

The naming scheme used to identify events differs between tools. In psichomics it starts with the type of event annotated, the strand and chromosome in which it is coded in the genome, the constitutive 5' splice site, the alternative splice(s) site(s) – can be one in the case of A5SS, A3SS, AFE or ALE, two in SE or multiple in MXE – the constitutive 3' splice site and the gene it reports to. Leafcutter produces a smaller unique identifier, composed by chromosome, genomic coordinates of the exon borders and the cluster in which it was placed.

Exploiting *HES4* as an example, this gene is coded in chromosome one in the reverse strand between genomic coordinates 998,962 and 1,000,172 for GRCh38 (Genome Reference Consortium Human Reference 38). It has three different protein coding transcripts, HES4-201 with 4 exons, HES4-203 with three exons due to intron 1 retention (in striped white) between would be exon 1 and 2, and HES4-204 also with three exons as a consequence of skipped exon 2 (Figure 3.3A). As the last intron is constitutively spliced out, lets focus on alternatively spliced exons (E1, E2 and E3) and consider C1,2, C2,3 and C1,3 as the read counts of exon-exon junctions between exons 1 and 2, exons 2 and 3 and exons 1 and 3. Psichomics annotation presents two different alternative splicing events, regarding the skipped event of exon 2, named SE_1_-

_999866_999787_999692_999613_HES4, where $PSI = \frac{\frac{C1,2 + C2,3}{2}}{\frac{C1,2 + C2,3}{2} + C1,3}$ and the intron retention that is reclassified as an alternative 5' splice site, A5SS_1_-_999866_999692_999613_HES4, with $PSI = \frac{C2,3}{C1,3 + C2,3}$ (Figure 3.3B). As Leafcutter produces clusters based on shared splice site coordinates present in the RNA-seq data, it will identify three events clustering them together as they share at least one common splice site. Computing PSI for the first splicing event, it will be the read counts of exon-exon junctions for that specific event divided by the total read counts of the cluster or $PSI = \frac{C1,2}{C1,2 + C2,3 + C1,3}$ with the rest of the events on this cluster having the same denominator (Figure 3.3C).

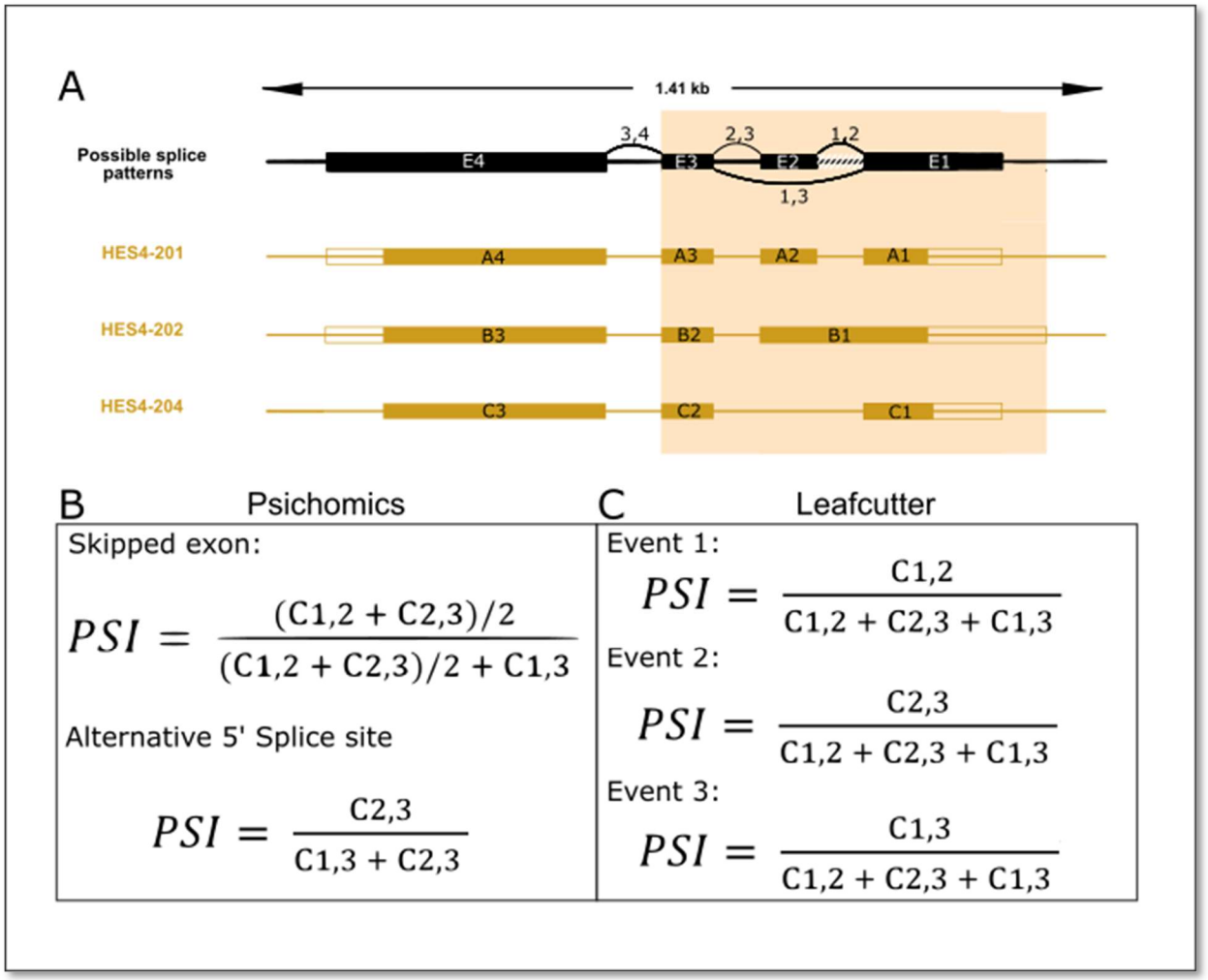


Figure 3.3 – *HES4* alternative splicing patterns and PSI. This figure shows how each software identifies alternative splice events and computes PSI. In panel A are represented possible splice patterns

from HES4 gene in black as well as coding the isoforms shown in gold. Each box is an exon and the line between each pair of boxes is an intron that is removed. UTRs are represented as the empty part of the box while coding sequence is the filled portion. In bisque background is alternatively spliced exons from gene HES4 for which PSI is going to be computed. In panel B are annotated events from psichomics and the exact formula to compute them and in panel C are the events discovered from available data. (image edited from ensembl.org)

3.6 Co-variates

In order to improve on QTL mapping accuracy and power, covariates need to be estimated and accounted for. Principal Component Analysis (PCA) is a dimension reduction technique that produces vectors of most variation across the data (Biswas, Storey and Akey, 2008; Ringnér, 2008). This estimates broad variation patterns across samples, usually produced by technical and/or biological artifacts (Fehrmann *et al.*, 2011). PCA was performed on the exon-exon junction read counts on both pre-processed and newly processed data independently. Correlation between principle components (PCs) and known biological and technical co-variates was assessed in an attempted to identify contributing components. Top five PCs were used as covariate, as well as known technical ones, such as pre-amplification prior to sequencing and sequencing platform.

3.7 sQTL mapping

QTL mapping was performed using TensorQTL, an improved version of fastQTL that uses GPUs (graphical processor units) to increase algorithm performance and decrease analysis cost (Ongen *et al.*, 2016; Taylor-Weiner *et al.*, 2019). FastQTL is a flexible and user-friendly QTL mapper. Both of these software search for associations between quantifiable molecular phenotypes, given in a BED (Browser Extensible Data) file and nearby genetic variants, provided as plink's BED (PLINK binary biallelic genotype table) or VCF (Variant Call Format) format for TensorQTL and FastQTL respectively. The BED file has a similar structure to UCSC standard – each row is a different observation, and the first four columns are Chromosome, Start, Stop, and Phenotype Identification with an extra column with a quantified phenotype per sample. The VCF and plink's BED files are used to store genomic variant information on samples. The VCF is composed by a header that provides information regarding file structure and how it was obtained and the body, a

table where each individual row is a different variant. There are 9 standardized columns, regarding chromosome, position, identification, which are the reference and alternative alleles, a quality score regarding inference, info on filtering steps and the genotype format followed an additional column per sample genotyped. Plink's BED file system is composed of 3 files, a .bed file that stores samples genotype as a couple of bits, a .bim file which stores information on each variant (chromosome, variant identifier, position in Morgan or centiMorgan, base pair coordinate and alleles 1 and 2) and a .fam file storing sample information. To do so, both software perform linear regressions between phenotype, P, and genotype dosage, g, by using $P = \beta g + \epsilon$, in which β is the effect size given by slope of the regression, and ϵ represents the associated error. Correlation is assessed by measuring Pearson product-moment correlation coefficient. In order to account for multiple testing, a permutation scheme for the best correlated genotype per phenotype is implemented. The number of permutations has a fixed maximum but stops early if significance is not achievable, saving time and computational resources. After the maximum number of permutations is reached, a beta approximation that is capable of simulating p-value distribution of independent uniformly distributed random variable is implemented, thus overcoming computational restrictions of higher order level of permutations. This allows the retrieval of p-values below 1/p, where p is the number of permutations performed, without increasing computational load. Previously determined covariates were introduced, and mapping of cis variants was executed defining a window of 1Mb up- and down-stream of phenotype alternatively spliced site and a maximum of 10,000 permutations. Output from this tool consists of a table where each row is a molecular phenotype and columns comprise the identification of the event as tested molecular phenotype, number of variants in cis, and parameters regarding the implemented beta distribution. In this tool only the best associated variant is identified, as well as the distance between variant and molecular phenotype, nominal p-value, slope of the regression line, p-value obtained after permutation and p-value obtained after beta-approximation. Three additional columns regarding number of minor allele samples, minor allele count and MAF are also provided. Resulting beta p-values were later corrected for multiple testing using false discovery rate of 0.05 (as implemented in R) (Storey and Tibshirani, 2003).

3.8 BC GWAS loci retrieval

GWAS Catalog is a publicly available manually curated online repository of user submitted GWAS results for a wide variety of traits (Buniello *et al.*, 2019). As of 9th of September of 2020, it contained 4,694 publications and 197,708 associations. Besides providing study trait and associated hit-SNP, it also provides summary statistics information, ancestry data, and other relevant materials. Taking advantage of GWAS Catalog API level access to its database, gwasrapidd is a R package that provides a client interface (Magno and Maia, 2019). It allows fast retrieval, filtering and data integration into bioinformatics pipelines. Accessing GWAS Catalog via gwasrapidd, all GWASs regarding breast cancer risk were retrieved and filtered for those performed on individuals of European ancestry.

3.9 Co-localization of sQTL and GWAS signals

To assess if there is co-localization of obtained sQTL and GWAS signals I searched for local LD patterns. LDLink (Machiela and Chanock, 2015) is an interactive web-based tool that computes pairwise LD between variants. It does so by relying on phase 3 haplotype data from 1000 Genome Project. Among other features it allows selection of populations of specific ancestry, obtain not only pairwise LD data but also retrieve a matrix of LD from a group of variants within the same chromosome and obtain traits associated with a variant. Recently, the authors implemented a R package, LDlinkR, that bypasses interaction with the web based tool allowing programmatic access to information (Myers, Chanock and Machiela, 2020). Using LDlinkR I retrieved chromosome-wise LD matrixes for all variants identified as sQTLs and GWAS hit-SNPs using individuals of European ancestry. These were filtered by LD > = 0.4 to retrieve co-localizations between results from different approaches.

4 Results

4.1 RNA-seq Analysis

4.1.1 RNA-seq quality control and pre-processing

Quality control analysis of the RNA-seq data was performed with FastQC, before and after pre-processing, to remove low quality samples and reads besides trimming adapters that can interfere with subsequent analysis. From the large number of graphical outputs FastQC produces, I selected the ones that are relevant to justify the methods previously described. All the outputs are available in annex 8.1.

From the report we can assess that the number of reads per run ranged from 28.6M to 100.1M, with a mean of 45.4M reads before pre-processing. After removal of low-quality runs, the maximum and the average number of reads was 70.3M and 44.6M, respectively, which is within the range advised for this type of analysis (recommended 30-40M reads per sample) (Sheng *et al.*, 2017). The slight decrease in number of reads is believed to be due to the removal of samples that were highly duplicated.

The percentage of duplicated reads ranged from 16.5% to 59.4% of the total reads per run, which also decreased to 14.9% to 39.1% after pre-processing, due to removal of highly duplicated samples and trimming of the last nt in all reads.

The mean quality score showed that there is some variation in quality between runs. As shown in Figure 4.1A, both fastq files from run SRR655852 presented a lower mean quality score than the rest of the runs, dipping below acceptable quality score values. It is also visible a decrease in base call quality towards the edges of the reads with the last nucleotides being more affected by this issue. After trimming and filtering there were four runs that kept warnings due to Phred scores below 28, but still within acceptable values, above 20.

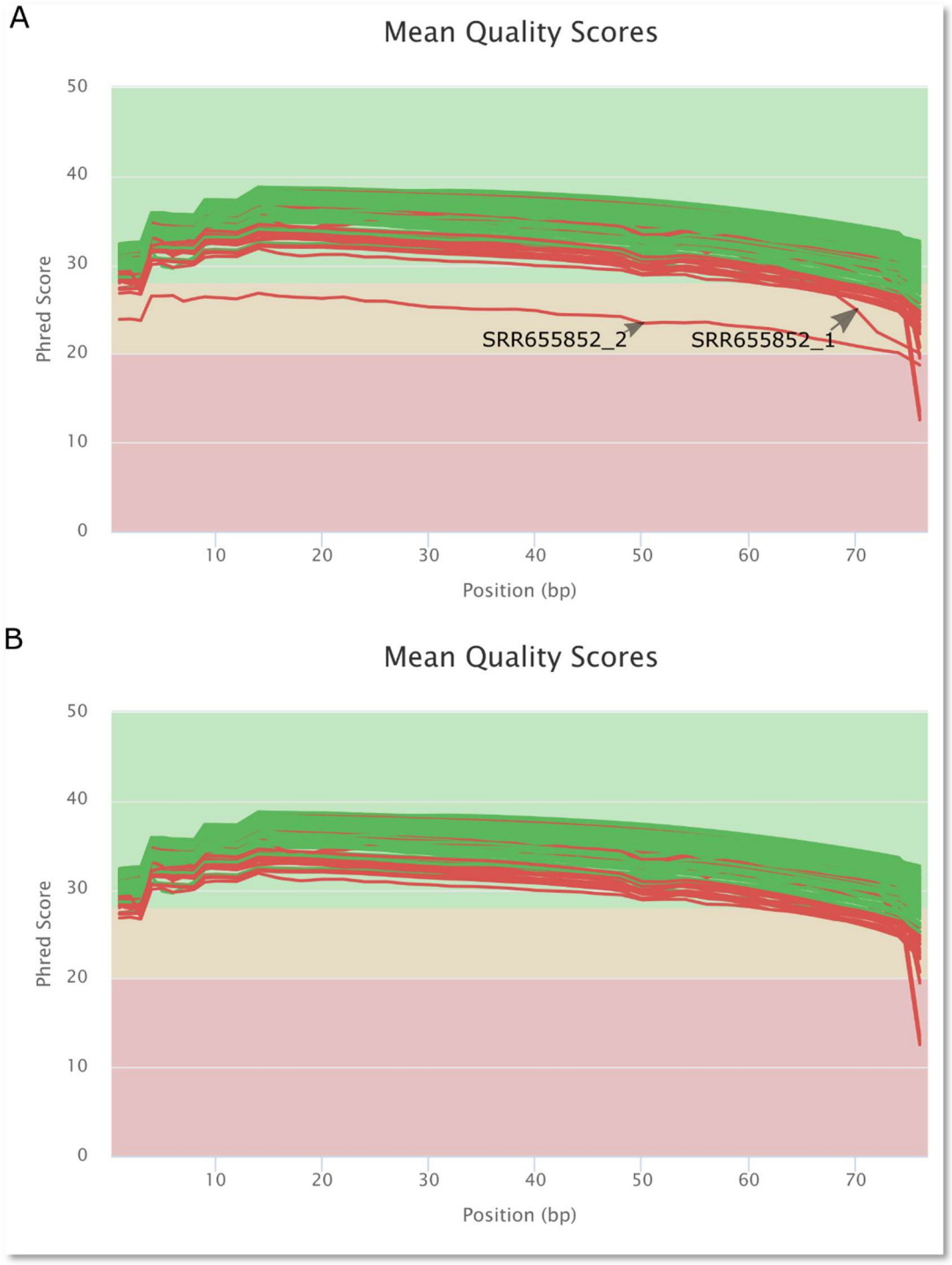


Figure 4.1 – Mean Quality Score per Sample. The above graph shows average quality of reads given by the Phred score (y-axis) per run (given by lines) in each position (x-axis), at pre-processing before (panel A) and after (panel B) trimming and removal of low-quality runs.

Per Sequence Quality scores showed a right shifted distribution of average quality per read. This shows that most of the sequences have a high quality, overcoming the necessity to perform per sequence mean quality filtering.

Per Base Sequence Content showed a weak bias in the first 11-13 nucleotides across all sequences and runs as represented in Figure 4.2, subpanels A and C, which is a bias introduced by the use of random hexamers in the amplification step. It appears to be towards the CTNAAATCYAT sequence, with different relative deviations from what would be expected if it were random, as shown in Figure 4.2D. Furthermore, in some runs there is a visible shift from random sequence, in dark grey, shown in Figure 4.2A. When showing a detailed view of a run with this type of problem, as in Figure 4.2B, a departure from randomness of sequence in these runs becomes clear, with spikes of a specific nucleotide throughout the read length. In some runs there is also some deviation from 24% to 26% nucleotide content range when looking at positions over 13, but it never goes above 29% or below 22% with complementary base showing the same trend.

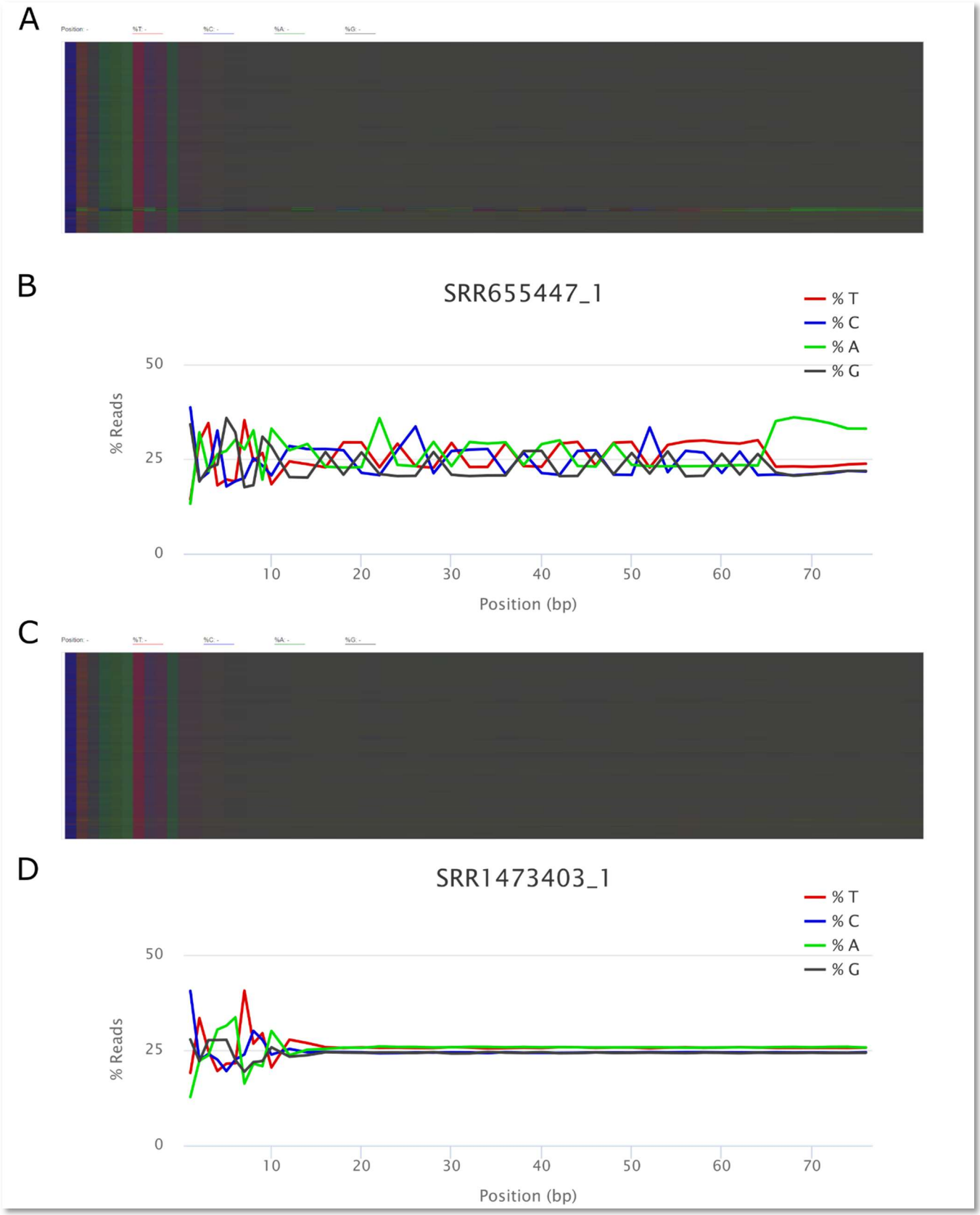

A
B
SRR655447_1
% T
% C
% A
% G
50
25
0
% Reads
10
20
30
40
50
60
70
Position (bp)
C
D
SRR1473403_1
% T
% C
% A
% G
50
25
0
% Reads
10
20
30
40
50
60
70
Position (bp)

Figure 4.2 – Per Base Sequence Content. In this figure is represented the proportion of each base per position related to DNA. T, C, A and G are represented as red, blue, green and black respectively and the x axis shows position in the sequence. In panels A and C each horizontal line represents a different RNA-seq run and is coloured by the average nucleotide content at each of the positions in pre-processed and post-processed runs. In panels B and D the y-axis represents the relative frequency of each base. Panel B shows a problematic run with nucleotide specific peaks in different positions. Panel D shows a run with peaks present in the first few positions.

Per sequence GC content of pre-processed RNA-seq reads showed a normal distribution with a mean of 49% GC content for the majority of the runs, with the exception of two as shown in Figure 4.3. The run SRR655852 showed a peak with 9.7% of reads having 45% GC content, while the run SRR655447 displays a smaller peak of 6.2% of all reads with 45% of GC content.

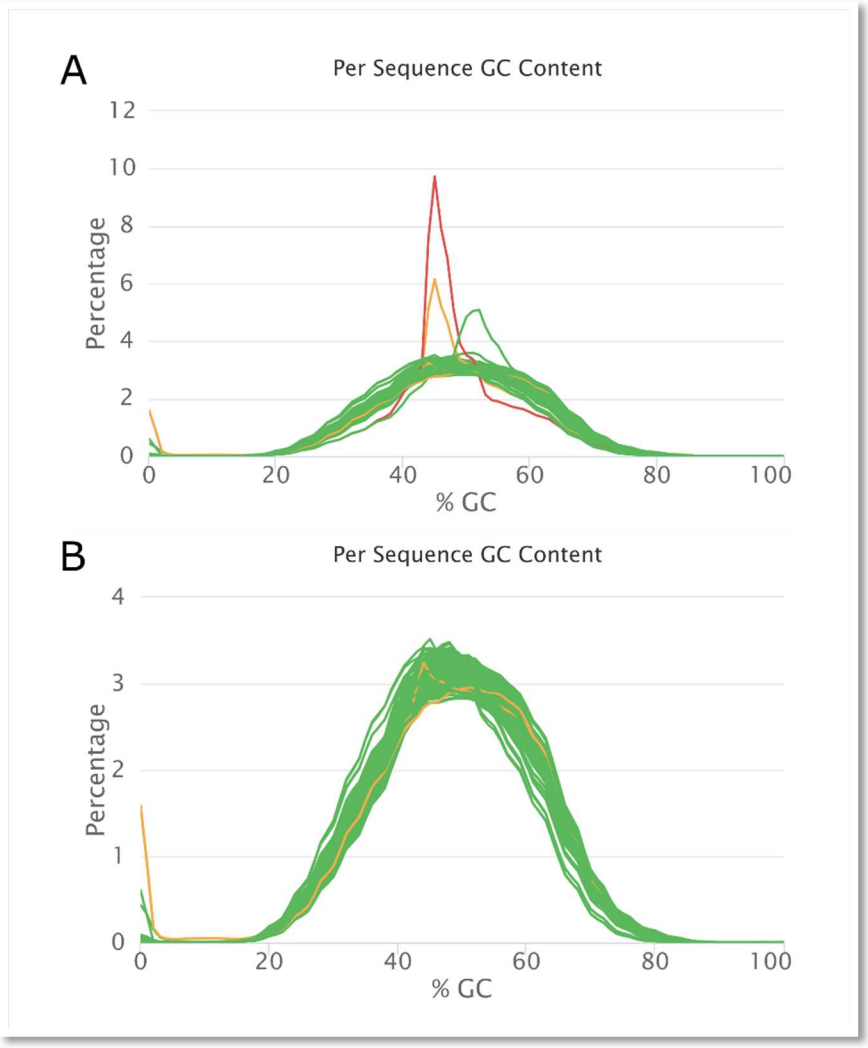


Figure 4.3 – Per Sequence GC Content. Distribution of GC content per sequence at each run before (A) and after (B) pre-processing step. In the y axis is the percentage of reads and in the x axis the percentage of CG content. Notice the change in scales between panels. In yellow are runs that triggered warnings (SRR655447) and in red failures (SRR655852).

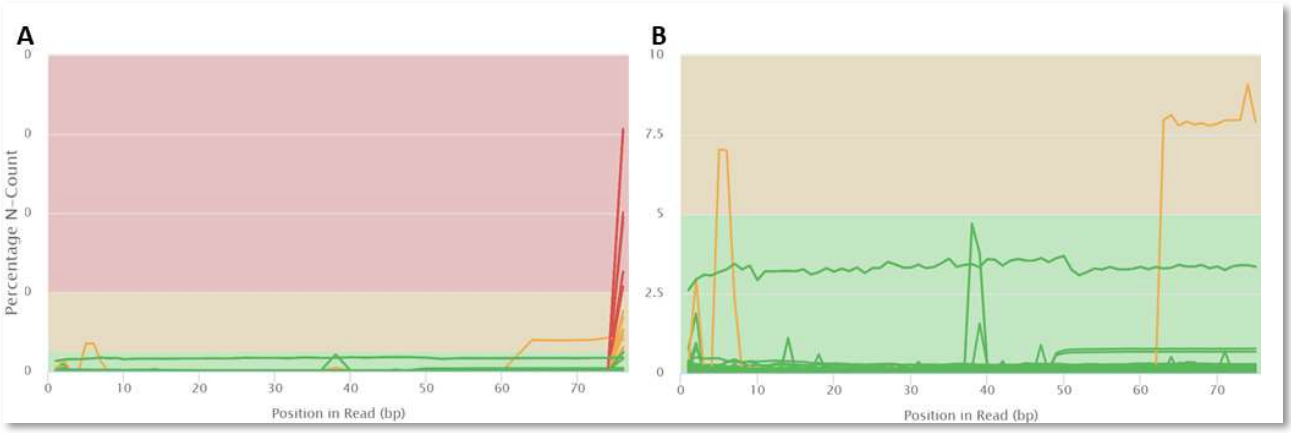


Figure 4.4 – Per Base N Content. In this graph we observe the percentage of N per position from all reads from each run. In panel A is shown the pre-processed and in panel B is post-processed. In red and yellow lines are runs that triggered failure or warning of this module. Notice that the y axis scale between both graphs is not the same equal.

Per base N content showed a decrease in quality at the last few positions in the sequences across most runs in pre-processed runs (**Figure 4.4**A), which was corrected through processing.

Analysis of sequence duplication levels showed a peak of duplications in some runs in the pre-processed data (Figure 4.5). In run SRR655852 the peak at 10k+ read counts accounts for 46.4% of all sequences from this run, while in run SRR655547 it accounts for 32.0% of all reads. These peaks disappeared after pre-processing steps were taken, as shown in Figure 4.5B.

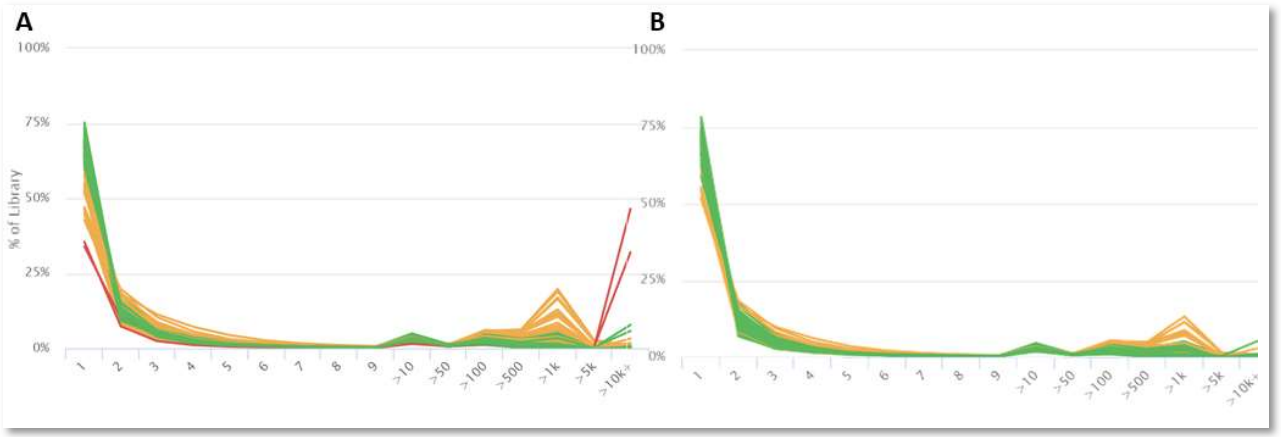


Figure 4.5 – Sequence Duplication Levels. This graph represents in the y axis the percentage duplicated sequences and in the x axis the number each sequence is represented. Panel A and B refers to pre-processed and post-processed data.

Analysis of overrepresented sequences showed a similar result to sequence duplication levels. Runs SRR655852 and SRR655447 showed approximately 20% and 10% for the top over-represented sequence, respectively, with the remaining over-represented sequences on these runs accounted for 8.25% and 0.29% of overall number of reads.

The adapter content analysis showed that run SRR655852 was enriched in adaptor sequences, with special incidence in the distal portion of the reads. Although no warning was triggered, this issue was solved in the post-processed runs as shown in Figure 4.6B, with a maximum of 0.44% of adapter content across all runs.

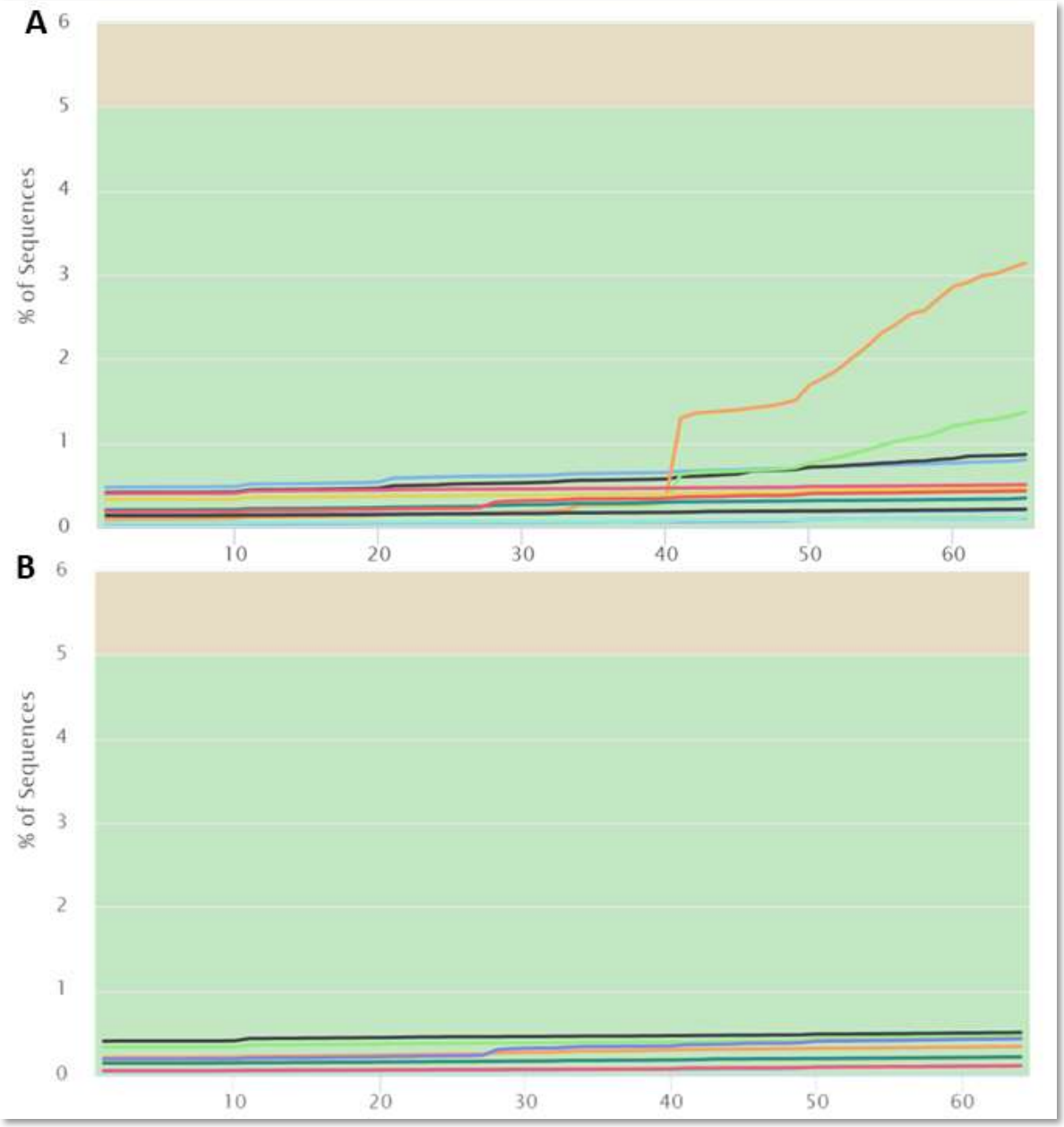


Figure 4.6 – Adapter Content. In this image is shown the percentage of sequences in each run that is identified as an adapter. Due to the number of runs, only those with higher percentage of adapters across each run are represented. Each run is displayed in a different colour. Panel A and B regard pre and post-processing steps respectively.

The final plots, **Figure 4.7**, are quality index heatmaps resuming all available quality scores across all available runs. Most important are the failures in pre-processed and post-processed data. Several failures are shown in red for: per base sequence quality, a single failure in per sequence

quality (run SRR655852), failure/warning in all runs from per base sequence content, a failure in per sequence GC content (SRR655852), a failure and warnings from some samples at per base N content, a failure for two runs in sequence duplication (SRR655852 and SRR655447) and three runs in overrepresented sequences. By comparing Figure 4.7 A and B a clear reduction in failures and warnings is clear, which confirms that processing was efficient.

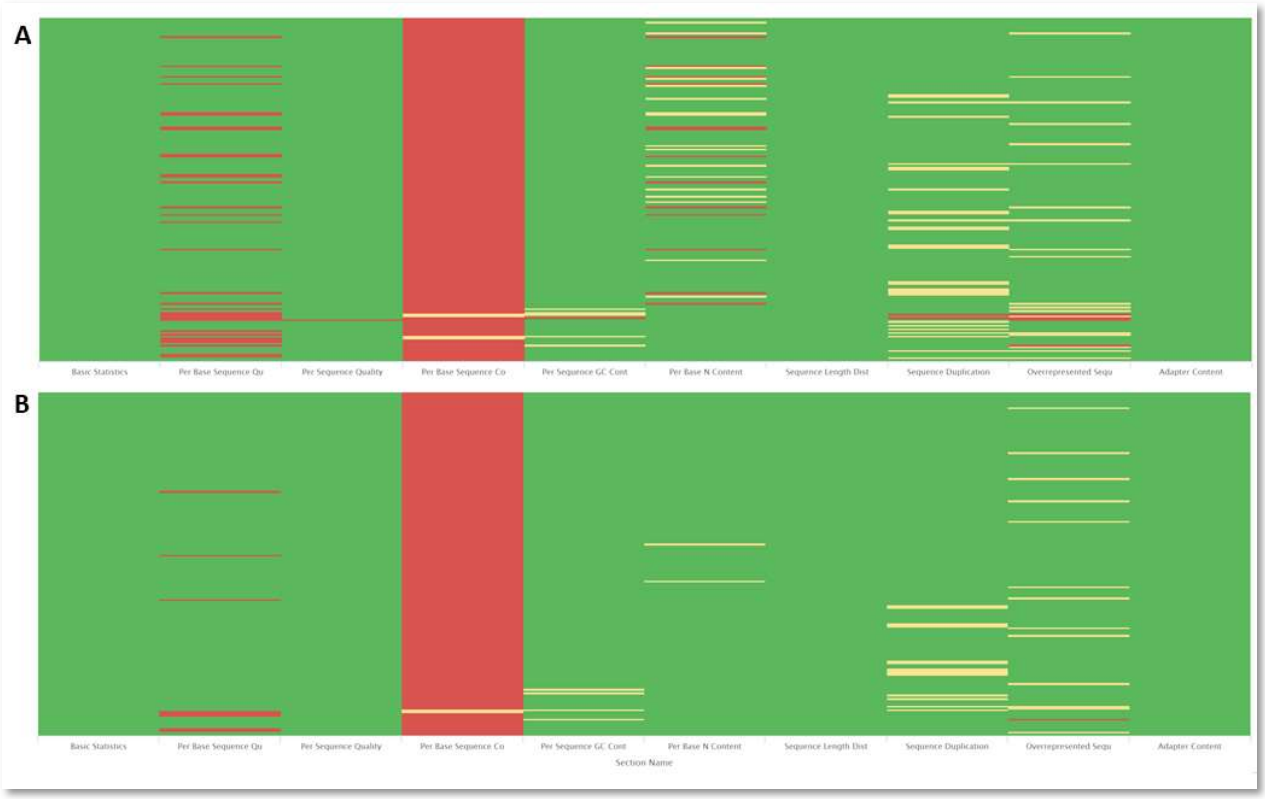


Figure 4.7 – Status Check. This image represents the overall status for each of our quality indexes generated by FastQC. Each line reports a different run and each column a different index. In green, yellow and red are represented the runs that passed, raised a warning or triggered a failure in each module. Panel A reports pre-processing results and panel B post-processing results.

After performing quality control, 78 samples with paired genotype information were used in subsequent analysis.

4.1.2 Read Alignment Quality Control

Read alignment was carried with STAR, which provides general alignment quality metrics. Percentage of aligned reads was very good (Dobin and Gingeras, 2015), ranging from 85.8% to

93.4%, corresponding to a minimum of 26.2M reads and a maximum of 62.7M reads aligned per sample. When considering the relative frequency of uniquely aligned reads (Figure 4.8A), all samples showed a high alignment rate averaging at 90.21%. Unaligned reads were mostly due to short read length, which impaired matching to the reference genome. The number of reads aligning to multiple loci was also small for all runs, averaging at 5.41%. The absolute count of reads aligned per sample was more variable across samples, with a visible dispersion in unique reads aligned per sample, showing a larger than two-fold difference between both extremes. The number of splice sites detected ranged from 9.5M to 23.1M per sample, with an average of 15.2M splice sites.

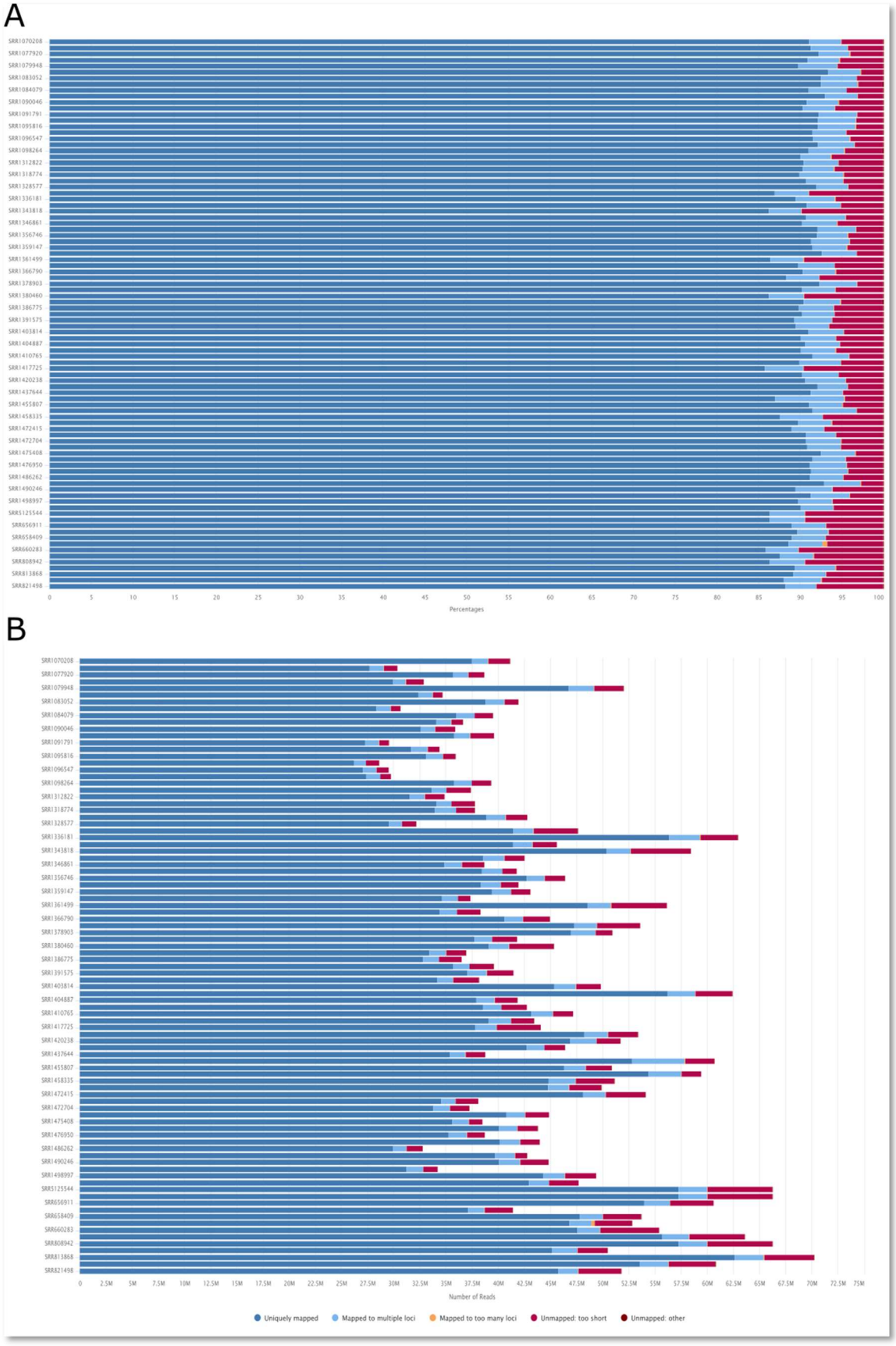

A
SRR1070208
SRR1077920
SRR1079948
SRR1083052
SRR1084079
SRR1090046
SRR1091791
SRR1095816
SRR1096547
SRR1098264
SRR1312822
SRR1318774
SRR1328577
SRR1336181
SRR1343818
SRR1346861
SRR1356746
SRR1359147
SRR1361499
SRR1366790
SRR1378903
SRR1380460
SRR1386775
SRR1391575
SRR1403814
SRR1404887
SRR1410765
SRR1417725
SRR1420238
SRR1437644
SRR1455807
SRR1458335
SRR1472415
SRR1472704
SRR1475408
SRR1476950
SRR1486262
SRR1490246
SRR1498997
SRR5125544
SRR656911
SRR658409
SRR660283
SRR808942
SRR813868
SRR821498
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percentages
B
0 2.5M 5M 7.5M 10M 12.5M 15M 17.5M 20M 22.5M 25M 27.5M 30M 32.5M 35M 37.5M 40M 42.5M 45M 47.5M 50M 52.5M 55M 57.5M 60M 62.5M 65M 67.5M 70M 72.5M 75M
Number of Reads
Uniquely mapped
Mapped to multiple loci
Mapped to too many loci
Unmapped: too short
Unmapped: other

Figure 4.8 – STAR alignment scores. Most reads were uniquely aligned to reference human genome with deeper sequencing providing a higher absolute number of mis-aligned or multiple aligned reads.

In here is visible the general alignment score retrieved from STAR. Each horizontal line represents one run. In panel A the relative count is reported while on panel B the absolute count is shown. In darker blue is show uniquely mapped sequences, in lighter blue and in yellow sequences that were aligned to multiple (> 1 and ≤ 10) or too many loci (> 10), in eggplant is shown unmapped reads when alignments are too short and in maroon are reads that were not mapped due to other motives.

4.2 Alternative Splicing Quantification

Once the data was appropriately processed and filtered, and the quality control measurements were satisfactory, I proceeded to identify and quantify all alternative splicing (AS) events.

4.2.1 Leafcutter

LeafCutter was used in order to quantify AS events across all samples. One important aspect of this tool is that it can identify new AS junctions from aligned BAM/SAM files without prior annotation. In total, 340,872 unique splice site junctions in autossomic chromosomes were identified. The size of each identified intron ranged between 22 bp to 498,598 bp, with an average intron size of 10,652 bp and a median of 2,303 (Figure 4.9).

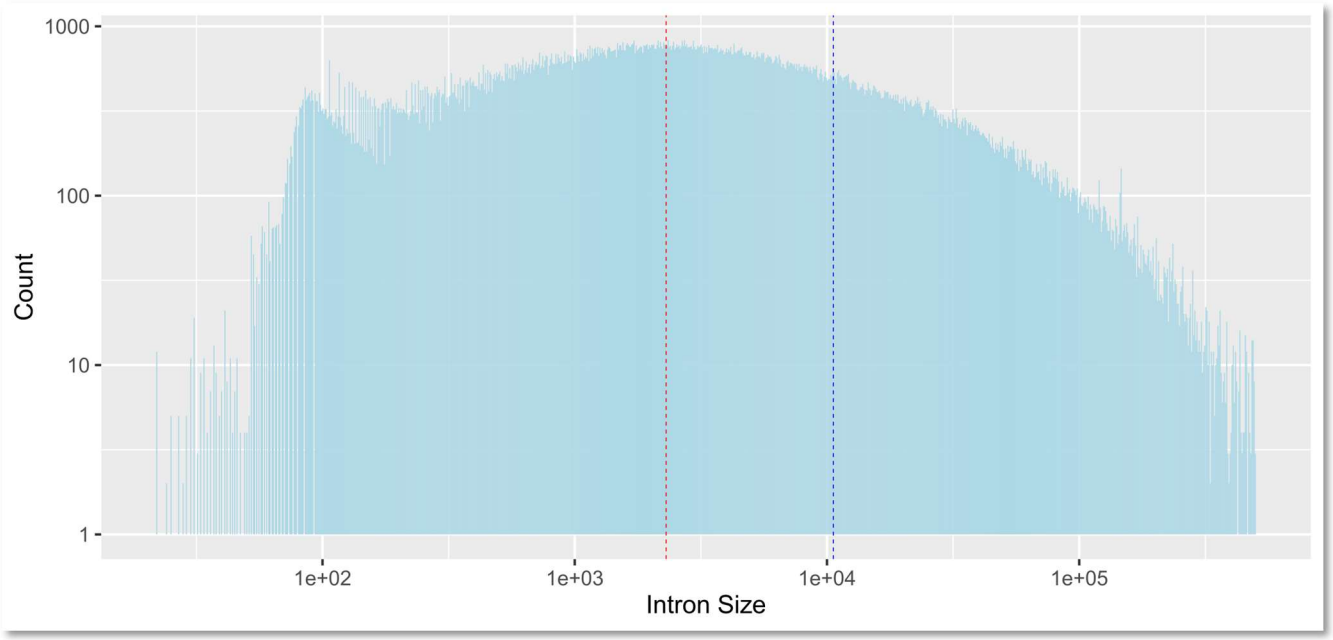


Figure 4.9 – Intron Size Distribution. The red and blue dashed line represent the median and the mean of intron size of 2,303 and 10,658 respectively. Notice that both x and y scales are logarithmized.

Read count per intron is defined as the sum of read counts for all samples available in an intron as identified by LeafCutter. It ranged from a minimum to five to 1,444,663 (Figure 4.10A). This maximum is regarding an intron in *RPL13A*, a component of the large ribosomal subunit 60S. When looking in detail at its distribution a left leaning curve is visible with few introns (0.029 of all introns detected or 9940 introns) with more than 10k reads (Figure 4.10B). In a detailed view of introns with less than 10k reads, we can observe a mean of 1,521 and a median of 50 (Figure 4.10C).

4 Results

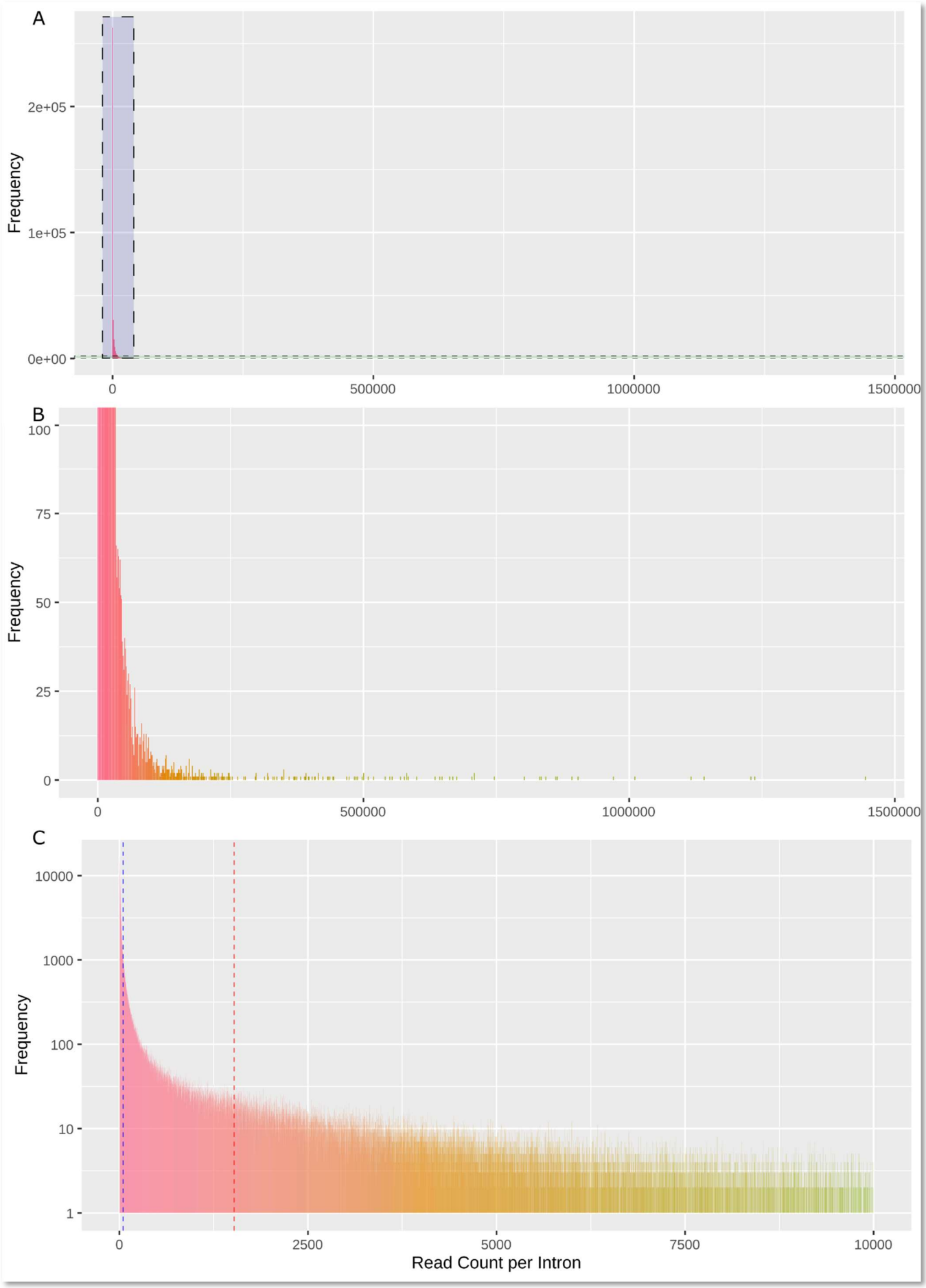

A
2e+05
1e+05
0e+00
Frequency
0
500000
1000000
1500000
B
100
75
50
25
0
Frequency
0
500000
1000000
1500000
C
10000
1000
100
10
1
Frequency
0
2500
5000
7500
10000
Read Count per Intron

Figure 4.10 – Frequency of Read Count per Intron. These panels show in the X axis the number of reads per intron and in the Y axis its absolute frequency. In panel A we can observe a strong left leaning histogram of read count per intron revealing that most introns were rarely detected. In B I increased the resolution by setting the scale of the y axis to a maximum of 100 reads. This detailed view shows that few introns have a large read count creating a long tail with an intron showing a maximum of 1,444,663 reads aligned to an intron. In panel C the binwidth decreased to one in order to obtain a better spread of the number of reads per intron on introns that have up to 10,000 reads. The vertical dashed lines in red and blue are the mean and median of the overall distribution at 1,52 and 50 reads per intron respectively.

These junctions were grouped by the software into 73,944 unique clusters with more than one splice site per cluster. The number of splice sites per cluster ranged from 185 to two, with an average of 4.61. Looking closer at the cluster with displaying the maximum number of alternative introns, 185, it spans just 2121 bps (chr9:137,101,870-137,103,991) and is aligned to gene *MAN1B1*. This cluster has 49 unique alternative 5' splice sites and 48 unique alternative 3' splice sites. Plotting the distribution of introns per cluster (Figure 4.11) we can observe that it is a left leaning distribution with most clusters (41,739 or 0.5645) having less than four alternative splice patterns and only 0.085 of all clusters presenting more than 10 introns.

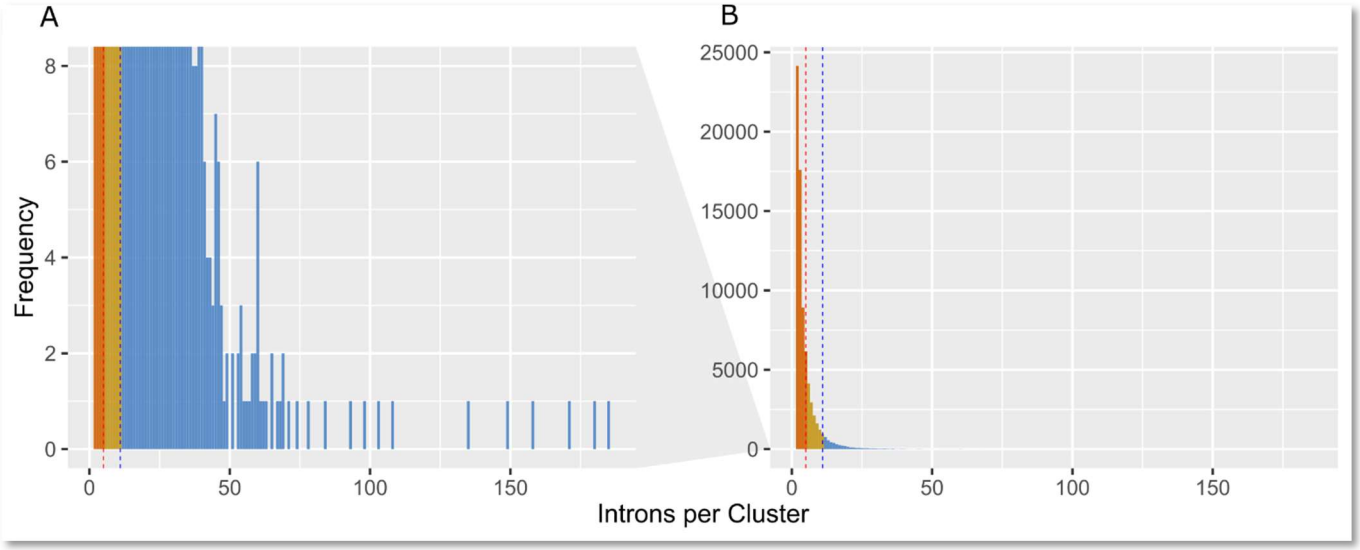


Figure 4.11 – Number of introns per cluster as identified by LeafCutter. In this two-part histogram is plotted the frequency of introns per cluster in different scales of the y axis to allow a better visualization. At each panel the vertical dashed lines are dividing the distribution at four and ten, showing to the left of them a cumulative distribution of 0.565 and 0.915 of all clusters. Panel B shows a general

view of the distribution, where we can see how strongly left leaning distribution is. In panel A a more detailed view displays in better detail the higher intron count clusters.

Next, clusters were filtered based on a minimum of 50 reads per cluster in order to increase specificity by removing misalignments and low splice junctions. This filter reduced the number of clusters to 68,306 and the number of unique splice sites to 314,875, representing a loss of approximately 7 % for both parameters (Figure 4.12).

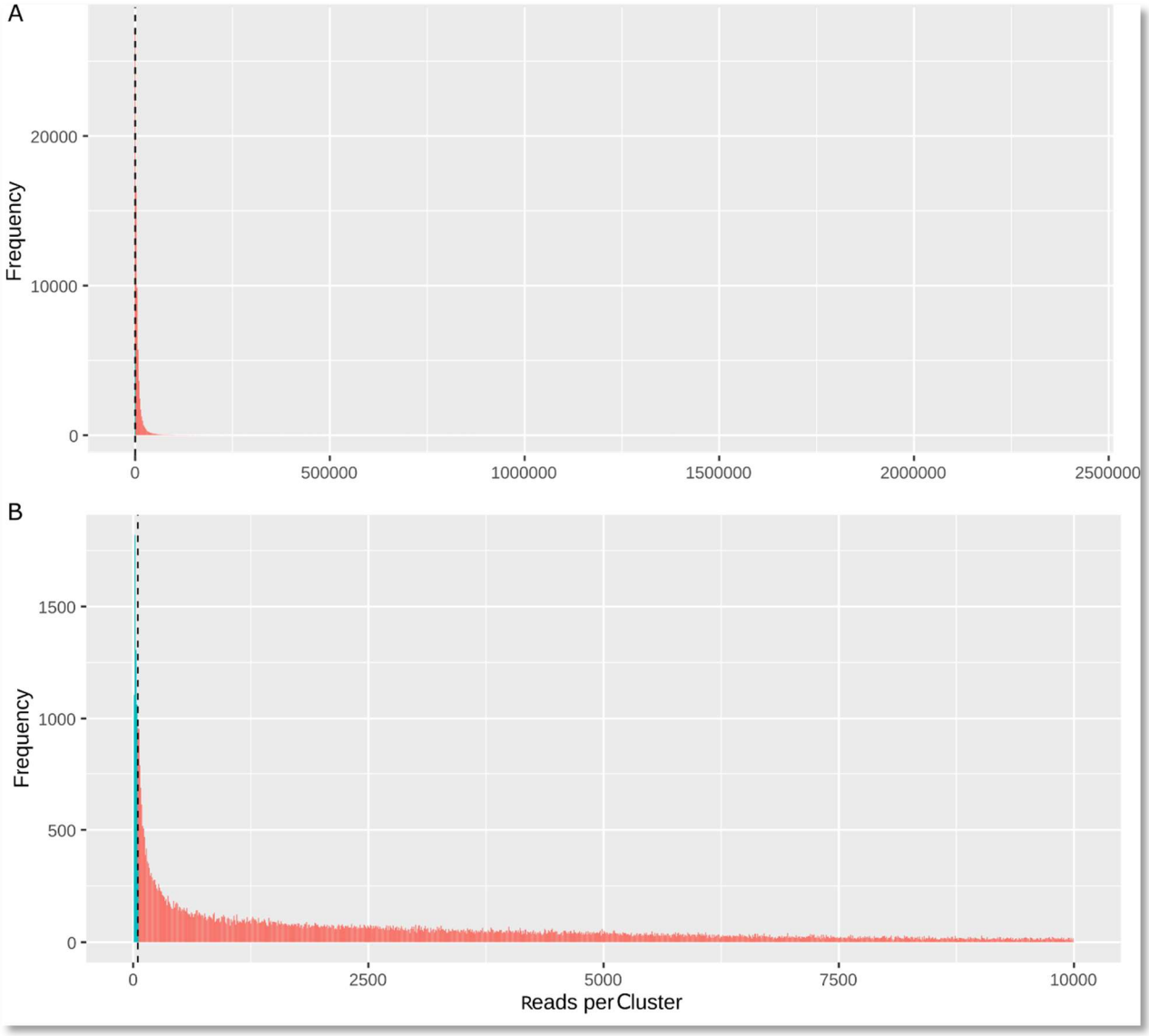


Figure 4.12 – Frequency of number of reads per cluster. In this image the frequency of the read counts per cluster is plotted. An uneven distribution was obtained as most reads (0.505) have less than

2,500 reads. In panel A is shown in the X axis the full range of reads per cluster, from ten to 2,390,898. In panel B the X axis is zoomed in, between zero to 10,000 showing 0.838 of all clusters. The vertical dashed line at x= 50 separates to the left the clusters that will be lost by the filtering step in blue.

A second filtering step was applied after PSI calculation, which removes all lines regarding unused introns (defined as introns with a PSI less than 0.001), as well as those junctions with little variation across samples, reducing the total count of alternatively spliced introns to 107,730, reporting to 55,813 unique clusters.

4.2.2 Psichomics

The second tool used was psichomics, which quantifies previously defined AS events, the list of which is provided in the software. Taking advantage of psichomics's features, I used GTEx pre-processed data from 83 samples which had matching genotype available. This dataset diverges in part from previously data with only 75 samples in common. There are eight new samples unique to GTEx pre-processed data while three samples were only available from RNA-seq data.

A total of 107,413 AS events were retrieved and divided into the six AS event types, as annotated by the package developers. On a closer look, many of these events were not quantified for some of the samples and were labelled as NAs (i.e., Not Available). In fact, a large proportion of them accounting for a full row of NAs, meaning that those events were not detected in any of the analysed samples. This is possibly due to gene expression and AS events being tissue-specific and, therefore, absent in breast tissue RNA, but also some rarer events that would require even deeper sequencing experiments to reach the minimum detection limit.

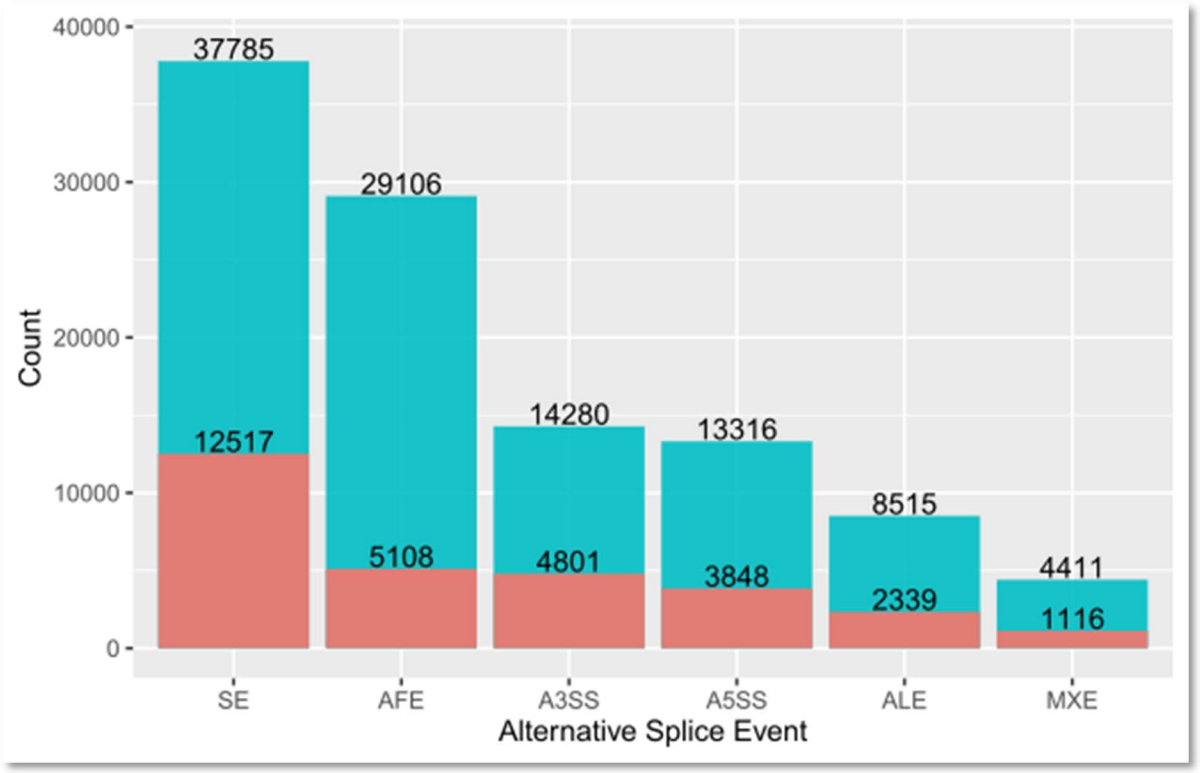


Figure 4.13 – Number of Alternative Splice Events detected with psichomics. In teal and salmon are visible the count of alternative splice events from each type before and after filtering out events containing NA values.

After all AS events with NA results were removed, the count of events was reduced to 29,729 (Figure 4.13), with a similar rank of frequency between all six classes as previously. The most common class detected was “skipped exon” (SE), accounting for approximately a third of the events, while the least common class was “mutually exclusive exons” (MXE), accounting for only three percent of all events. Importantly, some of these are not unique events and may be allocated to more than one class. For example, a skipped exon (SE event) might be due to the generation of an alternative 3’ splice site (A3SS), in which it would appear in both SE and A3SS classes, as exemplified in Figure 3.3.

4.3 Co-variates analysis

Principal Component Analysis was performed on intron read counts in order to assess co-variates. Cumulative variation per principal component (PC) showed that PC1 accounted for 27.11% of all variation in psichomics data (**Supplementary Figure 1**) while only 22.25% for LeafCutter data (**Supplementary Figure 3**). Plotting PC1 vs PC2 for each method of PSI quantification did not result in any obvious clustering between samples (**Supplementary Figure 2** & **Supplementary Figure 4**).

4.4 QTL mapping

As one of the central objectives of this thesis was to identify variants associated with AS, we next proceeded with the splicing QTL (sQTL) mapping exercise using the TensorQTL software. This was performed independently for the results generated via psichomics and LeafCutter. sQTL mapping results are shown below in separate for each tool employed to calculate PSI, and only after are integrated to give a general view of breast tissue splicing patterns. From this point onwards, an sQTL is defined as an alternative splicing event whose variation was correlated with alternative allele content at that position, and a sGene is any gene that has at least one sQTL identified.

4.4.1 Psichomics sQTL

TensorQTL analysis using the splicing phenotype measured by psichomics, was only applied to 28,603 alternative splicing events, with a loss of 1,126 events (0.038) comparing to the input because of low inter-sample variation events, which are filtered out during sQTL analysis.

Comparing significant events before and after applying false discovery rate correction we obtained 2,065 and 382 sQTLs when setting a threshold of $\alpha = 0.05$, and 859 and 287 sQTLs when lowering the threshold to $\alpha = 0.01$, as shown in Figure 4.14. Also, a total of 343 unique sGenes (using ensemble gene nomenclature) were identified in this analysis.

One of the features of tensorQTL is that it reports only the best correlated sQTL for each AS event. Inspecting these, I found that for 7,489 events there was only one sample with at least one alternative allele for the best correlated sQTLs. Nevertheless, many (7444 or 0.994) of these

results are subsequently corrected as their beta adjusted p-value is greater than 0.05, and so they are not significant when multiple test correction is applied.

The mean distance between the 5'splice site and the best correlated variant is 3.300 with a median of 101 bps. 0.693 of all significative sQTLs fall within a range of 25k bps upstream or downstream from AS event boundary. This metric may be skewed by the definition of an upper distance threshold of 1M bp from the mapped intron when performing cis-sQTL mapping.

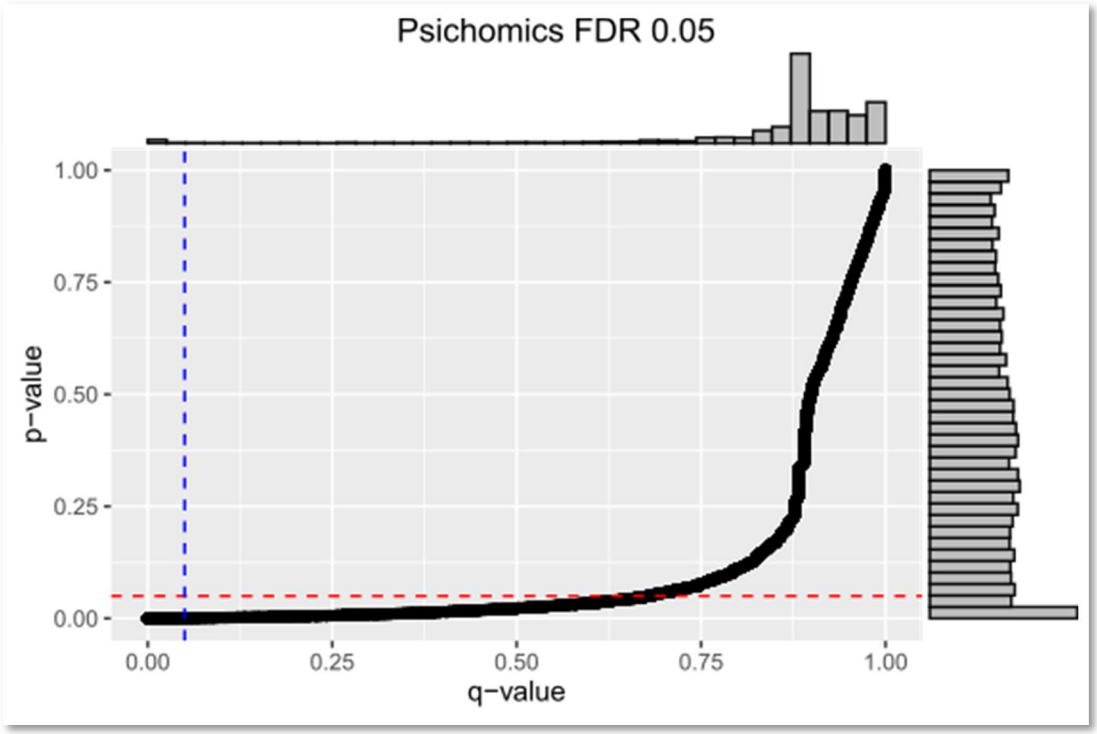


Figure 4.14 – Psichomics sQTL FDR correction. Comparison between beta-modeled p-values in the vertical axis and FDR corrected q-value in the horizontal axis. At the margins of the main plot are the histograms for p-values in the right and for q-value at the top.

The complete table of significant sQTLs after FDR implementation is available in Annex 8.7.

4.4.2 Leafcutter sQTL

When performing QTL mapping using the LeafCutter computed phenotype a small number of introns, 32, were lost due to low inter-sample variance. The total number of splice events inputted for QTL mapping was 107,698. Of these 8,484 and 3,327 achieved a beta-adjusted p-value

below 0.05 and 0.01, respectively. When adjusting for the number of multiple tests performed, we obtained 1,467 or 1,127 sQTLs, depending on the used threshold that was previously mentioned as visible in Figure 4.15.

Similarly to previous sQTL mapping, we found 24,671 splicing events whose best correlated variant only had one sample with an alternative allele. Likewise to sQTLs obtained using psichomics computed phenotype, none of these events reached significance after applying FDR correction. Here, 961 sGenes were identified with at least 1 alternative splicing event.

The best correlated variant for significant sQTLs is, on average, at 12.209 bp upstream of alternative splice site while 0.541 of all events range between -25k to 25k.

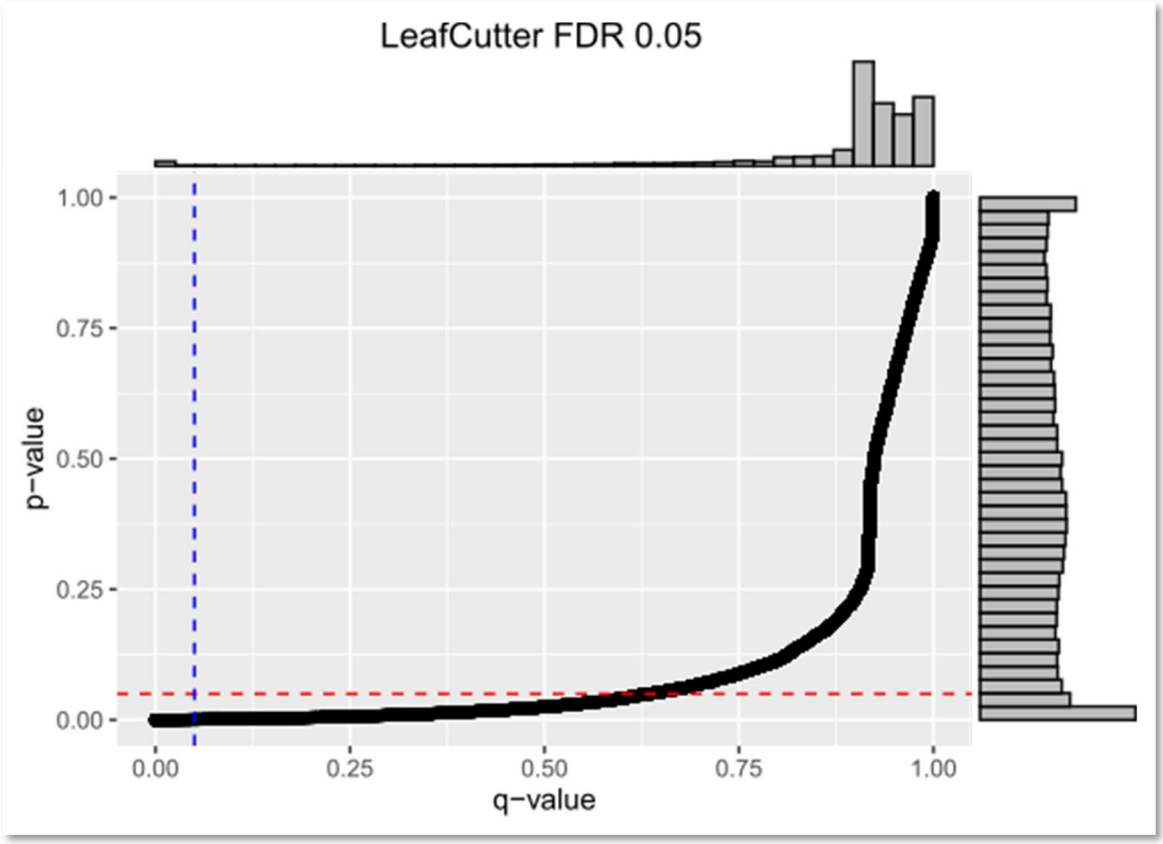


Figure 4.15 – LeafCutter sQTL FDR correction. P-values are represented in the y-axis and FDR corrected q-values are on the x-axis. On the right-hand margin, is the histogram of p values and at the top is the histogram of q-values.

As before, a table of significant sQTLs after FDR implementation is available in Annex 8.6.

4.4.3 Comparing sQTLs

As tensorQTL only reports the best correlated variant per splicing event, it can identify different variants in the same locus that are highly associated between themselves, as measured by a correlation coefficient. As such, when comparing variants identified as sQTL, we are comparing not only individual variants to verify overlap, but also assessing the linkage disequilibrium (LD) between them. There were 50 loci identified by both methods of calculating PSI, reporting to 63 and 61 unique sQTLs using phenotypes processed by psichomics and LeafCutter, respectively. There were also 222 sGenes in common between both analyses.

4.5 Identification of sQTLs associated with risk to BC

4.5.1 Retrieval of GWAS associated variants for BC risk

Next, 93 GWAS studies on breast carcinoma were identified in the GWAS Catalog online repository. Closer analysis revealed that some of these were not only on breast cancer, but also on other subjects related to it. Examples are adverse response to chemotherapy and progression free survival. Filtering by reported trait left 56 GWAS in total. Given GTEx samples used were of European ancestry (Gay *et al.*, 2020), I filtered out GWAS performed on other populations, reduced the number further to 41.

From these 41 studies 1,249 risk associated variants were identified, of which 942 were unique reported variants .Next I filtered variants to keep only those reaching the threshold for genome-wide statistical significant of 5×10^{-8}, which left 701variants, of which 475 were unique variants. For completeness purpose, all variants with a p-value higher than 1×10^{-5} were kept for further analysis as new GWAS may validate variants that previously did not reach statistical significance (Panagiotou *et al.*, 2012). The full list of variants retrieved is available at annex 8.8 (page CXXI). Looking at reported function, most (0.846) are intronic or intergenic variants (Figure 4.16).

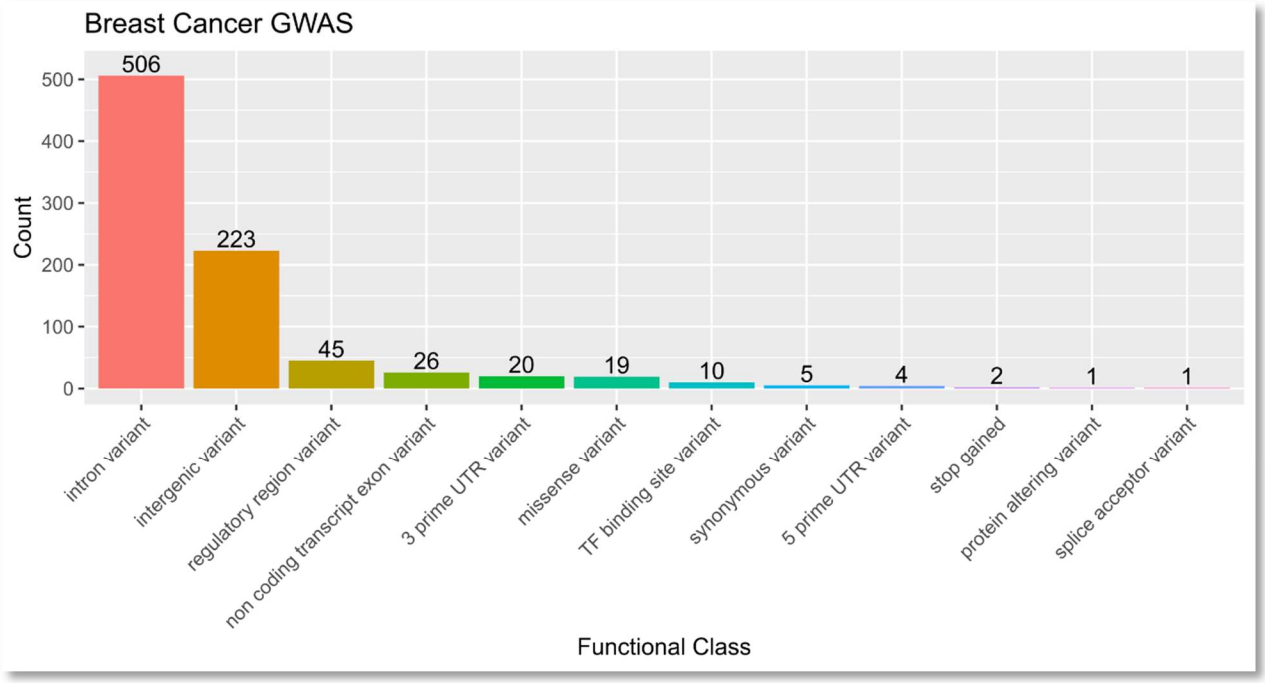


Figure 4.16 – Breast cancer genome wide association study hit-SNPs attibuted functional class.

Looking at the odds-ratio of the risk allele in available hit-SNPs, it ranges from 1.02 to 1.60 with a mean of 1.109.

4.5.2 sQTLs in LD with GWAS hit-SNPs

To assess if variants associated with changes in alternative splicing were also associated with changes in breast cancer risk, I used the LD measured as a correlation coefficient, defining a minimum threshold of $r^2 \geq 0.4$ to ascertain co-association.

From statistically significant sQTLs retrieved using LeafCutter approach, seventeen variants were in strong LD with 29 hit-SNPs retrieved from GWAS on BC in fourteen loci. These report to changes in splicing pattern in 19 genes (Table 1 – sQTLs obtained using LeafCutter's PSI in co-localization with BC GWAS hit-SNPs).

Regarding psichomics derived sQTLs, seven sQTLs were in strong LD with nine variants previously associated with BC, spread over seven loci. These changes in splicing patterns were detected in seven different genes. (Table 2 – sQTLs retrieved using PSI as calculated by psichomics in linkage disequilibrium with breast cancer GWAS hit-SNPs)

The intersection between both sQTLs and GWAS hit SNPs retrieved three risk-associated loci in chromosomes 1, 11 and 15, with significant sQTLs for genes *PARK7*, *BANF1* and *ULK3*.

4.5.3 Risk locus 1p36

Changes in splicing pattern of the gene Parkinsonism Associated Deglycase or *PARK7,* located in locus 1p36, were detected by both splicing analysis tools. This gene is coded in the forward strand and the measured events regard the size of the first intron, specifically whether the 5' splice site is at position chr1:7,961,735 or at chr1:7,961,793, using the 3' splice site at position chr1:7,962,763 (Figure 4.17, A and B). Using LeafCutter, the alternative T allele of rs4908724 is associated with two different events in the same cluster, with a significant increase in use of the larger first intron (indicated as spliced out intron D in Figure 4.17B), and independently with a decrease in the use of the smaller first intron (indicated as spliced out intron A in Figure 4.17B). The absolute effect size is 0.0946, with a q-value = 6.75×10^{-3}, for both events (Figure 4.17C, 2 and 4).

A concordant splicing event in the same intron was identified using psichomics, but mapped to a different variant, rs17229081 (Figure 4.17C, 1 and 3). According to psichomics' annotation, this is classified as an alternative first exon (AFE) and an alternative 5' splice site (A5SS). The association found is between the alternative allele A of rs17229081 the increase of usage of the D intron and a decrease of usage of the A intron. The absolute effect size retrieved per alternative allele is 0.105 in both associations, concordantly in opposite directions, and the FDR corrected p-value is 1.96×10^{-5} and 1.99×10^{-5} for AFE and A5SS, respectively. Both sQTLs, rs4908724 and rs17229081, are in strong LD with each other ($r^2 = 0.78$).

According to GTEx data, this gene is expressed in healthy female breast tissue (median of 185 transcripts per million,TPM), and produces six alternative splicing isoforms (defined as higher than 1 TPM (Wagner, Kin and Lynch, 2013)) (Figure 4.18). ENST00000338639 is the main transcribed isoform and is the only with the previously mentioned A intron. Intron D is only present in isoforms ENST00000493678 and ENST00000377493, which have a lower expression. Both sQTLs identified are also eQTLs for *PARK7* in several tissues including breast. rs4908724 is associated with a normalized effect size (NES) of 0.0847 with an association p-value of 5.6×10^{-4} (Supplementary Figure 5), while rs17229081 has a NES of 0.0635 and an association p-value of 0.02 (Supplementary Figure 6).

These sQTLs were also in strong LD with the BC risk-associated variant rs225132 ($r^2 = 0.964$ and $r^2 = 0.767$ for rs4908724 and rs17229081, respectively), whose G allele was associated with an increase in breast cancer risk in 2017 (p-value = 3×10^{-6} ,effect size = 0.0375) (Michailidou

et al., 2017). As the risk allele rs225132(G) is correlated with the alleles rs4908724(T) – D' = 0.9927 – and rs17229081(A) – D' = 0.8952 –, this suggests that BC risk is linked to the increase in usage of a larger first intron in gene *PARK7* (Figure 4.17, C). This implies that risk allele rs225132-G is associated with a lower expression of isoform ENST00000338639 in exchange for an increase in either ENST00000493678 and/or ENST00000377493.

Using Cobalt (Papadopoulos and Agarwala, 2007) to compare protein sequences between the isoforms of interest, I found that ENST00000377493 produces a smaller protein with only 169aa, due to exon 6 skipping that encodes for amino acids 65 to 84 (Supplementary Figure 7). Although there is no available annotation on protein domains, what we can report is that loss of the exon 6 impacts the secondary structure by removing a β-strand and a α-helix that would be coded by amino acids 68-72 and 76-84 (Lakshminarasimhan *et al.*, 2008; *PARK7 - Parkinson disease protein 7 precursor - Homo sapiens (Human) - PARK7 gene & protein*, 2020). Furthermore, UTRs are also impacted by splicing into different isoforms, changing not only their length but also sequence.

4 Results

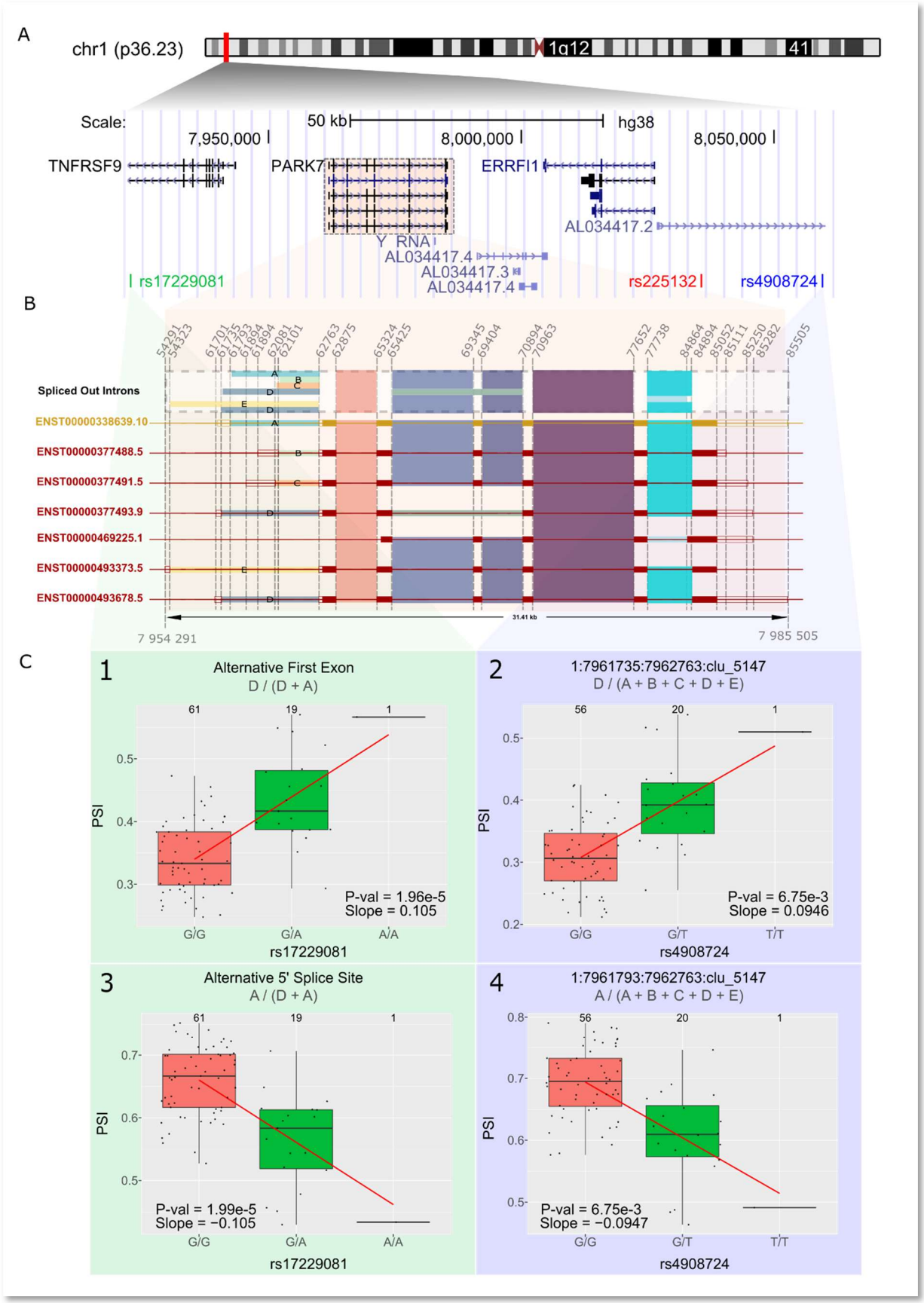

A
chr1 (p36.23)
1q12
41
Scale:
50 kb
hg38
7,950,000
8,000,000
8,050,000
TNFRSF9
PARK7
ERRFI1
AL034417.2
Y_RNA
AL034417.4
AL034417.3
AL034417.4
rs17229081
rs225132
rs4908724
B
Spliced Out Introns
ENST00000338639.10
ENST00000377488.5
ENST00000377491.5
ENST00000377493.9
ENST00000469225.1
ENST00000493373.5
ENST00000493678.5
31.41 kb
7 954 291
7 985 505
C
1
Alternative First Exon
D / (D + A)
P-val = 1.96e-5
Slope = 0.105
G/G
G/A
A/A
rs17229081
2
1:7961735:7962763:clu_5147
D / (A + B + C + D + E)
P-val = 6.75e-3
Slope = 0.0946
G/G
G/T
T/T
rs4908724
3
Alternative 5' Splice Site
A / (D + A)
P-val = 1.99e-5
Slope = −0.105
G/G
G/A
A/A
rs17229081
4
1:7961793:7962763:clu_5147
A / (A + B + C + D + E)
P-val = 6.75e-3
Slope = −0.0947
G/G
G/T
T/T
rs4908724
PSI

Figure 4.17 – sQTL mapping in the BC risk locus 1p36. **A** - Ideogram for chromosome one and the segment of interest is expanded underneath showing genes encoded. The relevant variants are represented in the bottom of panel A in green showing psichomics sQTLs, in red breast cancer GWAS hit-SNP and in indigo LeafCutter's sQTL. **B** - annotated coding transcripts for *PARK7* available from Ensembl. This annotation is provided either by Ensembl or Havana in red or by both in gold. The boxes in red and gold represent exons for which empty sections are UTRs. Between filled boxes of exons are introns identified by different colours representing possible alternative splicing patterns for the coding transcripts. These splicing patterns are summarized in the upper part of panel B, naming relevant introns. Relative size of introns and exons are not to scale. **C** - relevant sQTLs for this locus. In the vertical axis of each subpanel is PSI and on the horizontal axis is the genotypes measured at each variant, with the corresponding number of samples per genotype at the top of each box. At the top of each subpanel is the relevant designation for the event from each tool used to assess PSI and below in grey is how PSI is computed. The distribution of PSI per genotype group is represented by a box with each individual sample represented by a dot. The red line is the linear regression as implemented by TensorQTL and its FDR corrected p-value and slope is written in the bottom of each graph within drawing area.

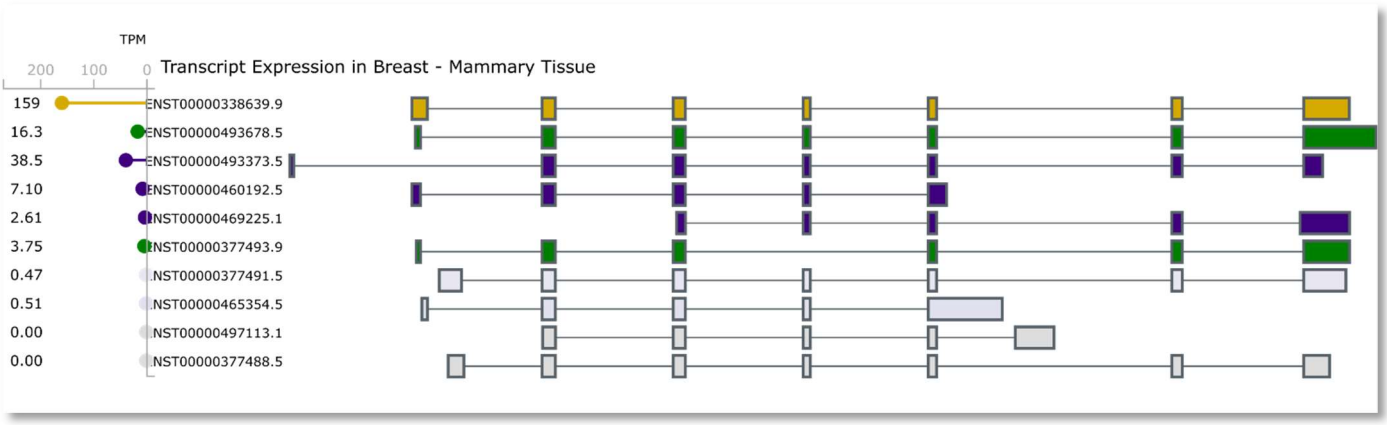


Figure 4.18 – *Park 7* isoform expression in breast tissue. In vertical order is from the most to least expressed isoform across all tissues. In gold and green are isoforms whose splicing changes were associated with the presence of alternative alleles. The expression of each isoform in breast measured as TPM is shown in the left side of the figure as obtained from GTEx project (Ardlie *et al.*, 2015).

4.5.4 Risk locus 11q13

Three variants were identified as sQTLs on chromosome 11 relating to changes in the splicing pattern of the gene *BANF1* (BAF Nuclear Assembly Factor 1) (Figure 4.19 – sQTL mapping identified splicing changes in BC risk associated locus 11q3. **A -** this region is located on the bigger arm of chromosome 11, on band 13.1. Near the bottom of the panel are represented variants that were identified as sQTLs, rs617791 from LeafCutter in indigo and rs6591195, rs9735063 from psichomics in green and rs56984820, a GWAS hit-SNP in red. **B -** Known *BANF1* isoforms. As in previous figure, well annotated isoforms are in gold while others are in red. Full boxes represent coding portion of exons and UTRs are hollowed. Introns are represented in different colours with special designation for introns of interest. Possible splicing patterns are represented at the top of panel B. Introns are not drawn to scale. **C** – sQTLs detected in this gene. rs56984820 is represented in subpanel 1 showing changes in spliced out intron D, rs6591195 in subpanel 2 regarding an alternative first exon and rs9735063 in subpanels 3 and 4 showing an alternative first exon and a skipped exon respectively. A and B), a protein coding gene encoded in the positive strand, between positions chr11:66,002,228 and chr11:66,004,149.

An sQTL, rs56984820, was identified with LeafCutter and specifically associated with changes in the splicing of intron D (chr11:66,003,070 to chr11:66,003,235) (Figure 4.19 – sQTL mapping identified splicing changes in BC risk associated locus 11q3. **A -** this region is located on the bigger arm of chromosome 11, on band 13.1. Near the bottom of the panel are represented variants that were identified as sQTLs, rs617791 from LeafCutter in indigo and rs6591195, rs9735063 from psichomics in green and rs56984820, a GWAS hit-SNP in red. **B -** Known *BANF1* isoforms. As in previous figure, well annotated isoforms are in gold while others are in red. Full boxes represent coding portion of exons and UTRs are hollowed. Introns are represented in different colours with special designation for introns of interest. Possible splicing patterns are represented at the top of panel B. Introns are not drawn to scale. **C** – sQTLs detected in this gene. rs56984820 is represented in subpanel 1 showing changes in spliced out intron D, rs6591195 in subpanel 2 regarding an alternative first exon and rs9735063 in subpanels 3 and 4 showing an alternative first exon and a skipped exon respectively. , panel C1). Deletion of the segment [CACTGAG] was correlated with a lower expression of isoform ENST00000533166 (effect size of -0.0103, FDR-corrected p-value of 1.52×10^{-7}), as detected by the decrease in PSI (calculated as the number of read counts of intron D divided by the sum of the read counts of introns A to E).

Another three sQTLs were identified with psichomics, two of which were alternative first exons (AFE) events, with the third relating to a skipped exon (SE) event. The first sQTL, rs6591195, related to an AFE in which the content of the minor T allele was associated with the

lower inclusion of intron A (chr11:66,002,570 to chr11:66,003,235) (effect size = -0.0318, q-value = 6.63×10^{-4}), as detected in a decrease in PSI calculated as the ratio between intron A read counts and the sum of intron A and D read counts (Figure 4.19 – sQTL mapping identified splicing changes in BC risk associated locus 11q3. **A -** this region is located on the bigger arm of chromosome 11, on band 13.1. Near the bottom of the panel are represented variants that were identified as sQTLs, rs617791 from LeafCutter in indigo and rs6591195, rs9735063 from psichomics in green and rs56984820, a GWAS hit-SNP in red. **B -** Known *BANF1* isoforms. As in previous figure, well annotated isoforms are in gold while others are in red. Full boxes represent coding portion of exons and UTRs are hollowed. Introns are represented in different colours with special designation for introns of interest. Possible splicing patterns are represented at the top of panel B. Introns are not drawn to scale. **C** – sQTLs detected in this gene. rs56984820 is represented in subpanel 1 showing changes in spliced out intron D, rs6591195 in subpanel 2 regarding an alternative first exon and rs9735063 in subpanels 3 and 4 showing an alternative first exon and a skipped exon respectively. , panel C2). The other two sQTLs correspond to the same variant, rs9735063, and related to an AFE and an SE. The AFE corresponded to the lesser usage of intron B (chr11:66,002,513 to chr11:66,003,235) associating with the increased number of rs9735063(C) minor alleles (effect size = -0.1427 per allele, multiple test corrected p-value = 1.13×10^{-4}), detected as a decrease in PSI (intron B read counts divided by the sum of B and D read counts). The SE event associated with rs9735063 refers to the exon located between introns D (chr11:66,002,513-66,002,844) and E (chr11:66,003,070-66,003,235), whose alternative splicing pattern is removal of intron B (chr11:66,002,513 to chr11:66,003,235). This splicing pattern is shown in Figure 4.19B comparing the first two exons of transcript ENST00000445560, where the exon is spliced out, and initial three exons of ENST0000053024, where it is spliced in. In this event the PSI is assessed as $\frac{(E+D)}{2} / (\frac{(E+D)}{2} + B)$. The effect size for this event is 0.0953 and p-value = 6.71×10^{-4} (Figure 4.19 panel C4).

Six isoforms of *BANF1*, out of eight, are expressed in breast tissue (Figure 4.20). The most expressed isoform is ENST00000312175, which contains the alternative intron A and for which rs6591195 is an sQTL. Isoform ENST00000445560 has a lower expression than the previous, corresponds to the splicing of intron B and has rs9735063 as an sQTL . Transcripts ENST00000530204 and ENST00000533166 both include intron D, but only the latter transcript is annotated as expressed in healthy breast tissue (Figure 4.20), to which corresponds the sQTLs rs56984820 and rs9735063.

The risk variant rs617791 has been identified in this region, with the minor C allele associated with an increase in breast cancer incidence (p-value = 7 x 10^{-7}, effect size = 0.0281) (Michailidou *et al.*, 2017). This risk variant is in high LD with all three sQTLs, namely rs56984820 (r^2 = 0.71, D' = 0.96), rs6591195 (r^2 = 0.48, D' = 0.93) and rs9735063 (r^2 = 0.51, D' = 0.97). This suggests that the risk C allele is associated with lower expression of isoforms ENST00000533166, ENST00000312175 and ENST00000445560.

Besides sQTLs, rs6591195 and rs9735063 are also eQTLs for *BANF1* in different tissues, including breast. GTEx data for rs6591195 reports an NES = 0.179 and p-value = 7.12×10^{-12} (Supplementary Figure 8), while for rs9735063 a greater NES of 0.197 is reported with a p-value = 1×10^{-13} (Supplementary Figure 9).

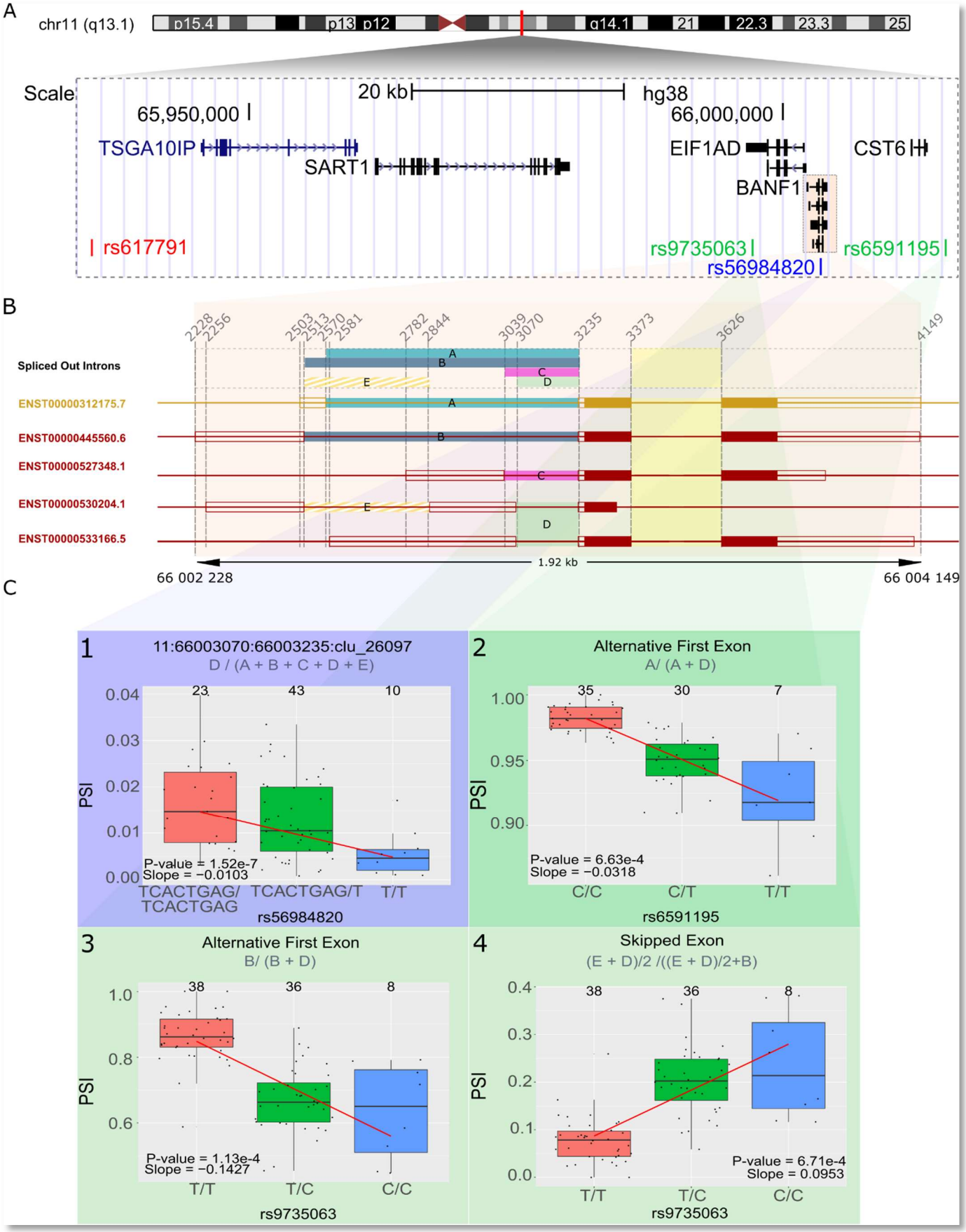

A
chr11 (q13.1)
p15.4
p13
p12
q14.1
21
22.3
23.3
25
Scale
20 kb
hg38
65,950,000
66,000,000
TSGA10IP
SART1
EIF1AD
CST6
BANF1
rs617791
rs9735063
rs6591195
rs56984820
B
Spliced Out Introns
ENST00000312175.7
ENST00000445560.6
ENST00000527348.1
ENST00000530204.1
ENST00000533166.5
1.92 kb
66 002 228
66 004 149
C
1
11:66003070:66003235:clu_26097
D / (A + B + C + D + E)
PSI
P-value = 1.52e-7
Slope = −0.0103
TCACTGAG/ TCACTGAG
TCACTGAG/T
T/T
rs56984820
2
Alternative First Exon
A/ (A + D)
P-value = 6.63e-4
Slope = −0.0318
C/C
C/T
T/T
rs6591195
3
Alternative First Exon
B/ (B + D)
P-value = 1.13e-4
Slope = −0.1427
T/T
T/C
C/C
rs9735063
4
Skipped Exon
(E + D)/2 /((E + D)/2+B)
P-value = 6.71e-4
Slope = 0.0953
T/T
T/C
C/C
rs9735063

Figure 4.19 – sQTL mapping identified splicing changes in BC risk associated locus 11q3. **A -** this region is located on the bigger arm of chromosome 11, on band 13.1. Near the bottom of the panel are represented variants that were identified as sQTLs, rs617791 from LeafCutter in indigo and rs6591195, rs9735063 from psichomics in green and rs56984820, a GWAS hit-SNP in red. **B -** Known *BANF1* isoforms. As in previous figure, well annotated isoforms are in gold while others are in red. Full boxes represent coding portion of exons and UTRs are hollowed. Introns are represented in different colours with special designation for introns of interest. Possible splicing patterns are represented at the top of panel B. Introns are not drawn to scale. **C** – sQTLs detected in this gene. rs56984820 is represented in subpanel 1 showing changes in spliced out intron D, rs6591195 in subpanel 2 regarding an alternative first exon and rs9735063 in subpanels 3 and 4 showing an alternative first exon and a skipped exon respectively.

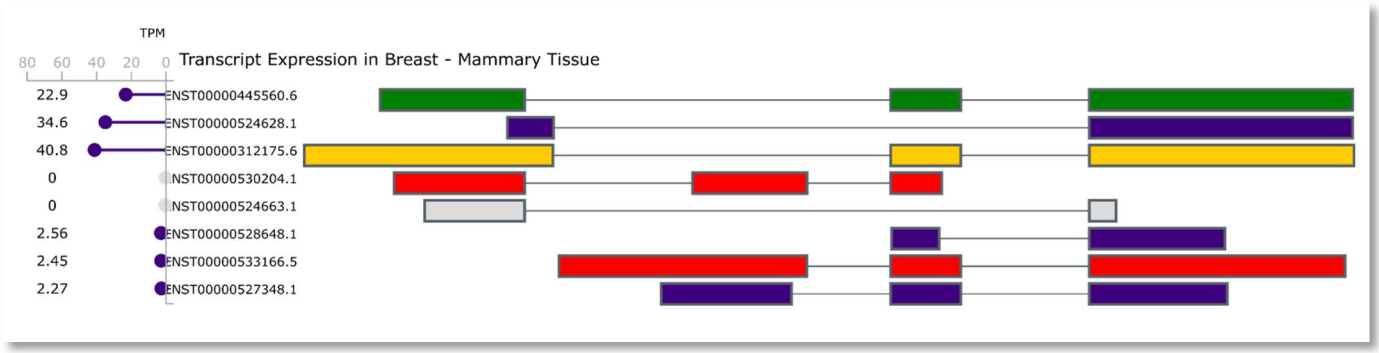


Figure 4.20 – *Banf1* isoform expression in breast tissue. In gold, green and red are isoforms whose splicing changes were associated with the presence of alternative alleles. The expression of each isoform in TPM is shown in the left side of the figure.

When inspecting the resulting proteins of the isoforms of interest (ENST0000012175, ENST00000445560 and ENST00000533166), I found that they differ only on their UTRs, sharing the entire coding sequence, hence producing identical proteins.

4.5.5 Risk locus 15q24

Two sQTLs were identified on chromosome 15, one per analysis tool, which associated with alternative splicing on the gene *ULK3* (Unc-51 like kinase 3) (Figure 4.21). *ULK3* is encoded in the reverse strand, between chr15:74,836,118 and chr15:74,843,346. Using psichomics, an alternative 5' splice site event was detected on the second to last intron, changing the intron boundary from chr15:74,837,751 to chr15:74,837,757. PSI for this event is calculated as the ratio of read counts of intron A (chr15:74,837,751-74,837,435) by the sum of read counts of introns A and B (chr15:74,837,757-74,837,435). The number of minor C alleles of rs12898397 was correlated with a decrease in PSI, with the corresponding linear regression showing the steepest slope (effect size = -0.5086) and the most significant association (FDR corrected p-value = 4.83×10^{-26}) (Figure 4.21C-1). This strong effect size led to the hypothesis that this variant could be disturbing a splice site, rendering it unrecognizable by spliceosome components. Using NetGene2, an *in silico* splice site predictor (Brunak, Engelbrecht and Knudsen, 1991; Hebsgaard *et al.*, 1996) I analysed the impact of each allele on splice site strength. Using a 100nt flaking sequence up and downstream of variant rs12898397 (Table 3). Using the reference rs12898397-T allele, NetGene2 predicted 2 potential donor splice sites on the complementary strand, a potential donor sequence [GCAGGTCAAG^gtgggcacat] is identified with a confidence of 0.91, along with [GAAGGAGCAG^gtcaaggtgg] with a lower confidence level of 0.71 (Supplementary Figure 10). The same analysis for the alternative allele rs12898397-C predicted only one donor splice site: [GAAGGAGCAG^gtcagggtgg], with a confidence level of 0.82 (Supplementary Figure 11). This suggests that rs12898397-C disrupts the canonical donor splice site recognised by spliceosome components. Loss of this cis-regulatory element (CRE) leads to a change in donor splice site from chr15:74,837,751 to a competing 5' splice site at chr15:74,837,757, leading to the skipping of 6bp in the processed mRNA.

The second sQTL, rs12591513, was detected using LeafCutter. The alternative splicing event is a decrease of usage of intron C, between chr15:74,839,050 and chr15:74,839,268, in a cluster composed of introns C and D. Intron D is an unannotated intron detected from raw RNA-seq, spanning from chr15:74,839,050 to chr15:74,839,204 (Figure 4.21B). sQTL mapping correlated the alternative allele content of rs12591513 with a PSI decrease (effect size = -0.0098, FDR corrected p-value = 4.23×10^{-2}), barely overcoming the threshold for significance of 5×10^{-2}.

According to GTEx, six annotated isoforms (out of 23) are significantly expressed in healthy breast tissue (Figure 4.22). From these, processing of ENST00000440863, ENST00000566479 and ENST00000567472 requires splicing out of intron A, while ENST00000561725 and ENST00000568718 are the only isoforms expressed whose splicing removed intron B. There is no isoform annotated with the presence of intron D.

The GWAS hit-SNP for this locus is rs6938, with the G allele associating with an increased risk of 0.0364 for breast cancer (p-value = 9 x 10^{-8}) (Michailidou *et al.*, 2017). The risk allele is correlated with the A allele of rs12591513 (r^2 = 0.57, D'=0.77), and with the C allele at rs12898397 (r^2 = 0.55, D' = 0.80). Thus, risk seems to be associated with a reduction of isoforms ENST00000440863, ENST00000566479 and/or ENST00000567472, and to a lesser extent with an increase in isoforms ENST00000561725, ENST00000569437 and/or ENST00000568718.

Similar to previous observations, both rs12591513 and rs12898397 are eQTLs for *ULK3* in breast tissue, as well as in several other tissues, with an estimated NES of -0.212 and -0.211 (p-value = 7.0×10^{-8} and 4.5×10^{-8}, respectively) (Supplementary Figure 12 & Supplementary Figure 13).

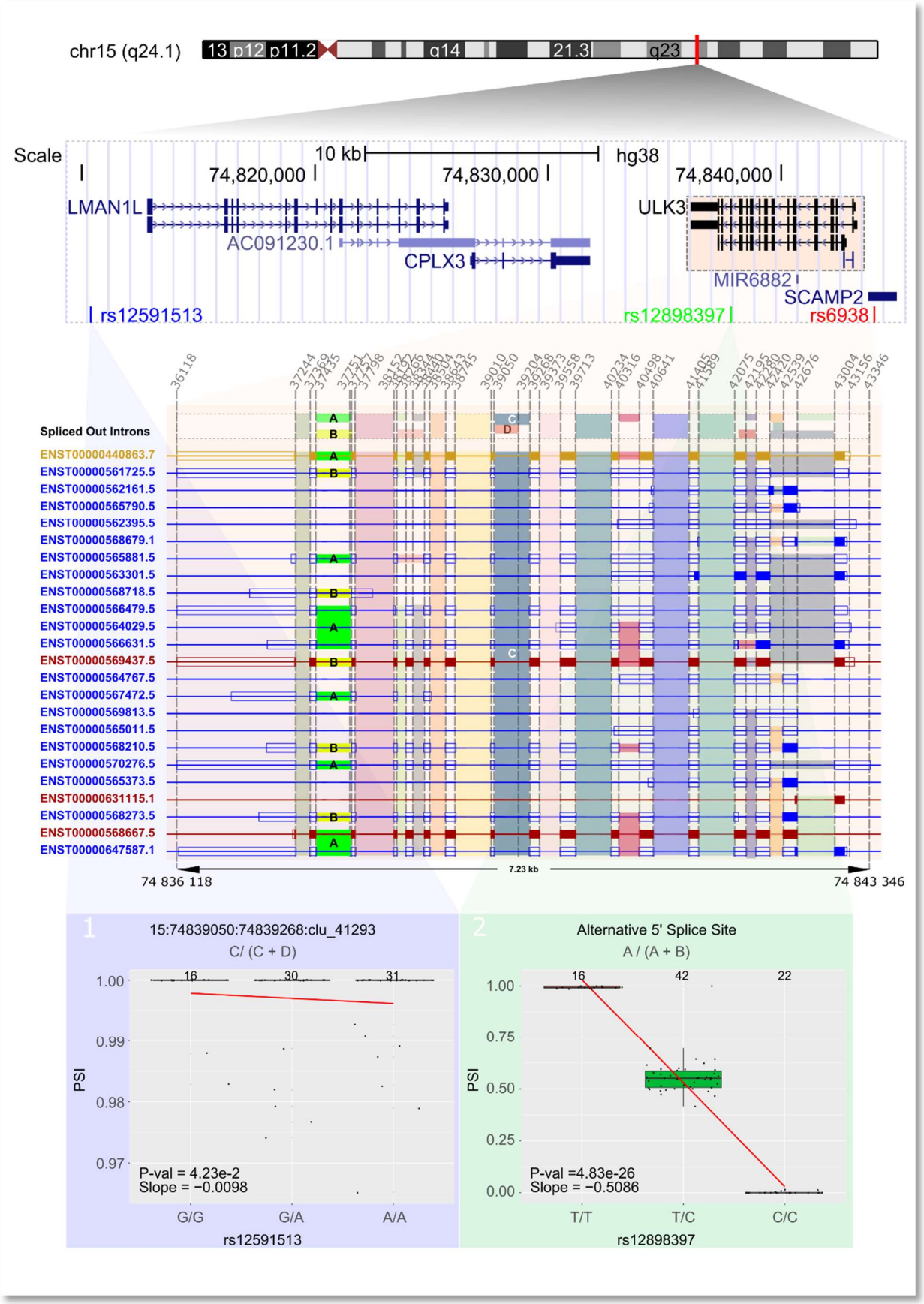

chr15 (q24.1)
Scale
10 kb
hg38
74,820,000
74,830,000
74,840,000
LMAN1L
AC091230.1
CPLX3
ULK3
MIR6882
SCAMP2
rs12591513
rs12898397
rs6938
Spliced Out Introns
ENST00000440863.7
ENST00000561725.5
ENST00000562161.5
ENST00000565790.5
ENST00000562395.5
ENST00000568679.1
ENST00000565881.5
ENST00000563301.5
ENST00000568718.5
ENST00000566479.5
ENST00000564029.5
ENST00000566631.5
ENST00000569437.5
ENST00000564767.5
ENST00000567472.5
ENST00000569813.5
ENST00000565011.5
ENST00000568210.5
ENST00000570276.5
ENST00000565373.5
ENST00000631115.1
ENST00000568273.5
ENST00000568667.5
ENST00000647587.1
7.23 kb
74 836 118
74 843 346
1
15:74839050:74839268:clu_41293
C/ (C + D)
PSI
P-val = 4.23e-2
Slope = −0.0098
G/G
G/A
A/A
rs12591513
2
Alternative 5' Splice Site
A / (A + B)
P-val =4.83e-26
Slope = −0.5086
T/T
T/C
C/C
rs12898397

Figure 4.21 – sQTL mapping of BC risk associated locus 15q24 identified changes in *ULK3*. **A -** Genomic locus encompassing *ULK3* is represented as well as surrounding genes. Relevant variants, rs12591513, rs12898397 and rs6938 are identified in blue, green and red, respectively. **B** - annotated transcripts are show in gold, red or blue for well annotated isoforms, partially annotated isoform and non-coding transcript, respectively. In between each exon are represented spliced-out introns with a summary of alternative splicing patterns at the top of panel B. Alternative splicing event identified by sQTL mapping are nominated for easier identification. Intron A and B are at the second to last intron, changing the five prime splice site by 6 bases and, as a consequence, reduces protein size by 2 amino acids. The other identified alternative splicing site are introns C and D in intron eight. In this case no annotated transcript is available showing intron D. **C** - sQTL for rs12591513 from LeafCutter and rs1289397 retrieved from psichomics showing changes in splicing patterns as measured by PSI.

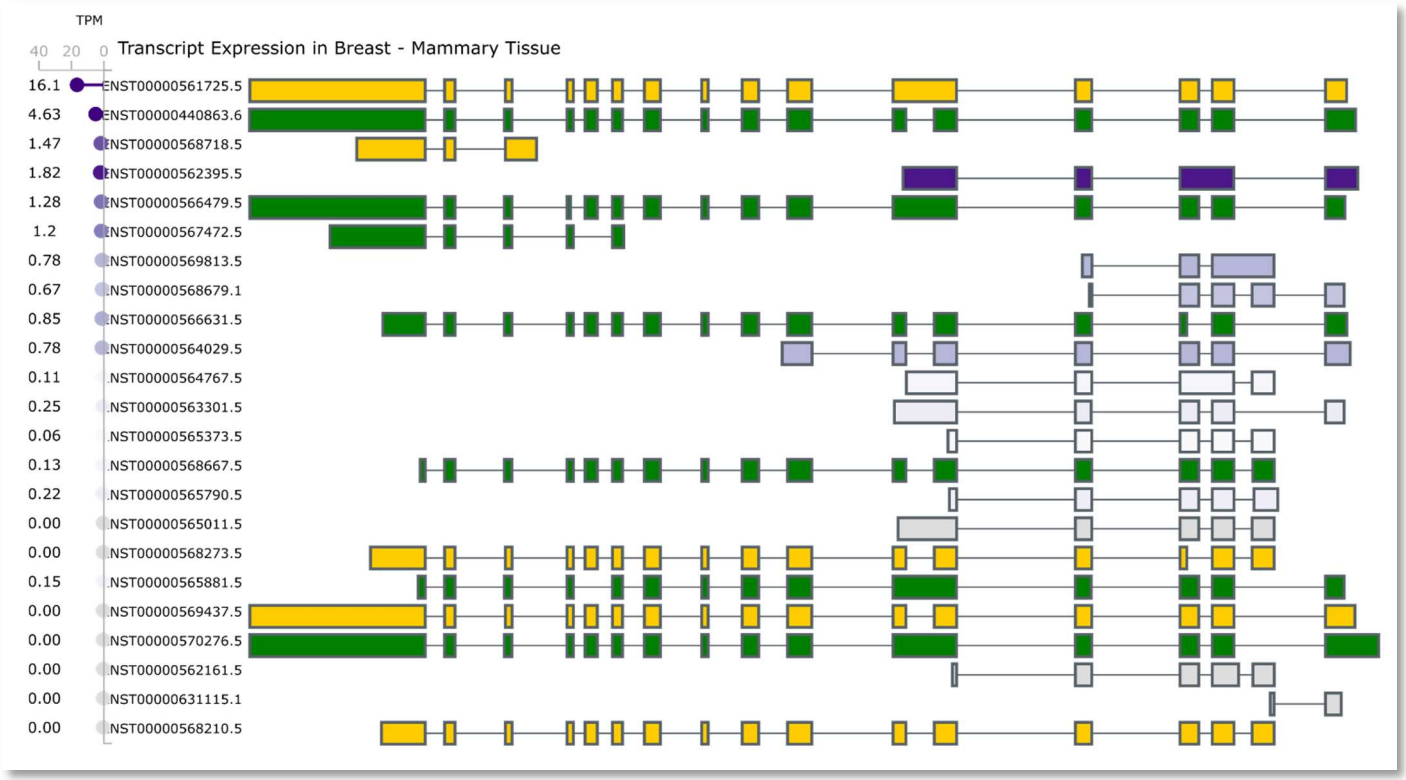


Figure 4.22 – *ULK3* isoform expression in breast tissue. In gold and green are isoforms where sQTLs were detected. The expression of each isoform in TPM is shown in the left side of the figure.

5 Discussion

In this study we performed splicing cis-QTL mapping using RNA-seq and genotyping data from female healthy breast tissue and integrated it with known risk loci for breast cancer, to assess the contribution to BC risk of cis-regulatory variation modulating alternative splicing.

5.1 Splicing Detection and Quantification

To reach my goals, I first quantified alternative splicing events across all samples, using two different computational tools.

Performing splicing quantification with LeafCutter, we observed that introns dimension ranged from 22 to 498,598 nucleotides with some genes spanning across multiple genomic loci. Large introns are known in the human genome – *ROBO2* gene includes an intron spanning 1,160,411 bp (Piovesan *et al.*, 2019) – but in order to reduce misaligned and multiple alignment of RNA-seq reads, a threshold of 500kb was set as maximum intron size artificially limiting the detected intron size. Although this can reduce the ability to detect alternative splicing events in larger introns, it allows a more reliable alignment of raw reads to the reference genome.

Clustering these introns based on splice sites showed that most alternative spliced introns in breast tissue had less than four different splicing patterns with rare clusters displaying up to 185 different intronic edges. On a closer look, the reads for this extreme cluster were aligned to the gene *MAN1B1*, known to have 16 different isoforms and a pseudogene on chromosome 11, which is known to hamper alignment (Raplee, Evsikov and De Evsikova, 2019).

It is interesting to observe that although 314,875 alternative splicing events were found with LeafCutter, passing through successive filters, only about a third (107,730) showed to have some level of variation between samples. This indicates that, although alternative splicing isoforms are available, this process is tightly regulated producing similar levels of transcripts across samples in the same tissue type (Wang *et al.*, 2008).

Psichomics detection of alternative splicing events is more skewed as pre-processed data with exon-exon junctions quantification from GTEx was used. Nonetheless, although less events were detected when compared to LeafCutter, most events detected showed some level of variation. This could be due many of the detected events did not pass removal of “Not Available” values

because of the denominator used to compute PSI being equal to zero. Given psichomics is dependent on previously annotated events it provides a clearer description on the event detected, as well as which are the alternatives. Nevertheless, this also reveals one of its caveats as novel splicing patterns are not identified.

Comparing ease of use for each tool, both were built with the purpose of performing differential analysis between groups. As such, some adaptations from well described documentation were made. As psichomics is implemented as an R package its use is simple and intuitive, providing an easy introduction to this type of analysis. Comparatively, LeafCutter's implementation required usage of bash and python languages besides requiring other bioinformatic tools, as STAR aligner.

By combining the outcome from psichomics and LeafCutter I expect to improve detection of alternative splicing events and its robustness. While LeafCutter's lower threshold for defining splicing clusters retrieves plenty of alternative splice sites, these appear to have low variability between samples. On the other hand, given only previously identified alternative splice sites are analysed by psichomics, most of the events it has information for present some level of variation.

One common aspect to the analysis using both tools, was the need to control for the different factors known to influence gene expression and splicing, such as age, quality of the RNA sample, sequencing technology, etc (Knight, 2004). Particularly, age could have skewed our results, as the distribution of age of donors ranged from 20 to 69. In terms of gene expression control in breast tissue, it would have been useful to be able to control for age at menarche, age at menopause, oral contraception use or HRT, but no such information was available (Supplementary Figure 14). In order to remove possible gender related confounding effects, only data regarding female samples was used (Aguet *et al.*, 2020b). Furthermore, RNA-seq was performed on bulk breast tissue RNA, retrieving an average from all cell present. Single-cell RNA-seq could further improve gene expression characterization by assessing differences between cell types (Aguet *et al.*, 2020a).

5.2 sQTL mapping

sQTL mapping detected associations between genotype and alternative splicing phenotypes in 1.3% of all assessed events, regardless of tool used to measure splicing. This suggests that common genetic variants have a relatively small contribution to the population variation observed for alternative splicing isoform expression.

Distance between best correlated variant and AS event showed that most sQTLs are within a small range of 25k from the regulated splice site, which is concordant with findings from other authors (Walker *et al.*, 2019). However, biologically it is expected that genetic variants acting upon alternative splicing are located closer to the splice sites, and this range may be enlarged in our analysis due to local LD patterns, as highly correlated variants (for example, $r^2 \geq 0.9$) will all exhibit association with the same splicing events. This is in fact a consequence of one of the features of tensorQTL, as only the best correlated variant is chosen for permutation analysis and beta modelling, leading to potential loss of multiple independent signals that influence the same alternative splicing event.

Looking at the intersection of sQTLs computed by both tools, only 50 loci were identified in common. This contrasts with the 222 common sGenes discovered. A possible explanation is that it is a consequence of numerous loci presenting more than one gene, with some overlapping or contiguous genes. Nonetheless some sQTLs identified the same event, as seen for gene *PARK7*. Comparing the common events detected, PSI levels computed with LeafCutter displayed a generally lower value. As a consequence, the effect size for each sQTL is smaller using the PSI from LeafCutter, when compared to the ones retrieved using psichomics.

It is important to remind that it is highly unlikely that the detected sQTLs are the causal variants for the changes in splicing observed, due to the LD effects discussed above. In order to detect the causal variants further *in silico* and *in vitro* studies are required: firstly, identify all variants in high LD with the sQTL within each locus, then characterize how they can modify splicing CRE and lastly determine its impact on splicing mechanism in the tissue of interest. Several databases of well annotated cis-regulatory elements and the impact from variants based on *in vivo* and *in vitro* experiments are available, such as ENCODE, RegulomeDB, HaploReg, but unfortunately for this work none includes RBP-binding essays performed in breast tissue (Boyle *et al.*, 2012; Ward and Kellis, 2012; Rojano *et al.*, 2019; Moore *et al.*, 2020; Van Nostrand *et al.*, 2020; Zhang *et al.*, 2020).

5.3 GWAS hit-SNPs

As one of the main aims of this work was to identify sQTLs potentially associated with risk to breast cancer, I looked for significant risk-associated loci in GWAS Catalog. From available studies on this database, most (41 out of 56) were performed on individuals of European ancestry,

creating a bias towards common variants with impact on this population (Mills and Rahal, 2019; Gay *et al.*, 2020). Hence, once GWAS are performed on other ancestries, new variants with populations-specific effects will be identified, providing new insights onto breast cancer genetic risk and aetiology, and works like the one presented here will have to be carried to extend our findings to such populations.

While 701 associations were found between variants and risk to breast cancer (any of the different phenotypes of risk studied), the same variant-phenotype association can be reported across multiple studies, reducing unique variant-phenotype pairs to 475. The most frequently identified variant was rs3803662, identified as significantly associated to breast cancer susceptibility in nine out of ten studies. This improves confidence in the results as the same variant is identified across multiple studies using different samples and methodology.

As anticipated, most hit-SNPs retrieved were characterized as intronic or intergenic, reinforcing our hypothesis that most variants identified in GWAS could impact cis-regulatory elements modulating gene expression. Risk increase evaluating metrics, either as OR, effect size or otherwise, of the variants included in this study were within literature expected values.

It is important to point out a trend of increased sample size on more recent GWAS (Supplementary Figure 15). Given biobanks collect and aggregate genotype and phenotype data from multiple samples, it reduces the costs and time allowing the inclusion of a bigger number of individuals, effectively increasing statistical power. One important gain is the increased capability to discover new hit-SNPs with smaller effect, whose biological influence is exerted over complex networks. To illustrate this point, a single GWAS meta-analysis performed was able to obtain information on nearly 150k individuals leveraging data from previous studies, identifying 803 loci associated with BC risk (Michailidou *et al.*, 2017).

As of the writing this manuscript, new GWAS were published using individuals of Sub-Saharan ancestry (Wang *et al.*, 2019) or employing pan-cancer GWAS methods to report new variants associated with cancer risk (Rashkin *et al.*, 2020). In the future, these should be included in a similar analysis as the one I present.

5.4 Risk loci co-localization with sQTLs

One of the caveats of this study is the co-localization method employed to assess the intersection between BC GWAS SNPs and sQTL signals. By relying solely on LD to cross sQTLs

and hit-SNPs from previous GWAS, I may be identifying co-localizations that are somehow fortuitous and a product of the genetic closeness which is not necessarily functional. Although co-localization is required to determine if a variant can be affecting BC susceptibility and splice pattern changes, it is not sufficient to prove it. Other methods as fine-mapping, mendelian randomization or more complex usage of association statistics and LD patterns can go further in providing the causal variant, reducing the number of causal SNPs for *in vitro* testing (Hormozdiari *et al.*, 2016; Schaid, Chen and Larson, 2018; Porcu *et al.*, 2019; Gay *et al.*, 2020; van der Graaf *et al.*, 2020). Moreover, here I only explored the common sQTLs, leaving out other loci of interest where sQTLs were only detected by one splicing measuring tool. Changes in splicing of known oncogenes or TSG, as *MUTYH* involved in DNA repair and whose germline mutations are associated with syndromic Familial Adenomatous Polyposis disease, were also detected using one of the two PSI computing procedures (Chakravarty *et al.*, 2017). Exploring these may return further insights.

5.5 Risk locus 1p36

The *PARK7* gene, also known as *DJ-1*, *DJ1* or *GATDF2*, encodes for a deglycase protein, part of peptidase C56 family. It has multiple functions pointing towards the role of oncogene: oxidative stress sensor protecting cells against oxidative stress and cell death, positive regulator of gene transcription, regulator of protein folding. Its expression is also known to be elevated in multiple tumours types (Ismail *et al.*, 2014; Zheng *et al.*, 2019).

Mutations and variants in this gene have been mostly associated with early onset Parkinson's Disease (PD), but also with different forms of cancer (Bonifati *et al.*, 2003; Nuytemans *et al.*, 2010). Although there was no colocalization between GWAS hit-SNPs for PD and BC, as of 2013, a relationship between PD and BC incidence, both in hereditary and sporadic forms, was found as most genes associated with PD are involved in cell cycle, with described functions as oncogenes or tumour suppressor genes (Kravitz *et al.*, 2013; Biosa *et al.*, 2017).

The translation of *PARK7*'s main isoform, ENST00000338639, results in a protein with 189 amino acids (aa), which forms a globular homodimer with a molecular weight of 20 kDa. Each monomer has three cysteine residues responsible for its catalytic activity: C46, C53 and C106. C106 is highly conserved amongst mammals and can be oxidated, rendering it as an oxidative stress sensor. Although C46 and C53 are mainly responsible for the interaction between monomers,

they can also participate in redox reactions (Biosa *et al.*, 2017). Although its location is mainly cytoplasmatic, a fraction can be found in the nucleus, as well as in the mitochondria where it acts as a respiratory chain stabilizer (Ejma *et al.*, 2020).

In the presence of oxidizing conditions, the protein Park7 interacts with several other proteins in different pathways: interaction with PTEN supresses its phosphatase activity allowing PI3K/Akt increased signalling for cellular survival (Kim *et al.*, 2005; Ismail *et al.*, 2014). Depending on the oxidative state of C106, Park7 can interact with p53 masking its DNA binding domain, hence preventing gene expression that would induce cell cycle arrest and apoptosis. Park7 also promotes ERk1/2 translocation to the nucleus, phosphorylating Elk and activating gene transcription of several target genes, including SOD. This interaction seems to be mediated by a domain near the aa 166, as mutations here prevent Elk activation. Park7 is still able to displace Keap1, allowing Nrf2 pathway activation and expression of proteins with antioxidant activity through antioxidant response element. Finally, Park7 sequesters Daxx in the nucleus preventing its interaction with Ask1 and inhibiting cell apoptosis (Ismail *et al.*, 2014; Biosa *et al.*, 2017; Schwab, 2017). This pleiotropy results in increases resistance to apoptosis and stimulates EMT enhancing invasion.

Park7 is known to provide response to oxidative stress either directly, by enhancing SOD1 enzymatic activity with ROS, or indirectly by increasing expression of proteins with antioxidant activity. It operates as a cooper chaperon for SOD1 where aa residues C106 and E16 are involved in metal coordination while geometry binding is dependent on G75 and H126, of which G75 is absent on ENST0000377493. It is also known that Park7increases SOD1 gene expression via MAPK pathway and Nqo1 or Trx1 via Nrf2, possible involved in longer-term response to stimuli (Biosa *et al.*, 2017).

Park7 also acts as a histone deglycase whose overexpression in tumours can originate epigenetic dysregulation in cancer cells. Glyoxals, as methylglyoxal (MGO), are toxic glycolytic by-products elevated in pathological of mitochondrial disfunction settings as diabetes, cardiovascular disease and cancer – Warburg effect. MGO can non-enzymatically glycolate nucleotides, in particular guanines, and proteins as histones. These glycations form adducts and replace other posttranslational modifications (methylation, acetylation and ubiquitination). At higher concentrations, histones lateral chain near DNA can also be glycated as well as nucleotides forming crosslinks. These changes disrupt nucleosome stability and unwrap DNA changing gene

expression (Trempe and Fon, 2013; Galligan *et al.*, 2018; Zheng *et al.*, 2019). Moreover, glycated proteins can be secreted to extra-cellular space and activate the receptor for advance glycation end-products and activate local inflammatory response. Park7 is also secreted and is a substrate for Matrix MetalloProtease (MMP) suggesting an additional extracellular role. (Ismail *et al.*, 2014; Biosa *et al.*, 2017).

As most studies focus on changes in gene expression, but not on splicing, it is not clear how the latter contributes to these phenotypes (Nuytemans *et al.*, 2010; Trempe and Fon, 2013). The haplotype associated with a higher risk of BC is also associated with changes in splicing increasing the production of the 169 aa isoform ENST00000377493. This alters the secondary structure of the protein, losing a β-strand and an α-helix, with unknown effect on any of its function as biological pathway modifier. From the literature, only stabilization of cooper binding site involves G75, part of the removed portion of the protein (P65 to E84). Furthermore, although the coding sequence is identical between ENST00000338639 and ENST00000493678, their 5' and 3'UTRs are not, possibly impacting other post-processing regulation mechanisms, as mRNA stability and/or translation efficiency, which is hinted by the dual classification of SNPs as eQTL and sQTL. Furthermore, mRNA localization and editing can also be impacted. Changes in any of Park7 functions has the potential to deregulate gene expression through epigenetic modifications, increase mutagenesis via decreased ROS response, enhance cell resistance to apoptosis via PI3K/Akt or MAPK pathways or persistent local inflammation, all of which are known hallmarks of cancer (Hanahan and Weinberg, 2000, 2011).

5.6 Risk locus 11q13

In locus 11q13 changes in the splicing pattern *of BANF1* were identified. Banf1, Barrier to Autointegration Factor 1 or BAF, is a DNA binding protein involved in genome damage repair by regulating PARP1 activity after oxidative damage or single strand brake.

BANF1 expression is increased in TNBC when compared to normal tissues and is positively correlated with the number of affected lymph nodes and TNM staging. Moreover its expression was directly correlated with *MKI67* and *MTA1*, biomarkers of proliferation and metastasis (Zhang, 2020). Analysis in other cancer types showed similar results proposing *BANF1* expression as a prognosis biomarker (J. Li *et al.*, 2018; Sears and Roux, 2020).

Translation of any of the identified isoforms produces a protein with 10kDa (89 residues) that forms an homodimer. It distributes through the nucleus and, in a lesser extension, the cytosol. Phosphorylation at T2,T3 or S4 regulates function and sub-cellular location, increasing cytoplasmatic fraction, inhibiting DNA binding and dimerization when phosphorylated (Jamin, Wicklund and Wiebe, 2014). The DNA binding domain is composed of a helix-turn-helix, composed of aa 20 to 35. During mitosis Banf1 is associated with the chromosomes, binding in an unspecific manner with double stranded DNA. Banf1 can also interact with other proteins containing LAP2/emerin/MAN1(LEM) domains as histones, laminins, transcription factors and DNA damage response proteins as PARP1 (Montes de Oca *et al.*, 2009). This allows the interaction with the inner nuclear membrane proteins, facilitating nuclear reassembly after mitosis and chromatin organization in the periphery of nuclear envelopment. This generates lamina associated domains where genes are either not expressed or expressed at very low quantities (van Steensel and Belmont, 2017). After oxidative damage or single-strand breaks, Banf1 leaves the nuclear periphery to assemble onto damaged chromatin, regulating PARP1 activity towards genome repair. Furthermore, overexpression of *BANF1* is shown to reduce PARP1 activity by interacting with PARP1's NAD^+ binding domain thus preventing catalytic activity inhibiting Base Excision Repair pathway (Bolderson *et al.*, 2019).

Defects or decrease expression of Banf1 lead to errors during chromosome segregation, nuclear envelop reassembly, mis-localization of LEM proteins, as well as embryonic lethality in *C. elegans* and *D. melanogaster* (Jamin, Wicklund and Wiebe, 2014). The mutation A12T is a known cause of severe progeria, characterized by premature aging and genomic instability, a characteristic of tumours enabling the increased rate of mutation accumulation.

Although no clear change in coding sequence is visible between annotated isoforms expressed in breast cancer, changes in UTRs are known to modify overall protein levels as hinted by dual eQTL and sQTL annotation of identified SNPs. An hypothesis is that altered expression levels of *BANF1* results in chromosomic instability or inefficient damage repair. Over-time increased accumulation of mutations enhance oncogenic processes.

5.7 Risk Locus 15q24

In locus 15q24, *ULK3* was identified as an sGene. It is also known as Unc-51 Like Kinase 3 or Serin/Threonine-Protein kinase, a kinase similar to Fused (FU/STK36) that acts as a regulator

of Hedgehog (Hh) signalling and autophagy. Its main isoform, ENST00000440863, is translated into a 472aa-long protein with a molecular mass of 53kDa. Annotation provided by Uniprot regarding the region ranging from aa 14 to 270, indicates this is a kinase domain, with particular reference to position 44 responsible for ATP binding. Microtubule interacting and transport domains (MIT) are also annotated between aa positions 280 to 348 and 375 and 444 (*ULK3 - Serine/threonine-protein kinase ULK3 - Homo sapiens (Human) - ULK3 gene & protein*, 2020).

Although generally regarded as a positive regulator, in the absence of Hh ligand Ulk3 interacts with SUFU, suppressor of Fused, inhibiting phosphorylation of glioma-associated factor Gli. This triggers cleavage of Gli into a transcriptional repressor, functionally inhibiting Gli target genes transcription (Maloverjan, Piirsoo, Kasak, *et al.*, 2010; Maloverjan, Piirsoo, Michelson, *et al.*, 2010). However, in the presence of Hh ligand it dissociates from SUFU, autophosphorylates and mediates Gli phosphorylation – in particular Gli2 but also Gli1 and Gli3 – activating them and promoting its nuclear translocation and target gene transcription (Goruppi *et al.*, 2017). Excessive Hedgehog pathway activation leads to carcinogenesis via upregulation of N-myc, Cyclins D and E and FOXM (Katoh and Katoh, 2009). Furthermore, Ulk3 can induce autophagy when cells are in cell cycle arrest contributing to homeostasis during senescence (Young *et al.*, 2009). Changes in Hh signalling has been associated with several types of cancer as activation of this pathway leads to increased mitotic rate (McMahon, Ingham and Tabin, 2003). Additionally, higher levels of ULK3 were detected in cancer associated fibroblast (CAF) than on normal fibroblast providing autophagy derived products. CAFs are part of local tumour micro-environment and contribute to tumour invasion and metastasis besides increasing resistance to drugs. They are recruited by tumoral cells through paracrine signalling of Hh ligands (Goruppi *et al.*, 2017).

rs12898397-C is associated with an increase isoform ENST00000569437 and decrease of ENST00000440863. Similar changes in splicing pattern of *ULK3* as consequence of this variant have been previously reported (Zhao *et al.*, 2013; Mazin *et al.*, 2018). This isoform has a similar coding sequence to ENST00000440863, with the exception of two aminoacids in positions 444V and 445K. My hypothesis is that rs12898397-C modifies the canonical splice site consensus sequence, by replacing the second intronic nucleotide T with a C, increasing the usage of an alternative/cryptic splice site 6bp upstream and producing a shorter transcript and protein. Looking at the annotated secondary structure and domains of Ulk3, these changes occur at an α-helix at the edge of the second MIT and may alter protein regulation, subcellular distribution and/or function.

Moreover, there is experimental evidence that deletion of the c-terminal domain of Ulk3 produces a permanent catalytic active protein, indicating that this region is required for Ulk3 kinase regulation (Maloverjan, Piirsoo, Kasak, *et al.*, 2010)

Although LeafCutter identified changes in the splicing pattern of *ULK3*, there is no known transcript whose splice pattern follows intron edges as detected (chr15:74,839,050-74,839,204). This brings focus onto one of LeafCutter's advantages, being able to identify novel splice patterns from RNA-seq data. Looking closer at the data, this intron was detected in 14 samples in very low amounts, ranging from 0 to 3 reads per sample, compared to a mean of 90 reads per sample for the canonical intron chr15:74839050-74839268. This can be a technical error derived from misalignment to the reference genome, or a truly novel splice pattern forming a very lowly expressed transcript. Further analysis is required to fully characterize this discovery.

Changes in *ULK3* splicing can contribute to enhanced CAF transformation providing optimal local conditions for tumorigenesis facilitating progression and thus increasing BC risk. This hypothesis points towards an ever more relevant role from the micro-environment on tumorigenesis as evident by recent publications (Quail and Dannenberg, 2019; Roma-Rodrigues *et al.*, 2019; Clara *et al.*, 2020). Single-cell genomic and transcriptomic analysis have the potential to shed some light into these mechanisms revealing novel actionable targets for new therapies. Finally, this protein is a kinase, a prime target for pharmacological action using small molecules.

6 Conclusion

The aim of this project was to identify variants that increased breast cancer susceptibility by modulating splicing CREs. In order to do so, I started by quantifying splicing usage using RNA-seq data from female breast samples using two different analysis tools, psichomics and LeafCutter. Although both approaches detected a vastly different amount of alternatively spliced introns, only about a third of all alternatively spliced introns detected by either tool showed some degree of variation amongst samples on breast tissue. Performing sQTL mapping, only a small fraction of variation of all splicing event was attributable to common variants reiterating how robust the splicing mechanism is. With the increasingly larger sequencing read length that high throughput equipment is capable, sequencing of full-length mRNA will become more widespread allowing researchers to better characterize not only gene expression but also each individual isoform, as their function can vary widely.

Retrieving GWAS hit-SNPs confirmed the trend in increasing power to detect variants associated with changes in phenotype through the inclusion of more and more individuals, with some recent BC risk GWAS surpassing 200k samples. It is now important to move beyond GWAS, not only pinpointing causal variant(s) and the target gene(s), but also defining the molecular mechanism involved. This will help develop better biomarkers to screen for disease, perform more accurate diagnostic and monitor therapy impact, as well as provide new actionable protein targets for medicinal chemistry.

Co-localization of BC risk GWAS hit-SNPs with sQTLs identified in this work revealed three loci where variants influencing splicing patterns can also be modulating the risk for BC. These changes in splicing are in *PARK7*, *BANF1* and *ULK3*, genes previously mentioned in cancer literature as promoting cell proliferation, evading growth suppression, resisting cell death and enhancing genomic instability. Exploring the meaning of these changes is challenging as most research focused on the effect of overall expression level of the gene, and not on specific isoforms. The risk alleles previously associated with breast cancer risk signals are also associated with splicing pattern changes detected on *PARK7* and *ULK3,* which affect protein coding sequencing. Moreover, even if protein coding sequence remains unaltered, modification of UTRs is known to modulate RNA stability, localization, editing and translation, having a compound impact through other mechanisms.

6 Conclusion

With this work I demonstrated that variants associated with breast cancer susceptibility are also associated with other linked to changes in alternative splicing. With limited statistical power, variants that modulate BC risk via alternative splicing regulation appear to be a minor mechanism in disease predisposition. Further analysis of sQTLs by characterization of impacted splicing cis-regulatory elements is warranted to identify regulatory variants. Better co-localization techniques in addition to *in vitro* and *in vivo* methods will ascertain whether the changes in splicing are responsible for the increase in BC risk as detected by GWAS hit-SNPs or if they are independent of each other.

7 References

Adams, S. *et al.* (2019) 'Current Landscape of Immunotherapy in Breast Cancer: A Review', *JAMA Oncology*, 5(8), pp. 1205–1214. doi: 10.1001/jamaoncol.2018.7147.

Agostinho, N. (2018) *alternativeSplicingEvents.hg38: Alternative splicing event annotation for Human (hg38))*. Available at: https://www.bioconductor.org/packages/release/data/annotation/html/alternativeSplicingEvents.hg38.html (Accessed: 7 October 2020).

Aguet, F. *et al.* (2020a) 'Cell type–specific genetic regulation of gene expression across human tissues', *Science*, 369(6509). doi: 10.1126/SCIENCE.AAZ8528.

Aguet, F. *et al.* (2020b) 'The impact of sex on gene expression across human tissues', *Science*, 369(6509). doi: 10.1126/SCIENCE.ABA3066.

Alberts, B. *et al.* (2017) *Molecular Biology of the Cell*. doi: 10.1201/9781315735368.

Altman, N. and Krzywinski, M. (2015) 'Points of Significance: Association, correlation and causation', *Nature Methods*. Nature Publishing Group, 12(10), pp. 899–900. doi: 10.1038/nmeth.3587.

Altshuler, D., Daly, M. J. and Lander, E. S. (2008) 'Genetic mapping in human disease', *Science*, 322(5903), pp. 881–888. doi: 10.1126/science.1156409.

Altshuler, D. L. *et al.* (2010) 'A map of human genome variation from population-scale sequencing', *Nature*, 467(7319), pp. 1061–1073. doi: 10.1038/nature09534.

Altshuler, D. M. *et al.* (2012) 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, 491(7422), pp. 56–65. doi: 10.1038/nature11632.

Andersson, R. and Sandelin, A. (2020) 'Determinants of enhancer and promoter activities of regulatory elements', *Nature Reviews Genetics*. Springer US, 21(2), pp. 71–87. doi: 10.1038/s41576-019-0173-8.

Ardlie, K. G. *et al.* (2015) 'The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans', *Science*, 348(6235), pp. 648–660. doi: 10.1126/science.1262110.

Atkinson, T. J. and Halfon, M. S. (2014) 'Regulation of gene expression in the genomic context', *Computational and Structural Biotechnology Journal*. Research Network of

Computational and Structural Biotechnology, p. e201401001. doi: 10.5936/csbj.201401001.

Audano, P. A. *et al.* (2019) 'Characterizing the Major Structural Variant Alleles of the Human Genome', *Cell*, 176(3), pp. 663-675.e19. doi: 10.1016/j.cell.2018.12.019.

Auton, A. *et al.* (2015) 'A global reference for human genetic variation', *Nature*, 526(7571), pp. 68–74. doi: 10.1038/nature15393.

Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data (no date). Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed: 23 September 2020).

Ban, K. A. and Godellas, C. V. (2014) 'Epidemiology of Breast Cancer', *Surgical Oncology Clinics of North America*. Elsevier Inc, 23(3), pp. 409–422. doi: 10.1016/j.soc.2014.03.011.

Barash, Y. *et al.* (2010) 'Deciphering the splicing code', *Nature*, 465(7294), pp. 53–59. doi: 10.1038/nature09000.

Bell, J. T. *et al.* (2011) 'DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines', *Genome Biology*, 12(6), p. 405. doi: 10.1186/gb-2011-12-6-405.

Bellanger, M. *et al.* (2020) 'Cost-Effectiveness of Lifestyle-Related Interventions for the Primary Prevention of Breast Cancer: A Rapid Review', *Frontiers in Medicine*, 6(February), pp. 1–10. doi: 10.3389/fmed.2019.00325.

Belmont, J. W. *et al.* (2005) 'A haplotype map of the human genome', *Nature*, 437(7063), pp. 1299–1320. doi: 10.1038/nature04226.

Biosa, A. *et al.* (2017) 'Recent findings on the physiological function of DJ-1: Beyond Parkinson's disease', *Neurobiology of Disease*. Elsevier Inc., 108, pp. 65–72. doi: 10.1016/j.nbd.2017.08.005.

Biswas, S., Storey, J. D. and Akey, J. M. (2008) 'Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis', *BMC Bioinformatics*. BioMed Central, 9(1), p. 244. doi: 10.1186/1471-2105-9-244.

Bolderson, E. *et al.* (2019) 'Barrier-to-autointegration factor 1 (Banf1) regulates poly [ADP-ribose] polymerase 1 (PARP1) activity following oxidative DNA damage', *Nature Communications*. Nature Research, 10(1), pp. 1–12. doi: 10.1038/s41467-019-13167-5.

Bonifati, V. *et al.* (2003) 'DJ-1 (PARK7), a novel gene for autosomal recessive, early onset parkinsonism', *Neurological Sciences*, 24(3), pp. 159–160. doi: 10.1007/s10072-003-0108-0.

Borrego-Soto, G., Ortiz-López, R. and Rojas-Martínez, A. (2015) 'Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer', *Genetics and Molecular Biology*, 38(4), pp. 420–432. doi: 10.1590/S1415-475738420150019.

Boyle, A. P. *et al.* (2012) 'Annotation of functional variation in personal genomes using RegulomeDB', *Genome Research*. Cold Spring Harbor Laboratory Press, 22(9), pp. 1790–1797. doi: 10.1101/gr.137323.112.

Bray, F; Ferlay, J; Soerjomataram, I; Siegel, RL; Torre, LA; Jemal, A. (2018) *World*.

Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) 'Prediction of human mRNA donor and acceptor sites from the DNA sequence', *Journal of Molecular Biology*, 220(1), pp. 49–65. doi: 10.1016/0022-2836(91)90380-O.

Buniello, A. *et al.* (2019) 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D1005–D1012. doi: 10.1093/nar/gky1120.

Bush, W. S. and Moore, J. H. (2012) 'Chapter 11: Genome-Wide Association Studies', *PLoS Computational Biology*. Edited by F. Lewitter and M. Kann, 8(12), p. e1002822. doi: 10.1371/journal.pcbi.1002822.

Bycroft, C. *et al.* (2018) 'The UK Biobank resource with deep phenotyping and genomic data', *Nature*, 562(7726), pp. 203–209. doi: 10.1038/s41586-018-0579-z.

Calabrese, C. *et al.* (2020) 'Genomic basis for RNA alterations in cancer', *Nature*, 578(7793), pp. 129–136. doi: 10.1038/s41586-020-1970-0.

Chaidarun, S. S. and Alexander, J. M. (1998) 'A Tumor-Specific Truncated Estrogen Receptor Splice Variant Enhances Estrogen-Stimulated Gene Expression', *Molecular Endocrinology*, 12(9), pp. 1355–1366. doi: 10.1210/mend.12.9.0170.

Chakravarty, D. *et al.* (2017) 'OncoKB: A Precision Oncology Knowledge Base', *JCO Precision Oncology*. American Society of Clinical Oncology (ASCO), (1), pp. 1–16. doi: 10.1200/po.17.00011.

Cheang, M. C. U. *et al.* (2015) 'Defining Breast Cancer Intrinsic Subtypes by Quantitative Receptor Expression', *The Oncologist*, 20(5), pp. 474–482. doi: 10.1634/theoncologist.2014-0372.

Chen, T. and Dent, S. Y. R. (2014) 'Chromatin modifiers and remodellers: Regulators of cellular differentiation', *Nature Reviews Genetics*. Nature Publishing Group, 15(2), pp. 93–106. doi: 10.1038/nrg3607.

Chen, Z. *et al.* (2017) 'Epigenetic Regulation: A New Frontier for Biomedical Engineers', *Annual Review of Biomedical Engineering*, 19(1), pp. 195–219. doi: 10.1146/annurev-bioeng-071516-044720.

Clara, J. A. *et al.* (2020) 'Targeting signalling pathways and the immune microenvironment of cancer stem cells — a clinical update', *Nature Reviews Clinical Oncology*. Springer US, 17(4), pp. 204–232. doi: 10.1038/s41571-019-0293-2.

Clark, S. L. *et al.* (2012) 'Structure-function of the tumor suppressor BRCA1', *Computational and Structural Biotechnology Journal*, 1(1), p. e201204005. doi: 10.5936/csbj.201204005.

Claussnitzer, M. *et al.* (2015) 'FTO obesity variant circuitry and adipocyte browning in humans', *New England Journal of Medicine*, 373(10), pp. 895–907. doi: 10.1056/NEJMoa1502214.

Claussnitzer, M. *et al.* (2020) 'A brief history of human disease genetics', *Nature 2020 577:7789*, 577(7789), pp. 179–189. doi: 10.1038/s41586-019-1879-7.

Collins, F. S., Morgan, M. and Patrinos, A. (2003) 'The Human Genome Project: Lessons from large-scale biology', *Science*, 300(5617), pp. 286–290. doi: 10.1126/science.1084564.

Conesa, A. *et al.* (2016) 'A survey of best practices for RNA-seq data analysis', *Genome Biology*. BioMed Central Ltd., pp. 1–19. doi: 10.1186/s13059-016-0881-8.

Cookson, W. *et al.* (2009) 'Mapping complex disease traits with global gene expression', *Nature Reviews Genetics*, 10(3), pp. 184–194. doi: 10.1038/nrg2537.

Delaneau, O. *et al.* (2017) 'A complete tool set for molecular QTL discovery and analysis', *Nature Communications*. Nature Publishing Group, 8(1), pp. 1–7. doi: 10.1038/ncomms15452.

Demirdjian, L. and Xing, Y. (2020) *PAIRADISE*. doi: https://bioconductor.org/packages/release/bioc/html/PAIRADISE.html.

Desmet, F.-O. *et al.* (2009) 'Human Splicing Finder: an online bioinformatics tool to predict splicing signals', *Nucleic Acids Research*, 37(9), pp. e67–e67. doi: 10.1093/nar/gkp215.

Dobin, A. and Gingeras, T. R. (2015) 'Mapping RNA-seq Reads with STAR', *Current Protocols in Bioinformatics*. John Wiley and Sons Inc., 51(1), pp. 11.14.1-11.14.19. doi: 10.1002/0471250953.bi1114s51.

Dow, J. (2003) 'The Drosophila genome: the way forward.', *Briefings in Functional Genomics and Proteomics*. Oxford Academic, 2(2), pp. 121–127. doi: 10.1093/bfgp.

Dunning, A. M. *et al.* (2009) 'Association of ESR1 gene tagging SNPs with breast cancer risk', *Human Molecular Genetics*, 18(6), pp. 1131–1139. doi: 10.1093/hmg/ddn429.

Dvinge, H. *et al.* (2016) 'RNA splicing factors as oncoproteins and tumour suppressors', *Nature Reviews Cancer*, 16(7), pp. 413–430. doi: 10.1038/nrc.2016.51.

Edwards, A. O. *et al.* (2005) 'Complement factor H polymorphism and age-related macular degeneration', *Science*, 308(5720), pp. 421–424. doi: 10.1126/science.1110189.

Ejma, M. *et al.* (2020) 'The links between parkinson's disease and cancer', *Biomedicines*, 8(10), pp. 1–26. doi: 10.3390/biomedicines8100416.

Ewels, P. *et al.* (2016) 'MultiQC: Summarize analysis results for multiple tools and samples in a single report', *Bioinformatics*, 32(19), pp. 3047–3048. doi: 10.1093/bioinformatics/btw354.

Ewen Callaway (2017) 'Genome studies attract criticism', *Nature*, 546(In Focus), p. 463. Available at: https://www.nature.com/polopoly_fs/1.22152!/menu/main/topColumns/topLeftColumn/pdf/nature.2017.22152.pdf.

Fehrmann, R. S. N. *et al.* (2011) 'Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla', *PLoS Genetics*, 7(8). doi: 10.1371/journal.pgen.1002197.

Ferreira, E. N. *et al.* (2007) 'Alternative splicing: A bioinformatics perspective', *Molecular BioSystems*. The Royal Society of Chemistry, 3(7), pp. 473–477. doi: 10.1039/b702485c.

Frazer, K. A. *et al.* (2007) 'A second generation human haplotype map of over 3.1 million SNPs', *Nature*, 449(7164), pp. 851–861. doi: 10.1038/nature06258.

Gallagher, M. D. and Chen-Plotkin, A. S. (2018) 'The Post-GWAS Era: From Association to Function', *American Journal of Human Genetics*. Cell Press, pp. 717–730. doi: 10.1016/j.ajhg.2018.04.002.

Galligan, J. J. *et al.* (2018) 'Methylglyoxal-derived posttranslational arginine modifications are abundant histone marks', *Proceedings of the National Academy of Sciences of the United States of America*, 115(37), pp. 9228–9233. doi: 10.1073/pnas.1802901115.

Gamazon, E. R. *et al.* (2018) 'Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation', *Nature Genetics*, 50(7), pp. 956–967. doi: 10.1038/s41588-018-0154-4.

Gay, N. R. *et al.* (2020) 'Impact of admixture and ancestry on eQTL analysis and GWAS

colocalization in GTEx', *Genome Biology*. BioMed Central Ltd, 21(1), p. 233. doi: 10.1186/s13059-020-02113-0.

Genome Reference Consortium (2019) *Genome Reference Consortium Human Build 38 patch release 13 (GRCh38. p13)*.

Goldman, D. and Domschke, K. (2014) 'Making sense of deep sequencing', *International Journal of Neuropsychopharmacology*, pp. 1717–1725. doi: 10.1017/S1461145714000789.

Goruppi, S. *et al.* (2017) 'The ULK3 Kinase Is Critical for Convergent Control of Cancer-Associated Fibroblast Activation by CSL and GLI', *Cell Reports*. ElsevierCompany., 20(10), pp. 2468–2479. doi: 10.1016/j.celrep.2017.08.048.

van der Graaf, A. *et al.* (2020) 'Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids', *Nature Communications*, 11(1), p. 4930. doi: 10.1038/s41467-020-18716-x.

Gretarsdottir, S. *et al.* (2015) 'A Splice Region Variant in LDLR Lowers Non-high Density Lipoprotein Cholesterol and Protects against Coronary Artery Disease', *PLoS Genetics*, 11(9), pp. 1–20. doi: 10.1371/journal.pgen.1005379.

Haberle, V. and Stark, A. (2018) 'Eukaryotic core promoters and the functional basis of transcription initiation', *Nature Reviews Molecular Cell Biology*. Springer US, 19(10), pp. 621–637. doi: 10.1038/s41580-018-0028-8.

Haines, J. L. (2005) 'Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration', *Science*, 308(5720), pp. 419–421. doi: 10.1126/science.1110359.

Hanahan, D. and Weinberg, R. A. (2000) 'The Hallmarks of Cancer', *Cell*, 100(1), pp. 57–70. doi: 10.1016/S0092-8674(00)81683-9.

Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*. Elsevier Inc., 144(5), pp. 646–674. doi: 10.1016/j.cell.2011.02.013.

Hansen, T. (2002) 'Genetics of type 2 diabetes', *Current Science*, 83(12), pp. 1477–1482. doi: 10.5005/jp/books/12626_22.

Harati, H. *et al.* (2019) 'No evidence of a causal association of type 2 diabetes and glucose metabolism with atrial fibrillation', *Diabetologia*. Diabetologia, 62(5), pp. 800–804. doi: 10.1007/s00125-019-4836-y.

Harbeck, N. *et al.* (2019) 'Breast cancer', *Nature reviews. Disease primers*. NLM (Medline), p. 66. doi: 10.1038/s41572-019-0111-2.

Head, S. R. *et al.* (2014) 'Library construction for next-generation sequencing: Overviews and challenges', *BioTechniques*. NIH Public Access, 56(2), pp. 61–77. doi: 10.2144/000114133.

Hebsgaard, S. M. *et al.* (1996) 'Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information', *Nucleic Acids Research*, 24(17), pp. 3439–3452. doi: 10.1093/nar/24.17.3439.

Hernandez, R. D. *et al.* (2019) 'Ultrarare variants drive substantial cis heritability of human gene expression', *Nature Genetics*. Springer US, 51(9), pp. 1349–1355. doi: 10.1038/s41588-019-0487-7.

Herzel, L. *et al.* (2017) 'Splicing and transcription touch base: co-transcriptional spliceosome assembly and function', *Nature Reviews Molecular Cell Biology*, 18(10), pp. 637–650. doi: 10.1038/nrm.2017.63.

Heyn, H. *et al.* (2013) 'DNA methylation contributes to natural human variation', *Genome Research*, 23(9), pp. 1363–1372. doi: 10.1101/gr.154187.112.

Hong, E. P. and Park, J. W. (2012) 'Sample Size and Statistical Power Calculation in Genetic Association Studies', *Genomics & Informatics*, 10(2), p. 117. doi: 10.5808/gi.2012.10.2.117.

Hormozdiari, F. *et al.* (2016) 'Colocalization of GWAS and eQTL Signals Detects Target Genes', *American Journal of Human Genetics*, 99(6), pp. 1245–1260. doi: 10.1016/j.ajhg.2016.10.003.

Hunt, S. E. *et al.* (2018) 'Ensembl variation resources', *Database : the journal of biological databases and curation*, 2018, pp. 1–12. doi: 10.1093/database/bay119.

Ismail, I. A. *et al.* (2014) 'DJ-1 upregulates breast cancer cell invasion by repressing KLF17 expression', *British Journal of Cancer*, 110(5), pp. 1298–1306. doi: 10.1038/bjc.2014.40.

Jaenisch, R. and Bird, A. (2003) 'Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals', *Nature Genetics*, 33(S3), pp. 245–254. doi: 10.1038/ng1089.

Jamin, A., Wicklund, A. and Wiebe, M. S. (2014) 'Cell- and Virus-Mediated Regulation of the Barrier-to-Autointegration Factor's Phosphorylation State Controls Its DNA Binding, Dimerization, Subcellular Localization, and Antipoxviral Activity', *Journal of Virology*. American Society for Microbiology, 88(10), pp. 5342–5355. doi: 10.1128/jvi.00427-14.

Jian, X., Boerwinkle, E. and Liu, X. (2014) 'In silico tools for splicing defect prediction: a

survey from the viewpoint of end users', *Genetics in Medicine*, 16(7), pp. 497–503. doi: 10.1038/gim.2013.176.

Kachuri, L. *et al.* (2020) 'Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction', *Nature Communications*. Springer US, 11(1), pp. 1–11. doi: 10.1038/s41467-020-19600-4.

Kalimutho, M. *et al.* (2019) 'Patterns of Genomic Instability in Breast Cancer', *Trends in Pharmacological Sciences*. Elsevier Ltd, 40(3), pp. 198–211. doi: 10.1016/j.tips.2019.01.005.

Karki, R. *et al.* (2015) 'Defining "mutation" and "polymorphism" in the era of personal genomics', *BMC Medical Genomics*. BMC Medical Genomics, 8(1), pp. 1–7. doi: 10.1186/s12920-015-0115-z.

Katoh, Y. and Katoh, M. (2009) 'Hedgehog Target Genes: Mechanisms of Carcinogenesis Induced by Aberrant Hedgehog Signaling Activation', *Current Molecular Medicine*, 9(7), pp. 873–886. doi: 10.2174/156652409789105570.

Key, T. J. *et al.* (2003) 'Body mass index, serum sex hormones, and breast cancer risk in postmenopausal women', *Journal of the National Cancer Institute*, 95(16), pp. 1218–1226. doi: 10.1093/jnci/djg022.

Key, T. J. *et al.* (2011) 'Circulating sex hormones and breast cancer risk factors in postmenopausal women: reanalysis of 13 studies', *British Journal of Cancer*, 105(5), pp. 709–722. doi: 10.1038/bjc.2011.254.

Kim, R. H. *et al.* (2005) 'DJ-1, a novel regulator of the tumor suppressor PTEN', *Cancer Cell*. Cell Press, 7(3), pp. 263–273. doi: 10.1016/j.ccr.2005.02.010.

Klein, R. J. *et al.* (2005) 'Complement factor H polymorphism in age-related macular degeneration', *Science*, 308(5720), pp. 385–389. doi: 10.1126/science.1109557.

Knight, J. C. (2004) 'Allele-specific gene expression uncovered', *Trends in Genetics*, 20(3), pp. 113–116. doi: 10.1016/j.tig.2004.01.001.

Kolovos, P. *et al.* (2012) 'Enhancers and silencers: An integrated and simple model for their function', *Epigenetics and Chromatin*. BioMed Central Ltd, 5(1), p. 1. doi: 10.1186/1756-8935-5-1.

Kravitz, E. *et al.* (2013) 'Parkinson's disease genes do not segregate with breast cancer genes' loci', *Cancer Epidemiology Biomarkers and Prevention*, 22(8), pp. 1464–1472. doi: 10.1158/1055-9965.EPI-13-0472.

Kruger, K. *et al.* (1982) 'Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena', *Cell*, 31(1), pp. 147–157. doi: 10.1016/0092-8674(82)90414-7.

Lakshminarasimhan, M. *et al.* (2008) 'Structural impact of three Parkinsonism-associated missense mutations on human DJ-1', *Biochemistry*, 47(5), pp. 1381–1392. doi: 10.1021/bi701189c.

Lappalainen, T. *et al.* (2013) 'Transcriptome and genome sequencing uncovers functional variation in humans', *Nature*, 501(7468), pp. 506–511. doi: 10.1038/nature12531.

Lee, J. K. *et al.* (2016) 'Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies', *Annual Review of Pathology: Mechanisms of Disease*, 11(February), pp. 283–312. doi: 10.1146/annurev-pathol-012615-044446.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: Emerging characteristics and insights into transcriptional regulation', *Nature Reviews Genetics*. Nature Publishing Group, 13(4), pp. 233–245. doi: 10.1038/nrg3163.

Levine, A. A. A. (2001) *BioInformatics Approaches to RNA Splicing*. University of Cambridge and The Sanger Centre. Available at: ftp://ftp.sanger.ac.uk/pub/resources/theses/levine/.

Lewis, C. M. and Knight, J. (2012) 'Introduction to genetic association studies', *Cold Spring Harbor Protocols*, 7(3), pp. 297–306. doi: 10.1101/pdb.top068163.

Li, D. *et al.* (2019) 'BRCA1—No Matter How You Splice It', *Cancer Research*. American Association for Cancer Research, 79(9), pp. 2091–2098. doi: 10.1158/0008-5472.CAN-18-3190.

Li, G. *et al.* (2012) 'Identification of allele-specific alternative mRNA processing via transcriptome sequencing', *Nucleic Acids Research*. Narnia, 40(13), pp. e104–e104. doi: 10.1093/nar/gks280.

Li, J. *et al.* (2018) 'Barrier-to-autointegration factor 1: A novel biomarker for gastric cancer', *Oncology Letters*. Spandidos Publications, 16(5), pp. 6488–6494. doi: 10.3892/ol.2018.9432.

Li, Y. I. *et al.* (2017) 'Annotation-free quantification of RNA splicing using LeafCutter Supplementary Note for Annotation-free quantification of RNA splicing using LeafCutter'. doi: 10.1038/s41588-017-0004-9.

Li, Y. I. *et al.* (2018) 'Annotation-free quantification of RNA splicing using LeafCutter', *Nature Genetics*. Springer US, 50(1), pp. 151–158. doi: 10.1038/s41588-017-0004-9.

Liudkovska, V. and Dziembowski, A. (2020) 'Functions and mechanisms of RNA tailing by metazoan terminal nucleotidyltransferases', *Wiley Interdisciplinary Reviews: RNA*, (June). doi: 10.1002/wrna.1622.

Ma, M. *et al.* (2015) 'Disease-associated variants in different categories of disease located in distinct regulatory elements', *BMC Genomics*. BioMed Central Ltd, 16(8), p. S3. doi: 10.1186/1471-2164-16-S8-S3.

Machiela, M. J. and Chanock, S. J. (2015) 'LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants', *Bioinformatics*, 31(21), pp. 3555–3557. doi: 10.1093/bioinformatics/btv402.

Magno, R. and Maia, A.-T. (2019) 'gwasrapidd: an R package to query, download and wrangle GWAS catalog data', *Bioinformatics*. Edited by J. Wren. Oxford University Press (OUP), 36(2), pp. 649–650. doi: 10.1093/bioinformatics/btz605.

Maloverjan, A., Piirsoo, M., Kasak, L., *et al.* (2010) 'Dual function of UNC-51-like kinase 3 (Ulk3) in the Sonic hedgehog signaling pathway', *Journal of Biological Chemistry*, 285(39), pp. 30079–30090. doi: 10.1074/jbc.M110.133991.

Maloverjan, A., Piirsoo, M., Michelson, P., *et al.* (2010) 'Identification of a novel serine/threonine kinase ULK3 as a positive regulator of Hedgehog pathway', *Experimental Cell Research*. Academic Press Inc., 316(4), pp. 627–637. doi: 10.1016/j.yexcr.2009.10.018.

Maney, D. L. (2017) 'Polymorphisms in sex steroid receptors: From gene sequence to behavior', *Frontiers in Neuroendocrinology*. Academic Press Inc., pp. 47–65. doi: 10.1016/j.yfrne.2017.07.003.

Manning, K. S. and Cooper, T. A. (2017) 'The roles of RNA processing in translating genotype to phenotype', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 18(2), pp. 102–114. doi: 10.1038/nrm.2016.139.

Mao, F. *et al.* (2016) 'RBP-var: A database of functional variants involved in regulation mediated by RNA-binding proteins', *Nucleic Acids Research*, 44(D1), pp. D154–D163. doi: 10.1093/nar/gkv1308.

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

Maurano, M. T. *et al.* (2012) 'Systematic localization of common disease-associated variation in regulatory DNA', *Science*, 337(6099), pp. 1190–1195. doi: 10.1126/science.1222794.

Mazin, P. V. *et al.* (2018) 'Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques', *Rna*, 24(4), pp. 585–596. doi: 10.1261/rna.064931.117.

McCann, K. E. and Hurvitz, S. A. (2018) 'Advances in the use of PARP inhibitor therapy for breast cancer', *Drugs in Context*, 7, pp. 1–30. doi: 10.7573/dic.212540.

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

McLaren, W. *et al.* (2016) 'The Ensembl Variant Effect Predictor', *Genome Biology*, 17(1), p. 122. doi: 10.1186/s13059-016-0974-4.

McMahon, A. P., Ingham, P. W. and Tabin, C. J. (2003) '1 Developmental roles and clinical significance of Hedgehog signaling', in *Current Topics in Developmental Biology*, pp. 1–114. doi: 10.1016/S0070-2153(03)53002-2.

Michailidou, K. *et al.* (2017) 'Association analysis identifies 65 new breast cancer risk loci', *Nature*. Nature Publishing Group, 551(7678), pp. 92–94. doi: 10.1038/nature24284.

Mignot, F. *et al.* (2017) 'Concurrent administration of anti-HER2 therapy and radiotherapy: Systematic review', *Radiotherapy and Oncology*. Elsevier B.V., 124(2), pp. 190–199. doi: 10.1016/j.radonc.2017.07.006.

Mills, M. C. and Rahal, C. (2019) 'A scientometric review of genome-wide association studies', *Communications Biology*. Springer US, 2(1). doi: 10.1038/s42003-018-0261-x.

Monlong, J. *et al.* (2014) 'Identification of genetic variants associated with alternative splicing using sQTLseekeR', *Nature Communications*. Nature Publishing Group, 5(May). doi: 10.1038/ncomms5698.

Montes de Oca, R. *et al.* (2009) 'Barrier-to-autointegration factor proteome reveals chromatin-regulatory partners', *PLoS ONE*, 4(9), p. 7050. doi: 10.1371/journal.pone.0007050.

Moo, T. A. *et al.* (2018) 'Overview of Breast Cancer Therapy', *PET Clinics*. Elsevier Inc, 13(3), pp. 339–354. doi: 10.1016/j.cpet.2018.02.006.

Moore, J. E. *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*. Nature Publishing Group, 583(7818), pp. 699–710. doi: 10.1038/s41586-020-2493-4.

Morgan, M. *et al.* (2017) 'Environmental estrogen-like endocrine disrupting chemicals and breast cancer', *Molecular and Cellular Endocrinology*. Elsevier Ireland Ltd, 457, pp. 89–102. doi:

10.1016/j.mce.2016.10.003.

Mostafavi, S. *et al.* (2013) 'Normalizing RNA-Sequencing Data by Modeling Hidden Covariates with Prior Knowledge'. doi: 10.1371/journal.pone.0068141.

Mucci, L. A. *et al.* (2016) 'Familial Risk and Heritability of Cancer Among Twins in Nordic Countries', *Jama*, 315(1), pp. 68–76. doi: 10.1001/jama.2015.17703.Familial.

Mudge, J. M. *et al.* (2019) 'Discovery of high-confidence human protein-coding genes and exons by whole-genome PhyloCSF helps elucidate 118 GWAS loci', *Genome Research*, 29(12), pp. 2073–2087. doi: 10.1101/gr.246462.118.

Myers, T. A., Chanock, S. J. and Machiela, M. J. (2020) 'LDlinkR: An R Package for Rapidly Calculating Linkage Disequilibrium Statistics in Diverse Populations', *Frontiers in Genetics*, 11(February), pp. 1–5. doi: 10.3389/fgene.2020.00157.

National Institutes of Health; and U.S. Department of Health and Human Services (2018) 'Breast Cancer Risk and Environmental Factors', (November).

Nguyen, H. T. *et al.* (2016) 'SRBreak: A Read-Depth and Split-Read Framework to Identify Breakpoints of Different Events Inside Simple Copy-Number Variable Regions.', *Frontiers in genetics*. Frontiers Media SA, 7, p. 160. doi: 10.3389/fgene.2016.00160.

Nickels, S. *et al.* (2013) 'Evidence of Gene-Environment Interactions between Common Breast Cancer Susceptibility Loci and Established Environmental Risk Factors', *PLoS Genetics*, 9(3). doi: 10.1371/journal.pgen.1003284.

Non, A. L. and Thayer, Z. M. (2019) 'Epigenetics and Human Variation', *A Companion to Anthropological Genetics*, pp. 293–308. doi: 10.1002/9781118768853.ch19.

Van Nostrand, E. L. *et al.* (2020) 'A large-scale binding and functional map of human RNA-binding proteins', *Nature*, 583(7818), pp. 711–719. doi: 10.1038/s41586-020-2077-3.

Nuytemans, K. *et al.* (2010) 'Genetic etiology of Parkinson disease associated with mutations in the SNCA, PARK2, PINK1, PARK7, and LRRK2 genes: A mutation update', *Human Mutation*, 31(7), pp. 763–780. doi: 10.1002/humu.21277.

Ongen, H. *et al.* (2016) 'Fast and efficient QTL mapper for thousands of molecular phenotypes', *Bioinformatics*, 32(10), pp. 1479–1485. doi: 10.1093/bioinformatics/btv722.

Pajares, M. J. *et al.* (2007) 'Alternative splicing: an emerging topic in molecular and clinical oncology', *Lancet Oncology*, pp. 349–357. doi: 10.1016/S1470-2045(07)70104-3.

Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human

transcriptome by high-throughput sequencing', *Nature Genetics*. Nature Publishing Group, 40(12), pp. 1413–1415. doi: 10.1038/ng.259.

Panagiotou, O. A. *et al.* (2012) 'What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations', *International Journal of Epidemiology*, 41(1), pp. 273–286. doi: 10.1093/ije/dyr178.

Papadopoulos, J. S. and Agarwala, R. (2007) 'COBALT: Constraint-based alignment tool for multiple protein sequences', *Bioinformatics*. Bioinformatics, 23(9), pp. 1073–1079. doi: 10.1093/bioinformatics/btm076.

Park, E. *et al.* (2018) 'The Expanding Landscape of Alternative Splicing Variation in Human Populations', *American Journal of Human Genetics*. Elsevier, 102(1), pp. 11–26. doi: 10.1016/j.ajhg.2017.11.002.

Park, S. *et al.* (2019) 'Differential Functions of Splicing Factors in Mammary Transformation and Breast Cancer Metastasis', *Cell Reports*, 29(9), pp. 2672-2688.e7. doi: 10.1016/j.celrep.2019.10.110.

PARK7 - Parkinson disease protein 7 precursor - Homo sapiens (Human) - PARK7 gene & protein (2020). Available at: https://www.uniprot.org/uniprot/Q99497 (Accessed: 9 December 2020).

Parkin, D. M., Boyd, L. and Walker, L. C. (2011) 'The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010', *British Journal of Cancer*. Nature Publishing Group, 105(S2), pp. S77–S81. doi: 10.1038/bjc.2011.489.

Paz, I. *et al.* (2014) 'RBPmap: a web server for mapping binding sites of RNA-binding proteins', *Nucleic Acids Research*, (1). doi: 10.1093/nar/gku406.

Pearson, C. E., Edamura, K. N. and Cleary, J. D. (2005) 'Repeat instability: Mechanisms of dynamic mutations', *Nature Reviews Genetics*, 6(10), pp. 729–742. doi: 10.1038/nrg1689.

Piovesan, A. *et al.* (2019) 'Human protein-coding genes and gene feature statistics in 2019', *BMC Research Notes*. BioMed Central Ltd., 12(1), p. 315. doi: 10.1186/s13104-019-4343-8.

Pompei, L. de M. and Fernandes, C. E. (2020) 'Hormone Therapy, Breast Cancer Risk and the Collaborative Group on Hormonal Factors in Breast Cancer Article', *Revista Brasileira de Ginecologia e Obstetrícia / RBGO Gynecology and Obstetrics*, 42(05), pp. 233–234. doi: 10.1055/s-0040-1712941.

Popejoy, A. B. and Fullerton, S. M. (2016) 'Genomics is failing on diversity', *Nature*,

538(7624), pp. 161–164. doi: 10.1038/538161a.

Porcu, E. *et al.* (2019) 'Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–12. doi: 10.1038/s41467-019-10936-0.

Provencio, M. *et al.* (2018) 'SEOM clinical guidelines for the treatment of non-small cell lung cancer (2018)', *Clinical and Translational Oncology*, 21(1), pp. 3–17. doi: 10.1007/s12094-018-1978-1.

Python Software Foundation (2020) *Welcome to Python.org*. Available at: https://www.python.org/ (Accessed: 8 October 2020).

Quail, D. F. and Dannenberg, A. J. (2019) 'The obese adipose tissue microenvironment in cancer development and progression', *Nature Reviews Endocrinology*. Springer US, 15(3), pp. 139–154. doi: 10.1038/s41574-018-0126-x.

R Core Team (2020) 'R: A Language and Environment for Statistical Computing'. Vienna, Austria. Available at: https://www.r-project.org.

Raplee, I. D., Evsikov, A. V. and De Evsikova, C. M. (2019) 'Aligning the aligners: Comparison of rna sequencing data alignment and gene expression quantification tools for clinical breast cancer research', *Journal of Personalized Medicine*, 9(2), pp. 1–18. doi: 10.3390/jpm9020018.

Rashkin, S. R. *et al.* (2020) 'Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts', *Nature Communications*. Springer US, 11(1). doi: 10.1038/s41467-020-18246-6.

Raychaudhuri, S. (2011) 'Mapping rare and common causal alleles for complex human diseases', *Cell*. Elsevier Inc., 147(1), pp. 57–69. doi: 10.1016/j.cell.2011.09.011.

Ridge, P. G. *et al.* (2013) 'Alzheimer's disease: Analyzing the missing heritability', *PLoS ONE*, 8(11), pp. 1–10. doi: 10.1371/journal.pone.0079771.

Ringnér, M. (2008) 'What is principal component analysis?', *Nature Biotechnology*, 26(3), pp. 303–304. doi: 10.1038/nbt0308-303.

Ripperger, T. *et al.* (2009) 'Breast cancer susceptibility: current knowledge and implications for genetic counselling', *European Journal of Human Genetics*. BioMed Central, 17(6), pp. 722–731. doi: 10.1038/ejhg.2008.212.

Rojano, E. *et al.* (2019) 'Regulatory variants: from detection to predicting impact',

Briefings in Bioinformatics. Oxford University Press, 20(5), pp. 1639–1654. doi: 10.1093/bib/bby039.

Roma-Rodrigues, C. *et al.* (2019) 'Targeting tumor microenvironment for cancer therapy', *International Journal of Molecular Sciences*, 20(4). doi: 10.3390/ijms20040840.

Romanoski, C. E. *et al.* (2015) 'Roadmap for regulation', *Nature*, 518(7539), pp. 314–316. doi: 10.1038/518314a.

Saraiva-Agostinho, N. and Barbosa-Morais, N. L. (2019) 'psichomics: graphical application for alternative splicing quantification and analysis', *Nucleic Acids Research*. Oxford University Press, 47(2), pp. e7–e7. doi: 10.1093/nar/gky888.

Schaid, D. J., Chen, W. and Larson, N. B. (2018) 'From genome-wide associations to candidate causal variants by statistical fine-mapping', *Nature Reviews Genetics*. Nature Publishing Group, pp. 491–504. doi: 10.1038/s41576-018-0016-z.

Schaub, M. A. *et al.* (2012) 'Linking disease associations with regulatory information in the human genome', *Genome Research*, 22(9), pp. 1748–1759. doi: 10.1101/gr.136127.111.

Schinzel, A. (2015) *Genetics and Genomics in Medicine*, *European Journal of Human Genetics*. doi: 10.1038/ejhg.2015.18.

Schmid, P. *et al.* (2018) 'Atezolizumab and Nab-Paclitaxel in Advanced Triple-Negative Breast Cancer', *New England Journal of Medicine*, 379(22), pp. 2108–2121. doi: 10.1056/NEJMoa1809615.

Schwab, M. (2017) *Encyclopedia of Cancer Fourth Edition*, *New York*.

Scotti, M. M. and Swanson, M. S. (2016) 'RNA mis-splicing in disease', *Nature Reviews Genetics*. Nature Publishing Group, 17(1), pp. 19–32. doi: 10.1038/nrg.2015.3.

Scriver, C. R. and Waters, P. J. (1999) 'Monogenic traits are not simple: Lessons from phenylketonuria', *Trends in Genetics*, 15(7), pp. 267–272. doi: 10.1016/S0168-9525(99)01761-8.

Sears, R. M. and Roux, K. J. (2020) 'Diverse cellular functions of barrier-to-autointegration factor and its roles in disease', *Journal of cell science*, 133(16). doi: 10.1242/jcs.246546.

Shen, S. *et al.* (2014) 'rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 111(51), pp. E5593-601. doi: 10.1073/pnas.1419161111.

Sheng, Q. *et al.* (2017) 'Multi-perspective quality control of Illumina RNA sequencing data

analysis', *Briefings in Functional Genomics*, 16(4), pp. 194–204. doi: 10.1093/bfgp/elw035.

Solovei, I., Thanisch, K. and Feodorova, Y. (2016) 'How to rule the nucleus: divide et impera', *Current Opinion in Cell Biology*. Elsevier Ltd, 40, pp. 47–59. doi: 10.1016/j.ceb.2016.02.014.

Sørlie, T. *et al.* (2001) 'Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications', *Proceedings of the National Academy of Sciences of the United States of America*, 98(19), pp. 10869–10874. doi: 10.1073/pnas.191367098.

SRA-Tools - NCBI (2015). Available at: https://ncbi.github.io/sra-tools/ (Accessed: 23 September 2020).

Stanta, G. and Bonin, S. (2018) 'Overview on clinical relevance of intra-tumor heterogeneity', *Frontiers in Medicine*, 5(APR), pp. 1–10. doi: 10.3389/fmed.2018.00085.

van Steensel, B. and Belmont, A. S. (2017) 'Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression', *Cell*. Elsevier Inc., 169(5), pp. 780–791. doi: 10.1016/j.cell.2017.04.022.

Stevens, M. and Oltean, S. (2019) 'Modulation of the Apoptosis Gene Bcl-x Function Through Alternative Splicing', *Frontiers in Genetics*. Frontiers Media S.A., 10(September), pp. 1–9. doi: 10.3389/fgene.2019.00804.

Storey, J. D. and Tibshirani, R. (2003) 'Statistical significance for genomewide studies', *Proceedings of the National Academy of Sciences of the United States of America*, 100(16), pp. 9440–9445. doi: 10.1073/pnas.1530509100.

Strumylaite, L., Mechonošina, K. and Tamašauskas, Š. (2010) 'Environmental factors and breast cancer', *Medicina*, 46(12), pp. 867–873. doi: 10.3390/medicina46120121.

Sudmant, P. H. *et al.* (2015) 'An integrated map of structural variation in 2,504 human genomes', *Nature*, 526(7571), pp. 75–81. doi: 10.1038/nature15394.

Szumilas, M. (2010) 'Explaining odds ratios', *Journal of the Canadian Academy of Child and Adolescent Psychiatry*. Canadian Academy of Child and Adolescent Psychiatry, 19(3), pp. 227–229. Available at: http://www.csm-oxford.org.uk/ (Accessed: 20 October 2020).

Tabor, H. K., Risch, N. J. and Myers, R. M. (2002) 'Candidate-gene approaches for studying complex genetic traits: practical considerations', *Nature Reviews Genetics*, 3(5), pp. 391–397. doi: 10.1038/nrg796.

Tang, R., Prosser, D. O. and Love, D. R. (2016) 'Evaluation of Bioinformatic Programmes

for the Analysis of Variants within Splice Site Consensus Regions', *Advances in Bioinformatics*. Hindawi Limited, 2016. doi: 10.1155/2016/5614058.

Taylor-Weiner, A. *et al.* (2019) 'Scaling computational genomics to millions of individuals with GPUs', *Genome Biology*. BioMed Central Ltd., 20(1), p. 228. doi: 10.1186/s13059-019-1836-7.

Tejedor, J. R. *et al.* (2015) 'Role of six single nucleotide polymorphisms, risk factors in coronary disease, in OLR1 alternative splicing', *RNA (New York, N.Y.)*, 21(6), pp. 1187–1202. doi: 10.1261/rna.049890.115.

The Sequence Read Archive (SRA): Getting Started (no date). Available at: https://www.ncbi.nlm.nih.gov/sra/docs/ (Accessed: 22 September 2020).

Trempe, J. F. and Fon, E. A. (2013) 'Structure and function of Parkin, PINK1, and DJ-1, the three musketeers of neuroprotection', *Frontiers in Neurology*, 4 APR(April), pp. 1–11. doi: 10.3389/fneur.2013.00038.

ULK3 - Serine/threonine-protein kinase ULK3 - Homo sapiens (Human) - ULK3 gene & protein (2020). Available at: https://www.uniprot.org/uniprot/Q6PHR2 (Accessed: 9 December 2020).

Venables, J. P. (2004) 'Aberrant and Alternative Splicing in Cancer', *Cancer Research*. American Association for Cancer Research, 64(21), pp. 7647–7654. doi: 10.1158/0008-5472.CAN-04-1910.

Venables, J. P. *et al.* (2008) 'Identification of alternative splicing markers for breast cancer', *Cancer Research*. American Association for Cancer Research, 68(22), pp. 9525–9531. doi: 10.1158/0008-5472.CAN-08-1769.

Venables, J. P. *et al.* (2009) 'Cancer-associated regulation of alternative splicing', *Nature Structural and Molecular Biology*. Nature Publishing Group, 16(6), pp. 670–676. doi: 10.1038/nsmb.1608.

Vonderheide, R. H., Domchek, S. M. and Clark, A. S. (2017) 'Immunotherapy for breast cancer: What are we missing?', *Clinical Cancer Research*, 23(11), pp. 2640–2646. doi: 10.1158/1078-0432.CCR-16-2569.

Wagner, G. P., Kin, K. and Lynch, V. J. (2013) 'A model based criterion for gene expression calls using RNA-seq data', *Theory in Biosciences*, 132(3), pp. 159–164. doi: 10.1007/s12064-013-0178-3.

Walker, R. L. *et al.* (2019) 'Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms', *Cell*. Elsevier Inc., 179(3), pp. 750-771.e22. doi: 10.1016/j.cell.2019.09.021.

Wall, J. D. and Pritchard, J. K. (2003) 'Haplotype blocks and linkage disequilibrium in the human genome', *Nature Reviews Genetics*, 4(8), pp. 587–597. doi: 10.1038/nrg1123.

Wang, E. T. *et al.* (2008) 'Alternative isoform regulation in human tissue transcriptomes', *Nature*. Nature Publishing Group, 456(7221), pp. 470–476. doi: 10.1038/nature07509.

Wang, S. *et al.* (2019) 'Germline variants and somatic mutation signatures of breast cancer across populations of African and European ancestry in the US and Nigeria', *International Journal of Cancer*, 145(12), pp. 3321–3333. doi: 10.1002/ijc.32498.

Wang, Z. and Burge, C. B. (2008) 'Splicing regulation: From a parts list of regulatory elements to an integrated splicing code', *RNA*. Cold Spring Harbor Laboratory Press, pp. 802–813. doi: 10.1261/rna.876308.

Ward, L. D. and Kellis, M. (2012) 'HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants', *Nucleic Acids Research*. Oxford University Press, 40(D1), p. D930. doi: 10.1093/nar/gkr917.

Weatherall, D. J. (2001) 'Phenotype-genotype relationships in monogenic disease: Lessons from the thalassaemias', *Nature Reviews Genetics*, 2(4), pp. 245–255. doi: 10.1038/35066048.

Webb, C.-H. T. *et al.* (2014) *Spliceosomal Pre-mRNA Splicing*, *Methods in Molecular Biology*. Edited by K. J. Hertel. Totowa, NJ: Humana Press (Methods in Molecular Biology). doi: 10.1007/978-1-62703-980-2.

White, A. J. *et al.* (2016) 'Exposure to multiple sources of polycyclic aromatic hydrocarbons and breast cancer incidence', *Environment International*. Elsevier Ltd, 89–90, pp. 185–192. doi: 10.1016/j.envint.2016.02.009.

Will, C. L. *et al.* (2001) 'A novel U2 and U11/U12 snRNP protein that associates with the pre-mRNA branch site.', *The EMBO journal*. European Molecular Biology Organization, 20(16), pp. 4536–46. doi: 10.1093/emboj/20.16.4536.

Wongpalee, S. P. and Sharma, S. (2014) 'The pre-mRNA splicing reaction', *Methods in Molecular Biology*. Humana Press, Totowa, NJ, 1126, pp. 3–12. doi: 10.1007/978-1-62703-980-2_1.

Xiao, R. and Scott, L. J. (2011) 'Detection of cis-acting regulatory SNPs using allelic

expression data', *Genetic Epidemiology*, 35(6), pp. 515–525. doi: 10.1002/gepi.20601.

Yates, A. D. *et al.* (2020) 'Ensembl 2020', *Nucleic Acids Research*. Oxford University Press, 48(D1), pp. D682–D688. doi: 10.1093/nar/gkz966.

Yee, B. A. *et al.* (2019) 'RBP-Maps enables robust generation of splicing regulatory maps', *Rna*, 25(2), pp. 193–204. doi: 10.1261/rna.069237.118.

Young, A. R. J. *et al.* (2009) 'Autophagy mediates the mitotic senescence transition', *Genes and Development*. Genes Dev, 23(7), pp. 798–803. doi: 10.1101/gad.519709.

Young, R. A. (2011) 'Control of the embryonic stem cell state', *Cell*. Elsevier, pp. 940–954. doi: 10.1016/j.cell.2011.01.032.

Yu, S. and Kim, V. N. (2020) 'A tale of non-canonical tails: gene regulation by post-transcriptional RNA tailing', *Nature Reviews Molecular Cell Biology*. Springer US, 21(9), pp. 542–556. doi: 10.1038/s41580-020-0246-8.

Zaborowska, J., Egloff, S. and Murphy, S. (2016) 'The pol II CTD: New twists in the tail', *Nature Structural and Molecular Biology*, 23(9), pp. 771–777. doi: 10.1038/nsmb.3285.

Zdravkovic, S. *et al.* (2002) 'Heritability of death from coronary heart disease: A 36-year follow-up of 20 966 Swedish twins', *Journal of Internal Medicine*, 252(3), pp. 247–254. doi: 10.1046/j.1365-2796.2002.01029.x.

Zhang, G. (2020) 'Expression and Prognostic Significance of BANF1 in Triple-Negative Breast Cancer', *Cancer Management and Research*, 12, p. 145. doi: 10.2147/CMAR.S229022.

Zhang, J. *et al.* (2020) 'RADAR: annotation and prioritization of variants in the post-transcriptional regulome of RNA-binding proteins', *Genome Biology*. BioMed Central, 21(1), p. 151. doi: 10.1186/s13059-020-01979-4.

Zhao, K. *et al.* (2013) 'GLiMMPS: Robust statistical model for regulatory variation of alternative splicing using RNA-seq data', *Genome Biology*. BioMed Central, 14(7), p. R74. doi: 10.1186/gb-2013-14-7-r74.

Zheng, Q. *et al.* (2019) 'Reversible histone glycation is associated with disease-related changes in chromatin architecture', *Nature Communications*. Nature Publishing Group, 10(1). doi: 10.1038/s41467-019-09192-z.

Ziegler, R. G. *et al.* (1993) 'Migration patterns and breast cancer risk in Asian-American women', *Journal of the National Cancer Institute*, 85(22), pp. 1819–1827. doi: 10.1093/jnci/85.22.1819.

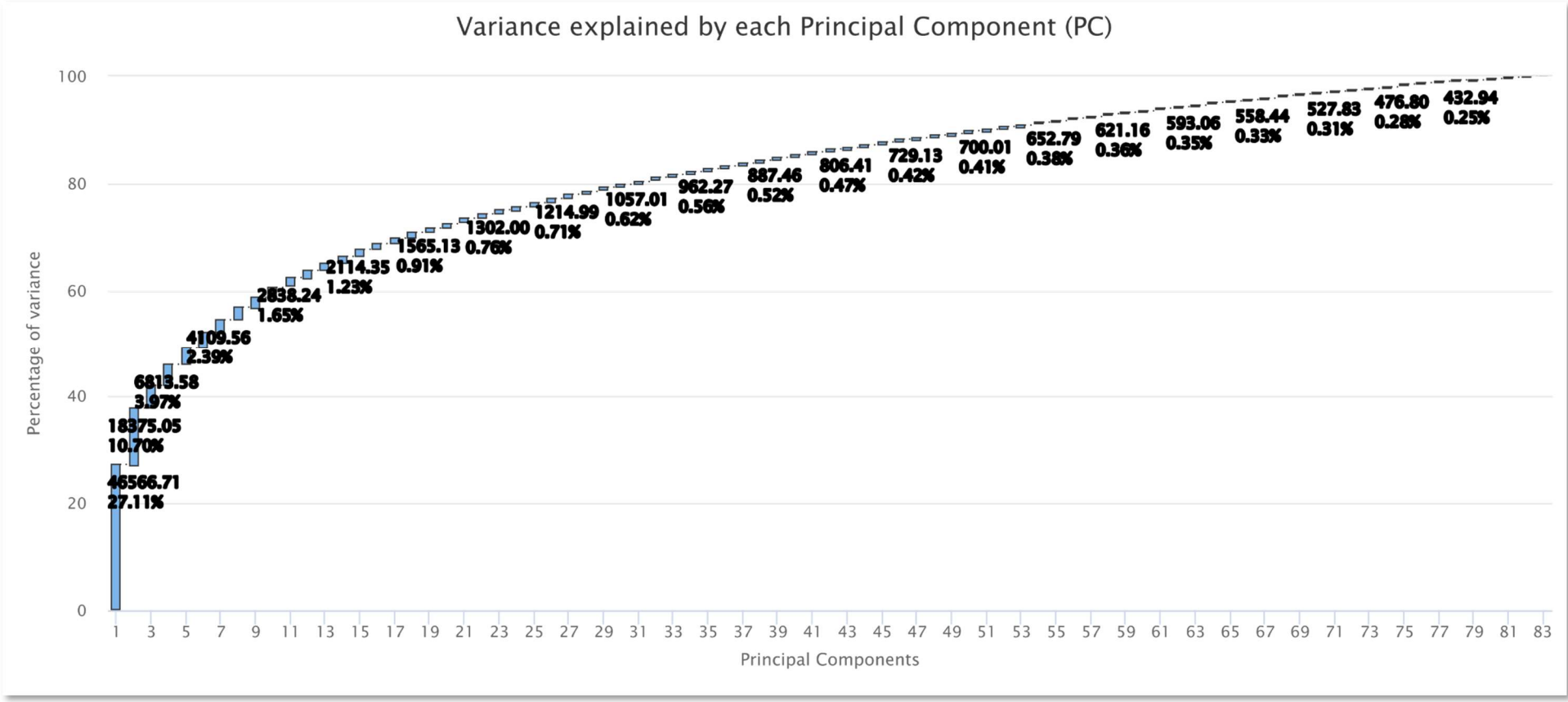
7 References

8 Annexes

8.1 Quality control pre- and post-processing of RNA-seq data

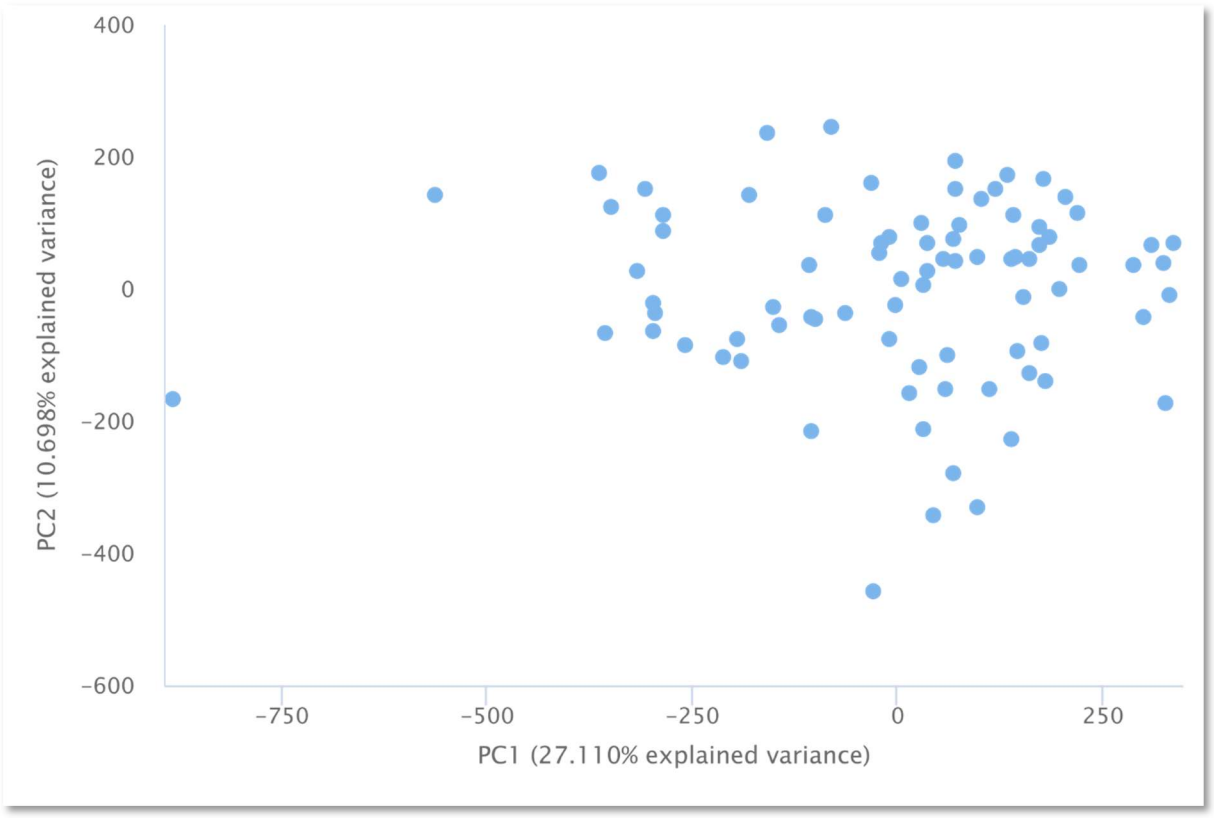
Available in digital support

8.2 Cumulative Principal Component Distribution on samples used on psichomics



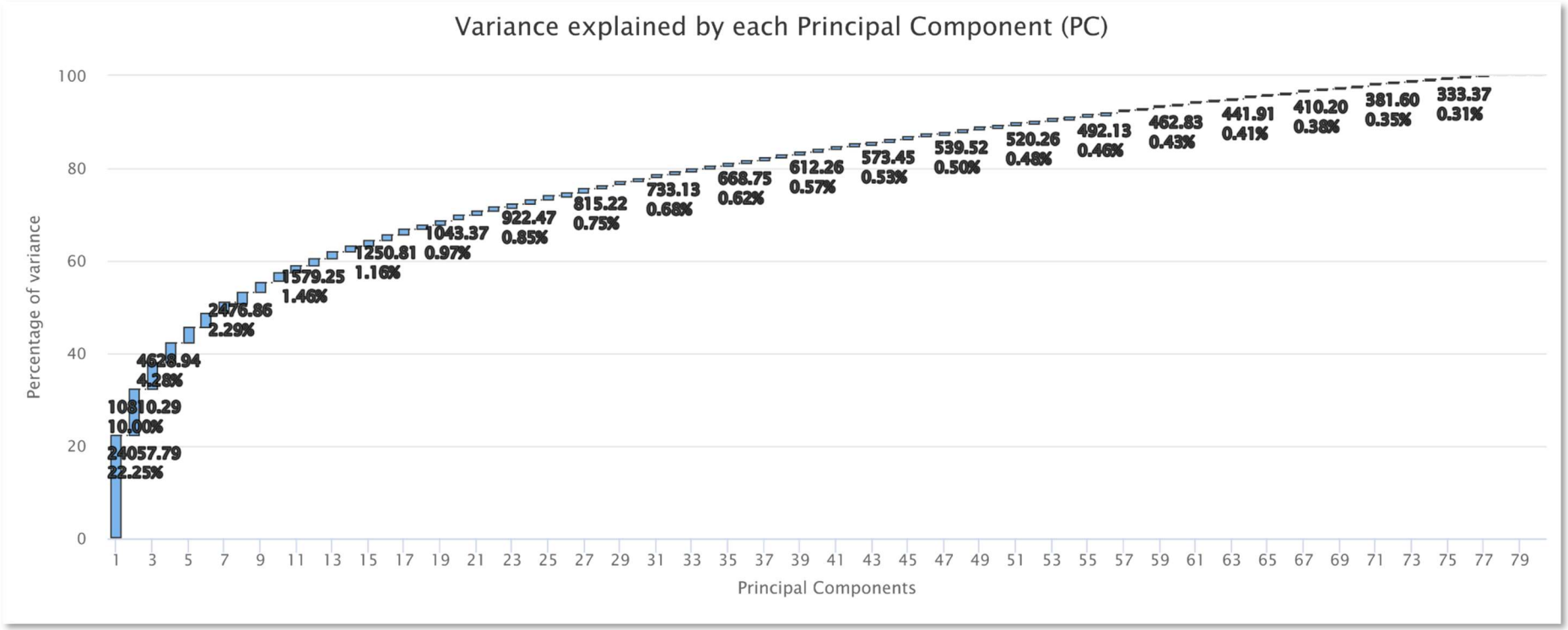
Supplementary Figure 1 – Cumulative variance of splicing explained by each Principal Component

8.3 Principal Component 1 vs Principal Component 2 – psichomics data



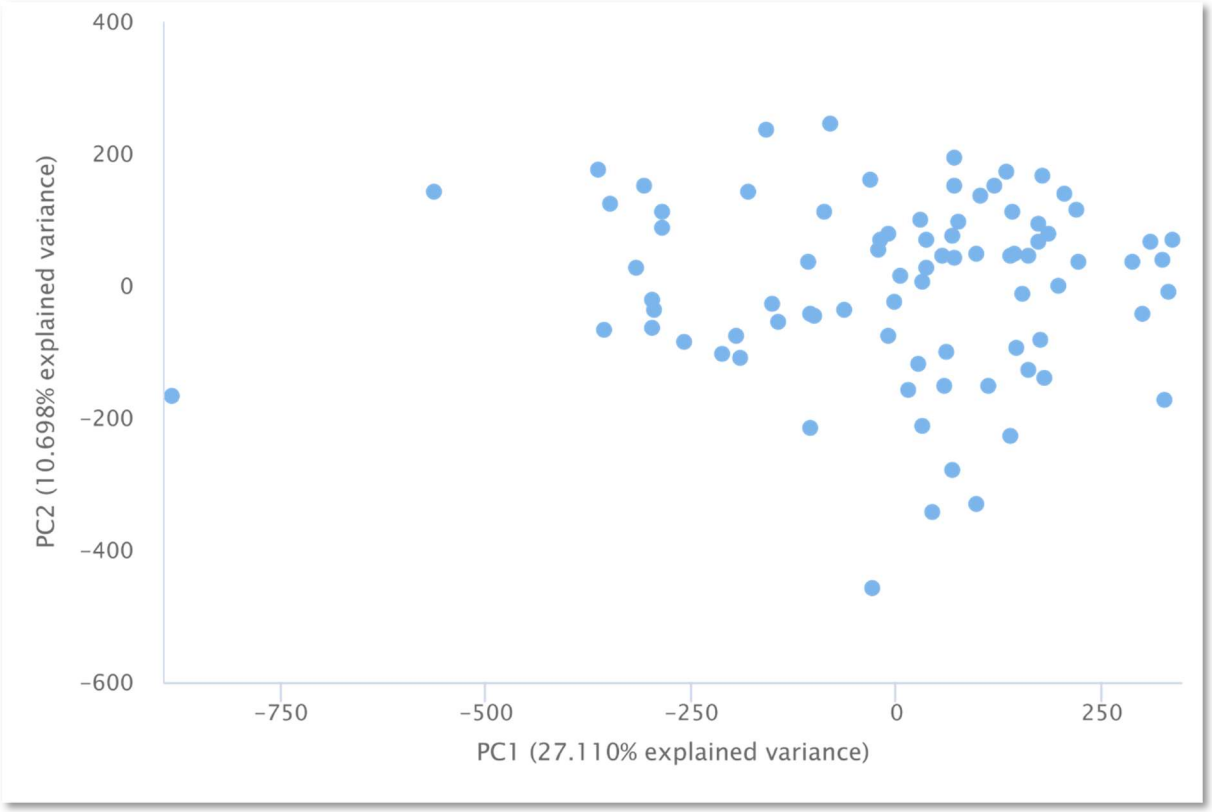
Supplementary Figure 2 – Principal Component 1 vs Principal Component 2 – psichomics data

8.4 Cumulative Principal Component Distribution on samples used on LeafCutter



Supplementary Figure 3 – Cumulative variance of splicing explained by each Principal Component

8.5 Principal Component 1 vs Principal Component 2 – LeafCutter data



Supplementary Figure 4 – Principal Component 1 vs Principal Component 2 – LeafCutter data

8.6 sQTLs identified using LeafCutter's PSI

Available in digital support

8.7 sQTLs identified using psichomics's PSI

Available in digital support

8.8 List of retrieved BC risk GWAS

Available in digital support

8.9 sQTLs obtained using LeafCutter's PSI in co-localization with BC GWAS hit-SNPs

Table 1 – sQTLs retrieved using PSI as calculated from LeafCutter in linkage disequilibrium with breast cancer GWAS hit-SNPs

Phenotype Id	ma samples	ma count	slope	slope se	q-value	Chromosome	Ref	Alt	sQTL	Position	hit-SNP	Ld (r^2)	Gene Ensemble Id	Gene Hgnc Symbol
1:7961735:7962763:clu_5147	22	23	0.09461954	0.010728523	1,96E-05	1	G	T	rs4908724	8059191	rs225132	0,964	ENSG00000116288	PARK7
1:7961793:7962763:clu_5147	22	23	-0.09468271	0.010733087	1,99E-05	1	G	T	rs4908724	8059191	rs225132	0,964	ENSG00000116288	PARK7
1:45330068:45330516:clu_7079	52	64	-0.08847021	0.011810859	0,000733	1	C	T	rs1771550	45417893	rs144366570	0,67	ENSG00000132781	MUTYH
1:45330068:45330516:clu_7079	52	64	-0.08847021	0.011810859	0,000733	1	C	T	rs1771550	45417893	rs144366570	0,67	ENSG00000288208	
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs9257809	0,432	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs1233480	0,404	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3116813	0,518	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs1611579	0,51	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3094146	0,533	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3094054	0,455	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3132615	0,442	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3132610	0,449	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs9262142	0,463	NA	NA
6:29792305:29942757:clu_73023	26	27	0.094161764	0.0027753743	1,15E-28	6	G	A	rs5013089	29852671	rs3129984	0,42	NA	NA
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs9257809	0,432	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs1233480	0,404	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3116813	0,518	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs1611579	0,51	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3094146	0,533	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3094054	0,455	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3132615	0,442	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3132610	0,449	ENSG00000206341	HLA-H
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs9262142	0,463	ENSG00000206341	HLA-H

8 Annexes

Phenotype Id	ma samples	ma count	slope	slope se	q-value	Chromosome	Ref	Alt	sQTL	Position	hit-SNP	Ld (r^2)	Gene Ensemble Id	Gene Hgnc Symbol
6:29890296:29890439:clu_73026	49	65	0.009342863	0.001055156	0,000161	6	C	T	rs1611745	29864739	rs3129984	0,42	ENSG00000206341	HLA-H
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs9257809	0,432	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs9257809	0,432	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs1233480	0,404	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs1233480	0,404	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3116813	0,518	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3116813	0,518	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs1611579	0,51	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs1611579	0,51	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3094146	0,533	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3094146	0,533	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3094054	0,455	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3094054	0,455	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3132615	0,442	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3132615	0,442	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3132610	0,449	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3132610	0,449	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs9262142	0,463	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs9262142	0,463	ENSG00000227766	
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3129984	0,42	ENSG00000206503	HLA-A
6:29942626:29942757:clu_73023	26	27	-0.09422108	0.0027763515	1,15E-28	6	G	A	rs5013089	29852671	rs3129984	0,42	ENSG00000227766	
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs9257809	0,453	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3116813	0,542	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs1611579	0,534	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3094146	0,557	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3094054	0,462	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3132615	0,448	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3132610	0,47	NA	NA

Phenotype Id	ma samples	ma count	slope	slope se	q-value	Chromosome	Ref	Alt	sQTL	Position	hit-SNP	Ld (r^2)	Gene Ensemble Id	Gene Hgnc Symbol
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs9262142	0,484	NA	NA
6:29944616:30009145:clu_73028	39	48	0.10332991	0.0064829784	8,25E-16	6	C	T	rs1655905	29950245	rs3129984	0,44	NA	NA
8:100300651:100303160:clu_45073	7	8	0.04257313	0.005092336	0,01021	8	C	G	rs1660338	1E+08	rs35748790	0,856	ENSG00000034677	RNF19A
10:37177756:37181322:clu_21536	47	55	-0.1946853	0.0312365	0,004639	10	T	C	rs2486117	37097506	rs1200921	0,508	ENSG00000148513	ANKRD30A
10:38365629:38365910:clu_21568	34	38	0.3783606	0.02052056	2,68E-18	10	A	G	rs2997802	38281835	rs143072280	0,479	ENSG00000099251	HSD17B7P2
10:38365629:38365910:clu_21568	34	38	0.3783606	0.02052056	2,68E-18	10	A	G	rs2997802	38281835	rs11146838	0,407	ENSG00000099251	HSD17B7P2
11:428199:428387:clu_23926	20	26	-0.06001266	0.007355452	0,00975	11	T	C	rs6598022	433135	rs11410354	0,956	ENSG00000185101	ANO9
11:66003070:66003235:clu_26097	46	51	0.01027151	0.001031038	1,52E-07	11	TCACTGAG	T	rs56984820	66003580	rs617791	0,712	ENSG00000175334	BANF1
15:74839050:74839268:clu_41293	19	22	-0.009791329	0.001111905	0,0423	15	G	A	rs12591513	74810373	rs6938	0,568	ENSG00000140474	ULK3
15:75352230:75352324:clu_41321	4	4	0.1718545	0.02804546	0,03958	15	A	G	rs28610581	75520133	rs8027365	0,904	ENSG00000140398	NEIL1
15:75352230:75352324:clu_41321	4	4	0.1718545	0.02804546	0,03958	15	A	G	rs28610581	75520133	rs67079557	0,816	ENSG00000140398	NEIL1
17:46094701:46170855:clu_52322	23	27	-0.2413105	0.01290614	4,88E-13	17	T	C	rs62055700	45681376	rs4763	0,679	ENSG00000120071	KANSL1
17:46094701:46170855:clu_52322	23	27	-0.2413105	0.01290614	4,88E-13	17	T	C	rs62055700	45681376	rs2532263	0,724	ENSG00000120071	KANSL1
17:46094701:46170855:clu_52322	23	27	-0.2413105	0.01290614	4,88E-13	17	T	C	rs62055700	45681376	rs2732699	0,565	ENSG00000120071	KANSL1
17:46094701:46170855:clu_52322	23	27	-0.2413105	0.01290614	4,88E-13	17	T	C	rs62055700	45681376	rs199498	0,618	ENSG00000120071	KANSL1
17:46152966:46170855:clu_52322	22	26	0.1722549	0.009073295	1,06E-14	17	T	G	rs17577159	46111111	rs4763	0,665	ENSG00000120071	KANSL1
17:46152966:46170855:clu_52322	22	26	0.1722549	0.009073295	1,06E-14	17	T	G	rs17577159	46111111	rs2532263	0,74	ENSG00000120071	KANSL1
17:46152966:46170855:clu_52322	22	26	0.1722549	0.009073295	1,06E-14	17	T	G	rs17577159	46111111	rs2732699	0,58	ENSG00000120071	KANSL1
17:46152966:46170855:clu_52322	22	26	0.1722549	0.009073295	1,06E-14	17	T	G	rs17577159	46111111	rs199498	0,632	ENSG00000120071	KANSL1
17:54999451:54999632:clu_52687	38	41	0.4460218	0.04322078	1,52E-08	17	G	A	rs9892976	54982968	rs6504950	1	ENSG00000166263	STXBP4
17:54999451:54999632:clu_52687	38	41	0.4460218	0.04322078	1,52E-08	17	G	A	rs9892976	54982968	rs2787486	0,747	ENSG00000166263	STXBP4
19:2131702:2131760:clu_12456	52	63	-0.2794161	0.04655629	0,020266	19	G	A	rs62128400	2133832	rs3815308	0,522	ENSG00000065000	AP3D1
20:35626800:35627280:clu_64875	9	9	0.09100068	0.004064969	1,70E-07	20	G	GC	rs3841688	35617468	rs201177249	0,409	ENSG00000214078	CPNE1
20:35626803:35627280:clu_64875	9	9	-0.1910873	0.01600074	5,76E-05	20	G	GC	rs3841688	35617468	rs201177249	0,409	ENSG00000214078	CPNE1
20:35627413:35630439:clu_64875	9	9	0.1013164	0.01502616	0,049127	20	G	GC	rs3841688	35617468	rs201177249	0,409	ENSG00000214078	CPNE1

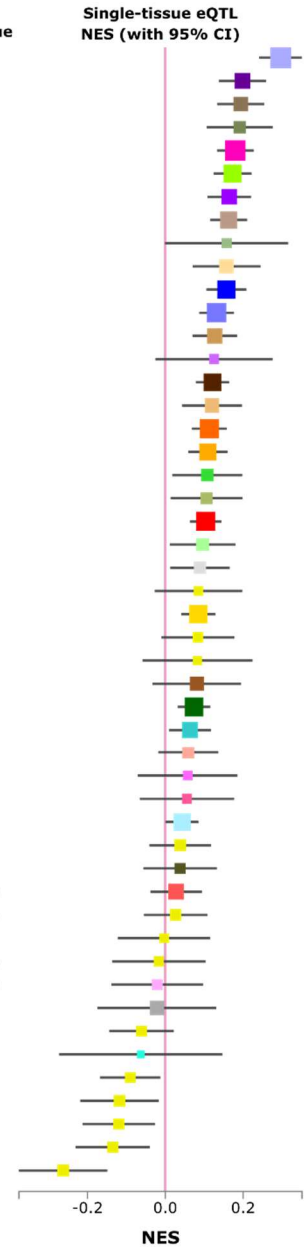
8.10 sQTLs obtained using psichomics's PSI in co-localization with BC GWAS hit-SNPs

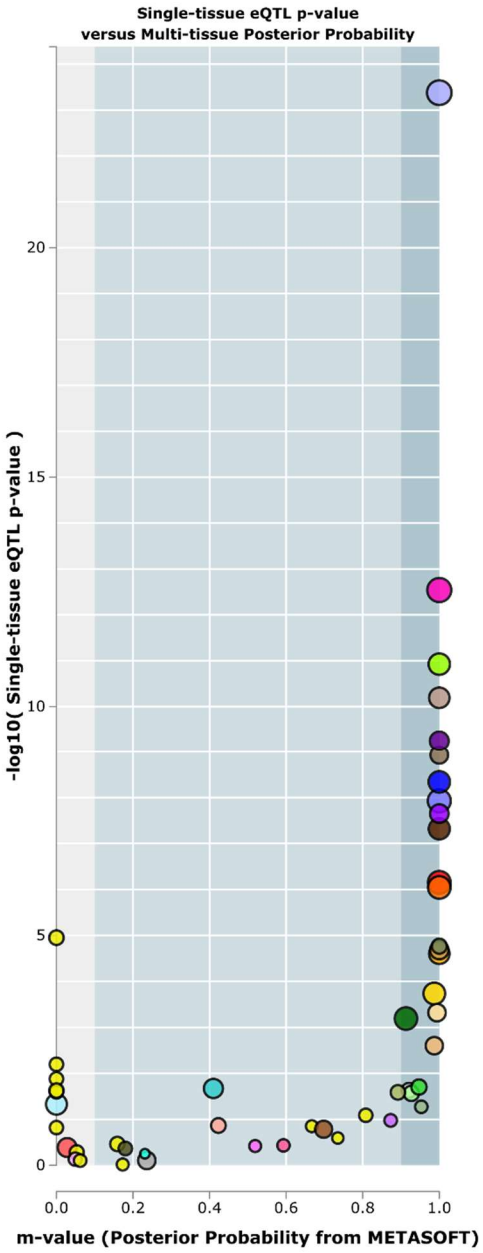
Table 2 – sQTLs retrieved using PSI as calculated by psichomics in linkage disequilibrium with breast cancer GWAS hit-SNPs

Phenotype id	ma samples	ma count	slope	slope se	q-value	Chromosome	Ref	Alt	sQTL	Position	hit-SNP	ld_r^2	gene ensembl id	gene hgnc symbol
A5SS_1_+_7961735_7961793_7962763_PARK7	20	21	-0.105005965	0.015196962	0,006753	1	G	A	rs17229081	7922756	rs225132	0,767	ENSG00000116288	PARK7
AFE_1_+_7961793_7961735_7962763_PARK7	20	21	0.105005965	0.015196962	0,006753	1	G	A	rs17229081	7922756	rs225132	0,767	ENSG00000116288	PARK7
AFE_1_-_155261657_155262086_155260659_SCAMP3	51	65	0.042250495	0.006925904	0,030414	1	C	T	rs2075569	1,55E+08	rs2974935	0,437	ENSG00000116521	SCAMP3
AFE_1_-_155261657_155262086_155260659_SCAMP3	51	65	0.042250495	0.006925904	0,030414	1	C	T	rs2075569	1,55E+08	rs7524950	0,629	ENSG00000116521	SCAMP3
SE_1_-_155262086_155261734_155261657_155260659_SCAMP3	51	65	-0.046766892	0.0071116965	0,00682	1	C	T	rs2075569	1,55E+08	rs2974935	0,437	ENSG00000116521	SCAMP3
SE_1_-_155262086_155261734_155261657_155260659_SCAMP3	51	65	-0.046766892	0.0071116965	0,00682	1	C	T	rs2075569	1,55E+08	rs7524950	0,629	ENSG00000116521	SCAMP3
SE_4_+_83456958_83459740_83459797_83460973_MRPS18C	54	64	0.035141487	0.005140761	0,00416	4	T	C	rs13110130	83508843	rs1963045	0,502	ENSG00000163319	MRPS18C
SE_4_+_83456958_83459740_83459797_83460973_MRPS18C	54	64	0.035141487	0.005140761	0,00416	4	T	C	rs13110130	83508843	rs1963045	0,502	ENSG00000163322	ABRAXAS1
SE_11_+_66002513_66002844_66003070_66003235_BANF1	45	52	0.09525603	0.01282016	0,000671	11	T	C	rs9735063	65997067	rs617791	0,513	ENSG00000175334	BANF1
AFE_11_+_66003070_66002513_66003235_BANF1	45	52	-0.1427121	0.01846617	0,000113	11	T	C	rs9735063	65997067	rs617791	0,513	ENSG00000175334	BANF1
AFE_11_+_66003070_66002570_66003235_BANF1	48	50	-0.03178853	0.003644385	0,000663	11	C	T	rs6591195	66015297	rs617791	0,476	ENSG00000175334	BANF1
A5SS_11_-_66438710_66438632_66438259_MRPL11	37	43	-0.07952543	0.008391142	8,24E-05	11	G	T	rs2282640	66478380	rs11344495	0,417	ENSG00000174547	MRPL11
A5SS_11_-_66438710_66438632_66438259_MRPL11	37	43	-0.07952543	0.008391142	8,24E-05	11	G	T	rs2282640	66478380	rs55908905	0,487	ENSG00000174547	MRPL11
A5SS_11_-_66438710_66438632_66438259_MRPL11	37	43	-0.07952543	0.008391142	8,24E-05	11	G	T	rs2282640	66478380	rs7570	0,476	ENSG00000174547	MRPL11
A5SS_15_-_74837757_74837751_74837435_ULK3	61	76	-0.5085934	0.01928051	4,83E-26	15	T	C	rs12898397	74837752	rs6938	0,554	ENSG00000140474	ULK3

8.11 eQTL rs17229081

Tissue	Samples	NES	p-value	m-value
Muscle - Skeletal	706	0.296	4.2e-24	1.00
Heart - Left Ventricle	386	0.198	5.6e-10	1.00
Esophagus - Gastroesophageal Junction	330	0.194	1.2e-9	1.00
Spleen	227	0.191	1.7e-5	1.00
Whole Blood	670	0.180	2.9e-13	1.00
Lung	515	0.173	1.2e-11	1.00
Heart - Atrial Appendage	372	0.164	2.2e-8	1.00
Esophagus - Muscularis	465	0.163	6.4e-11	1.00
Minor Salivary Gland	144	0.158	0.05	0.953
Stomach	324	0.157	4.8e-4	0.994
Skin - Not Sun Exposed (Suprapubic)	517	0.157	4.6e-9	1.00
Skin - Sun Exposed (Lower leg)	605	0.132	1.2e-8	1.00
Colon - Transverse	368	0.127	2.1e-5	1.00
Cells - EBV-transformed lymphocytes	147	0.125	0.1	0.873
Esophagus - Mucosa	497	0.121	4.8e-8	1.00
Colon - Sigmoid	318	0.120	2.5e-3	0.987
Adipose - Subcutaneous	581	0.114	9.0e-7	1.00
Adipose - Visceral (Omentum)	469	0.110	2.5e-5	1.00
Adrenal Gland	233	0.108	0.02	0.947
Liver	208	0.106	0.03	0.892
Artery - Tibial	584	0.104	7.0e-7	1.00
Pituitary	237	0.0960	0.03	0.927
Prostate	221	0.0892	0.02	0.921
Brain - Spinal cord (cervical c-1)	126	0.0853	0.1	0.667
Nerve - Tibial	532	0.0851	1.8e-4	0.987
Brain - Putamen (basal ganglia)	170	0.0839	0.08	0.808
Brain - Substantia nigra	114	0.0827	0.3	0.735
Pancreas	305	0.0810	0.2	0.698
Thyroid	574	0.0738	6.4e-4	0.913
Breast - Mammary Tissue	396	0.0635	0.02	0.410
Artery - Coronary	213	0.0591	0.1	0.423
Uterus	129	0.0575	0.4	0.519
Vagina	141	0.0558	0.4	0.593
Cells - Cultured fibroblasts	483	0.0430	0.05	0.00
Brain - Nucleus accumbens (basal ganglia)	202	0.0385	0.3	0.159
Small Intestine - Terminal Ileum	174	0.0380	0.4	0.180
Artery - Aorta	387	0.0279	0.4	0.0280
Brain - Caudate (basal ganglia)	194	0.0268	0.5	0.0530
Brain - Amygdala	129	-0.00320	1	0.173
Brain - Anterior cingulate cortex (BA24)	147	-0.0166	0.8	0.0620
Ovary	167	-0.0206	0.7	0.0490
Testis	322	-0.0214	0.8	0.236
Brain - Hypothalamus	170	-0.0611	0.2	0.00
Kidney - Cortex	73	-0.0628	0.6	0.231
Brain - Hippocampus	165	-0.0899	0.03	0.00
Brain - Cortex	205	-0.118	0.02	0.00
Brain - Cerebellar Hemisphere	175	-0.119	0.01	0.00
Brain - Frontal Cortex (BA9)	175	-0.135	0.006	0.00
Brain - Cerebellum	209	-0.263	1.1e-5	0.00

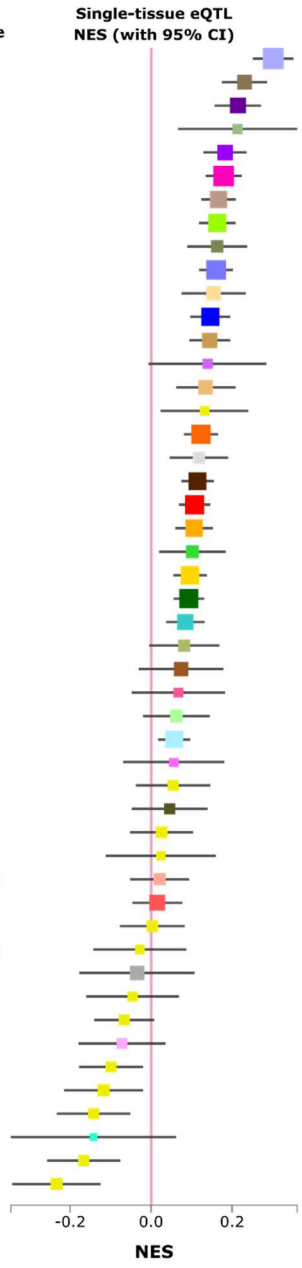


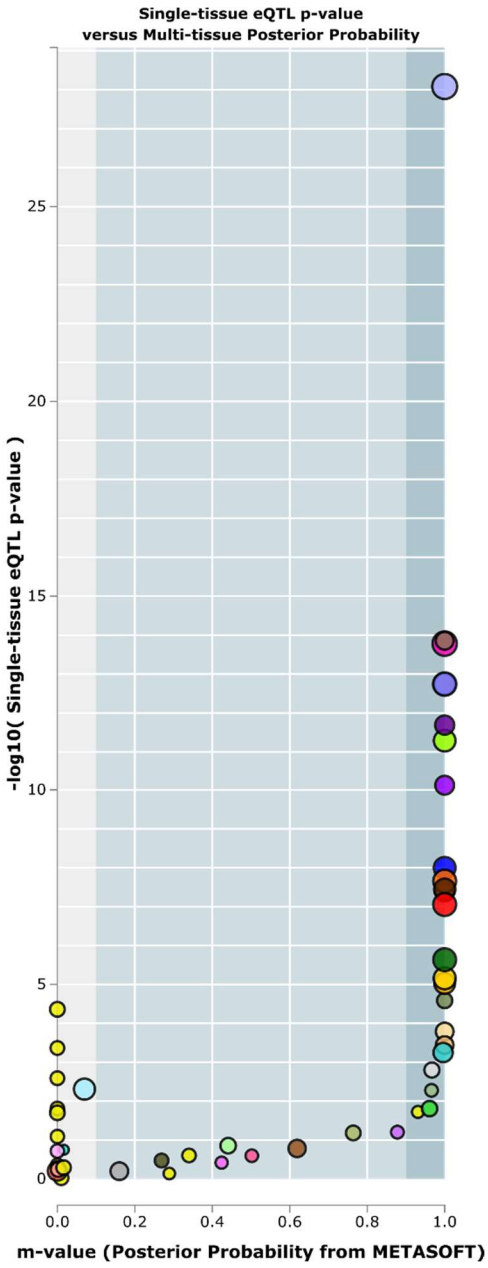


Supplementary Figure 5 – rs17229081 as eQTL of *PARK7* in multiple tissues.

8.12 eQTL rs4908724

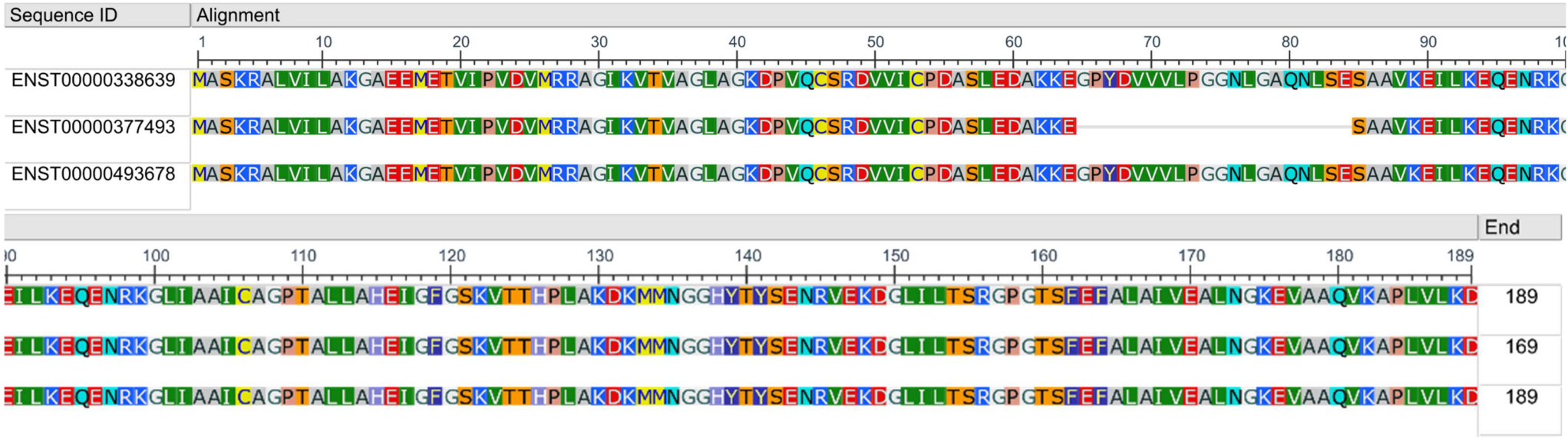
Tissue	Samples	NES	p-value	m-value
Muscle - Skeletal	706	0.302	8.2e-29	1.00
Esophagus - Gastroesophageal Junction	330	0.231	1.4e-14	1.00
Heart - Left Ventricle	386	0.215	2.0e-12	1.00
Minor Salivary Gland	144	0.214	0.005	0.966
Heart - Atrial Appendage	372	0.183	7.6e-11	1.00
Whole Blood	670	0.179	1.7e-14	1.00
Esophagus - Muscularis	465	0.167	1.8e-13	1.00
Lung	515	0.164	5.1e-12	1.00
Spleen	227	0.163	2.6e-5	1.00
Skin - Sun Exposed (Lower leg)	605	0.161	1.8e-13	1.00
Stomach	324	0.154	1.6e-4	1.00
Skin - Not Sun Exposed (Suprapubic)	517	0.146	1.0e-8	1.00
Colon - Transverse	368	0.145	4.2e-8	1.00
Cells - EBV-transformed lymphocytes	147	0.139	0.06	0.878
Colon - Sigmoid	318	0.135	3.6e-4	1.00
Brain - Spinal cord (cervical c-1)	126	0.132	0.02	0.931
Adipose - Subcutaneous	581	0.123	2.2e-8	1.00
Prostate	221	0.118	1.6e-3	0.967
Esophagus - Mucosa	497	0.115	3.6e-8	1.00
Artery - Tibial	584	0.108	8.8e-8	1.00
Adipose - Visceral (Omentum)	469	0.106	9.4e-6	1.00
Adrenal Gland	233	0.102	0.02	0.961
Nerve - Tibial	532	0.0964	7.0e-6	1.00
Thyroid	574	0.0932	2.3e-6	1.00
Breast - Mammary Tissue	396	0.0847	5.6e-4	0.996
Liver	208	0.0819	0.07	0.764
Pancreas	305	0.0741	0.2	0.619
Vagina	141	0.0674	0.3	0.502
Pituitary	237	0.0624	0.1	0.441
Cells - Cultured fibroblasts	483	0.0570	4.9e-3	0.0700
Uterus	129	0.0560	0.4	0.424
Brain - Putamen (basal ganglia)	170	0.0544	0.2	0.340
Small Intestine - Terminal Ileum	174	0.0459	0.3	0.269
Brain - Nucleus accumbens (basal ganglia)	202	0.0261	0.5	0.0160
Brain - Substantia nigra	114	0.0241	0.7	0.289
Artery - Coronary	213	0.0211	0.6	0.00200
Artery - Aorta	387	0.0153	0.6	0.00
Brain - Caudate (basal ganglia)	194	0.00300	0.9	0.0100
Brain - Amygdala	129	-0.0278	0.6	0.00800
Testis	322	-0.0346	0.6	0.160
Brain - Anterior cingulate cortex (BA24)	147	-0.0456	0.4	0.00
Brain - Hippocampus	165	-0.0663	0.08	0.00
Ovary	167	-0.0718	0.2	0.00
Brain - Hypothalamus	170	-0.0989	0.02	0.00
Brain - Cortex	205	-0.117	0.02	0.00
Brain - Cerebellar Hemisphere	175	-0.142	2.6e-3	0.00
Kidney - Cortex	73	-0.142	0.2	0.0170
Brain - Frontal Cortex (BA9)	175	-0.166	4.3e-4	0.00
Brain - Cerebellum	209	-0.233	4.4e-5	0.00





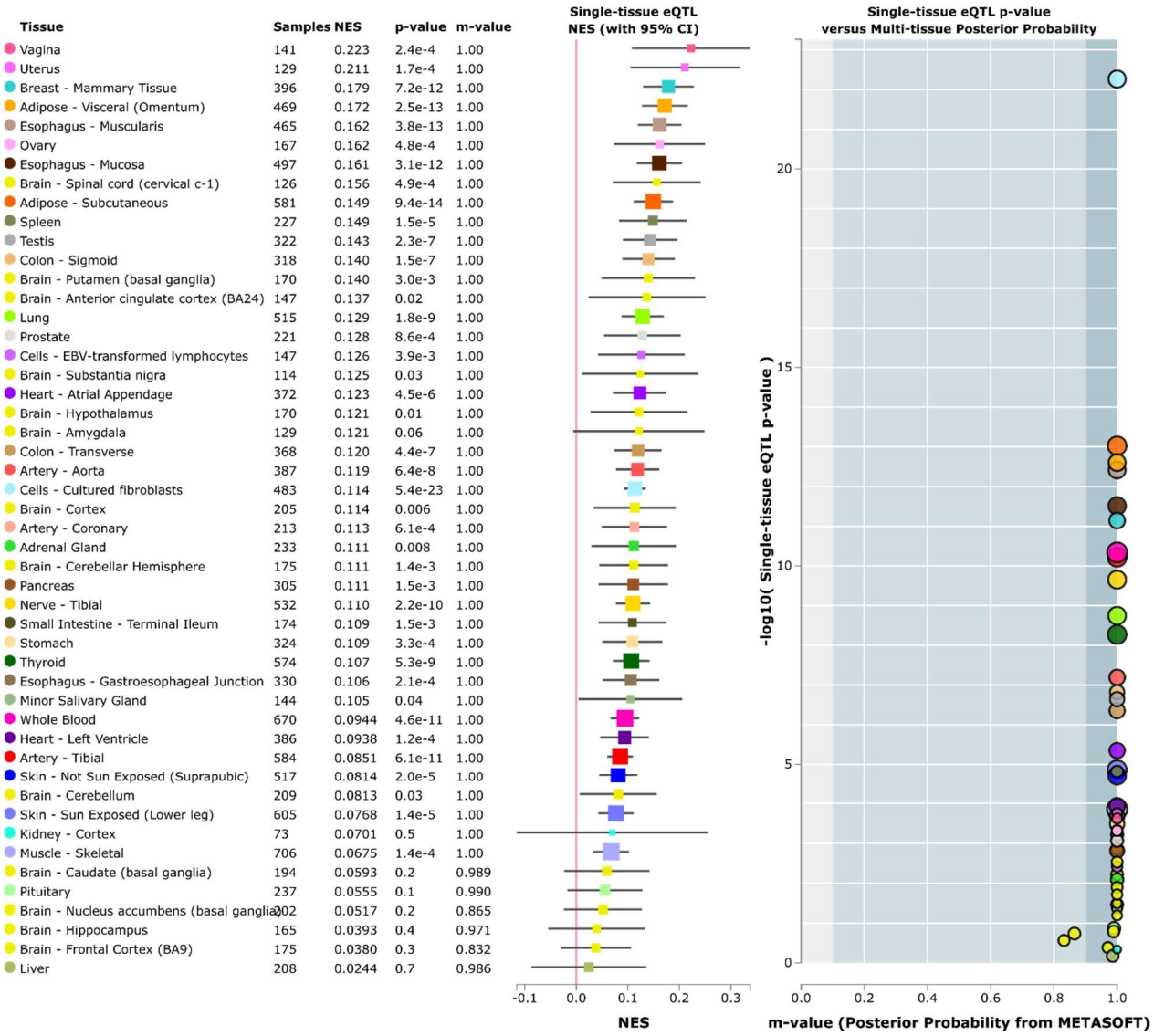
Supplementary Figure 6 – rs4908724 as eQTL of *PARK7* in multiple tissues.

8.13 DJ-1 Protein sequence from each isoform



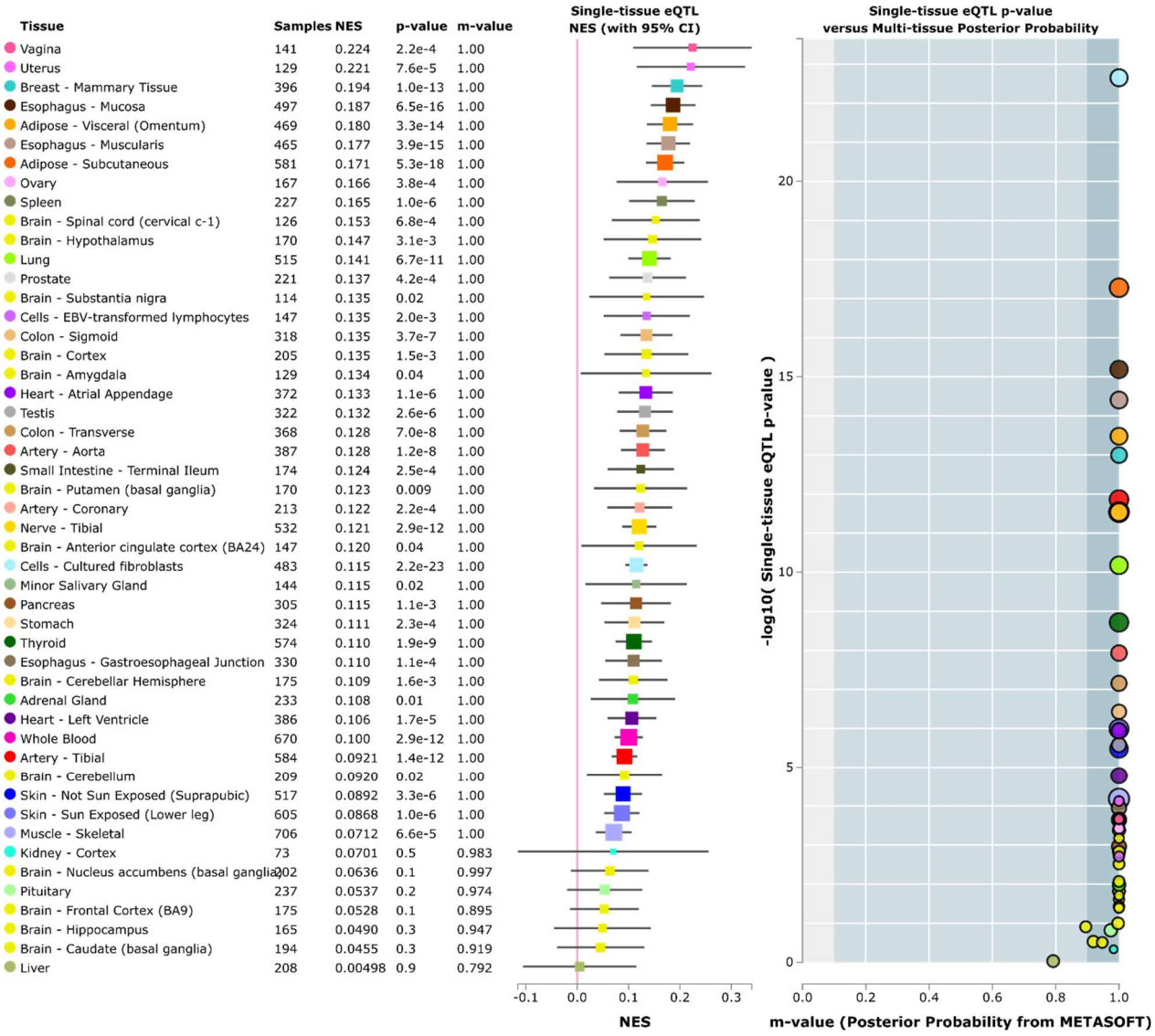
Supplementary Figure 7 – Aminoacids sequence of translated isoforms of interest

8.14 eQTL rs6591195



Supplementary Figure 8 – rs6591195 as eQTL *Banf1* in multiple tissues.

8.15 eQTL rs9735063



Supplementary Figure 9 – rs9735063 as eQTL of *Banf1* in multiple tissues

8.16 rs12898397 reference and alternative sequence

Table 3 – rs12898397 flanking sequences

Upstream	Variant	Downstream
GAGAGAGCAG ACATACACCA GCAGGGGGGG ACGACAGCCA CAAGCAGAGA GCAACAGAAC CCACAGACCC TAGCCCCAGC GTGGCCCCCA TGTGCCCACC	T (reference)	TGACCTGCTC CTTCAAGTAT TCAGCTCGGG CCATGAGGTT CTGAACCTGC CAAGGAAAGA TATGCCCTTG GGTGTGGGGG AGAGTGGCGG GGGGTGGGGT
	C (alternative)	

8.17 NetGene2 splice site prediction of rs12898397 reference allele

```
********************* NetGene2 v. 2.4 ************************


The sequence: sequence1 has the following composition:

Length: 201 nucleotides.
24.4% A, 27.9% C, 33.3% G, 14.4% T, 0.0% X, 61.2% G+C

Donor splice sites, direct strand
---------------------------------
   No donor site predictions above threshold.

Donor splice sites, complement strand
-------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'      exon intron    3'
       105          97       0     -        0.71       GAAGGAGCAG^GTCAAGGTGG
        99         103       0     -        0.91       GCAGGTCAAG^GTGGGCACAT H

Acceptor splice sites, direct strand
------------------------------------
   No acceptor site predictions above threshold.

Acceptor splice sites, complement strand
----------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'    intron exon      3'
       148          54       0     -        0.00       TCCTTGGCAG^GTTCAGAACC

-------------------------------------------------------------------------------
```

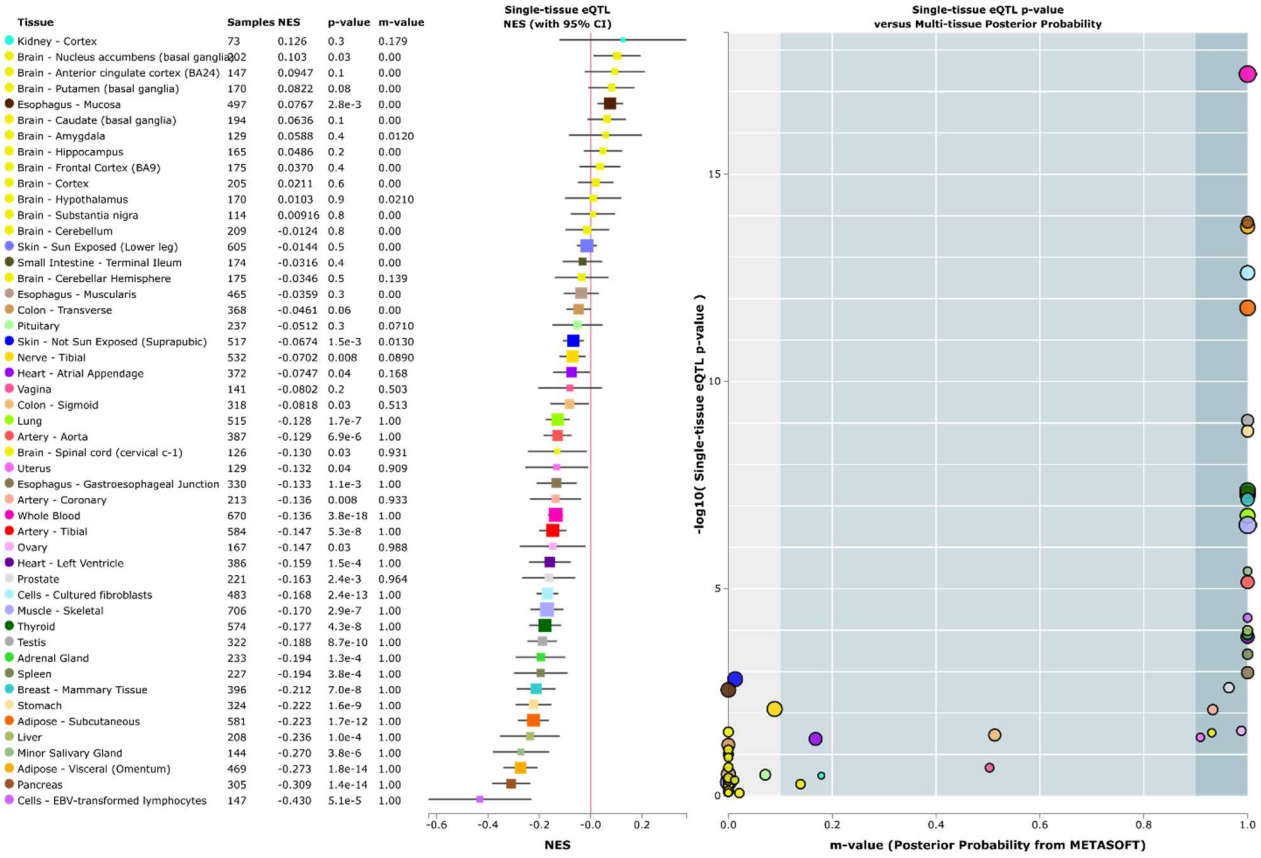
Supplementary Figure 10 - NetGene2 splice site prediction using genomic sequence surrounding rs12898397 reference allele T.

8.18 NetGene2 splice site prediction of rs12898397 alternative allele

```
********************* NetGene2 v. 2.4 ************************


The sequence: sequence1 has the following composition:

Length: 201 nucleotides.
24.4% A, 28.4% C, 33.3% G, 13.9% T, 0.0% X, 61.7% G+C

Donor splice sites, direct strand
---------------------------------
   No donor site predictions above threshold.

Donor splice sites, complement strand
-------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'      exon intron    3'
       105          97       0     -        0.82       GAAGGAGCAG^GTCAGGGTGG

Acceptor splice sites, direct strand
------------------------------------
   No acceptor site predictions above threshold.

Acceptor splice sites, complement strand
----------------------------------------
   pos 3'->5'  pos 5'->3'  phase strand  confidence  5'    intron exon      3'
       148          54       0     -        0.00       TCCTTGGCAG^GTTCAGAACC

--------------------------------------------------------------------------------
```

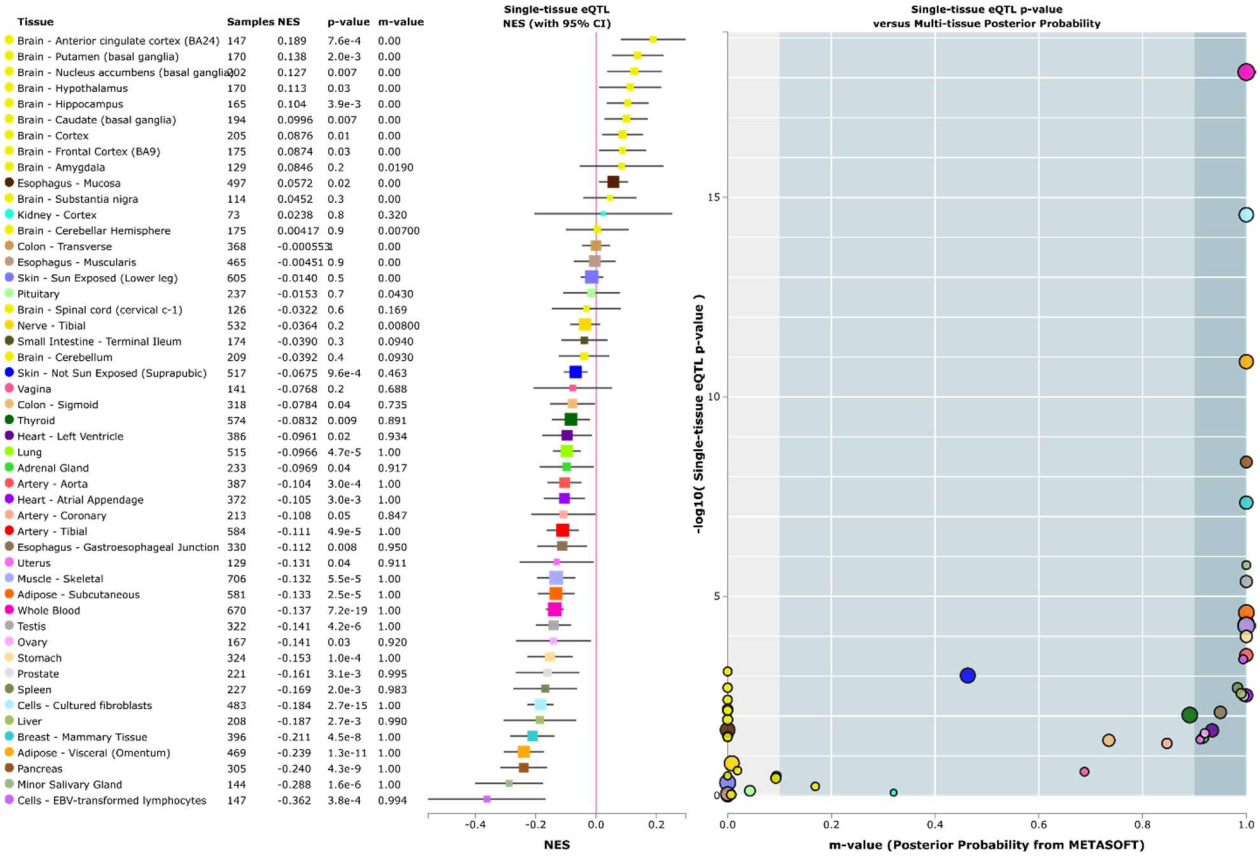
Supplementary Figure 11 – NetGene2 splice site prediction using genomic sequence surrounding rs12898397 alternative allele C.

8.19 eQTL rs12591513



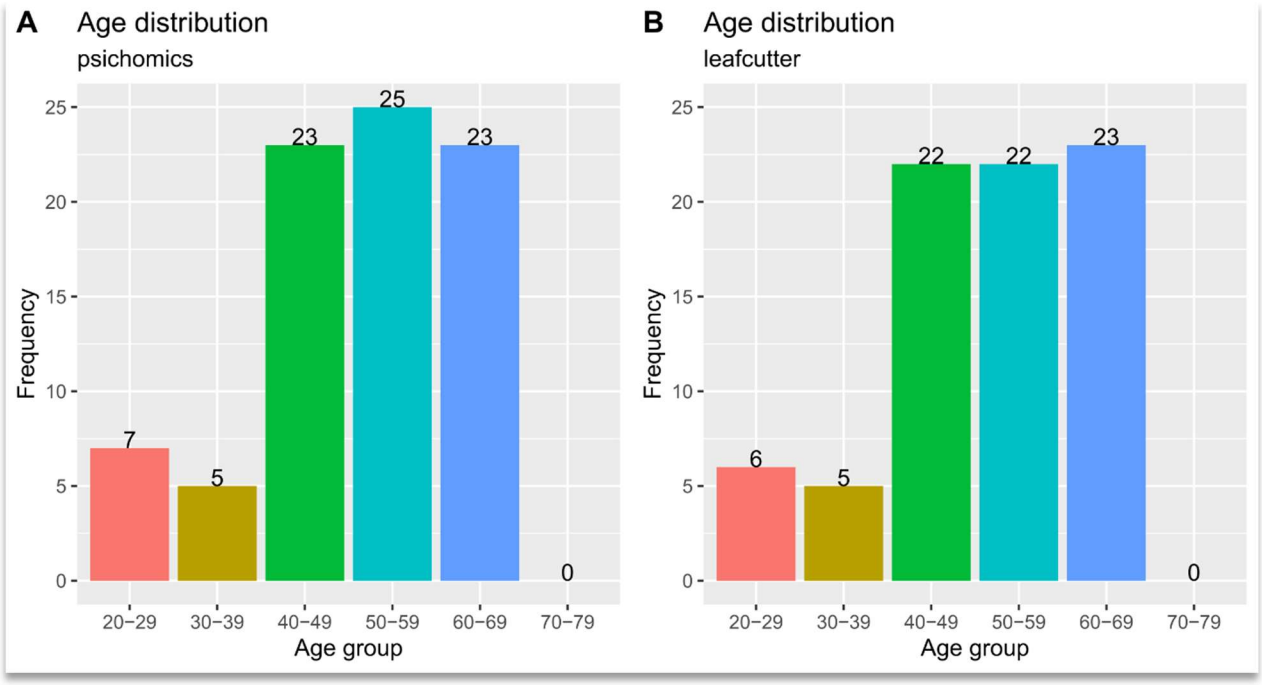
Supplementary Figure 12 – rs12591513 is an eQTL of *ULK3* in multiple tissues

8.20 eQTL rs12898397



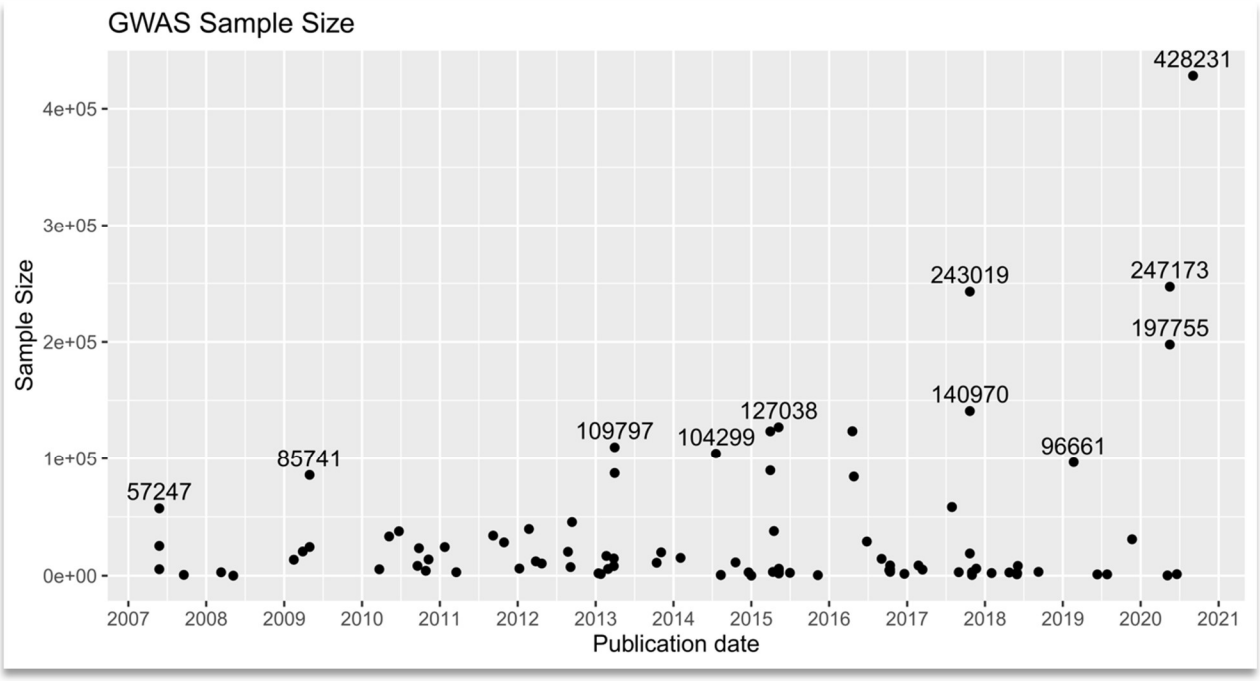
Supplementary Figure 13 – rs12898397 is an eQTL for *ULK3* in multiple tissues.

8.21 Age distribution of RNA-seq samples used to calculate PSI by each tool



Supplementary Figure 14 – Age distribution of samples used in sQTL analysis, presented by analysis tool

8.22 Date vs GWAS_sample_size



Supplementary Figure 15 – GWAS sample size through time. An increase in sample size has been observed over the years.