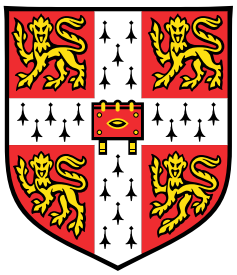
Using single-cell RNA-seq to assess the effect of common genetic variants on gene expression during development



Anna Cuomo

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Peterhouse College

September 2020

"Nothing in life is to be feared, it is only to be understood."

Marie Curie

Ai miei genitori, Daniela e Renato.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Contribution sections at the beginning of each chapter. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Anna Cuomo
September 2020

ACKNOWLEDGEMENTS

"It takes a village (to raise a child)."

African proverb

First and foremost, I would like to acknowledge my co-supervisors Oliver Stegle and John Marioni, for their guidance and support through my entire PhD. I cannot emphasise enough how much I have learnt in the last four years, and how much I have grown, as a person and as a scientist. I am thankful to EMBL-EBI and the University of Cambridge for providing an outstanding environment to perform my research - I have been truly spoilt. I am thankful to my thesis advisory committee, Eileen Furlong and Duncan Odom, and to my collaborators, Ludovic Vallier, Mariya Chhatriwala, Daniel Gaffney, Natsuhiko Kumasaka and Florian Merkle.

The help and guidance of several people was instrumental in making my PhD productive and exciting. A special thank you goes to Marc Jan Bonder and Daniel Seaton, for being the most amazing colleagues and for their continued presence and support, at work and outside, especially in the second part of my PhD. Thank you for the morning coffees, the board games, and just for always being there for me.

I would also like to thank Paolo Casale, who in the first few months of my PhD introduced and explained to me key concepts within the field of Statistical Genetics; Davis McCarthy for introducing me to and patiently helping me with all things R and single cell RNA-seq related; and Nils Eling for helping me navigate life as an EBI and Cambridge PhD student in those early days. I would also like to thank Danilo Horta, for his availability to answer all my questions to do with software and maths. I would like to thank Julie Jerber, Ximena Ibarra-Soria, Arianne Richard, Rebecca Berrens, Na Cai, Rachel Moore and Sara Pulit for teaching me all the biology and genetics I know, and for being fantastic role models for me, whether they are aware or not! I am especially thankful to Julie with whom I have worked on one of the projects during my PhD, it was a truly wonderful collaboration and I could not have wished for a better partner. I would also like to thank all other past and present members of the Marioni group, particularly Aaron, Chris, Mike, Rebecca, Shila and Tom.

I would like to thank the following people for reading and commenting this thesis: Shila Ghazanfar, Nils Eling, Julie Jerber, Daniel Seaton, Michael Morgan, Ximena Ibarra-Soria, Paula Frampton, Kendall MacMillan, Oliver Stegle and John Marioni.

I would like to thank Nils Eling and Krishna Kumar for this amazing template. I would also like to thank my viva examiners Professor Cecilia Lindgren and Dr. Chris Wallace for reading my thesis in detail and for fruitful discussion and comments.

Finally, I would like to thank Hannah Currant and Elissavet (Elsa) Kentepozidou, for being the best companions in this PhD journey. From the very first days in Heidelberg, to wedding dress shopping, to discussing work, science, equality, pandemics and everything in between, I could not have asked for two better people to do this thing with. I would also like to thank the rest of the amazing community that are the EBI predocs, including the rest of my batch, besides Hannah and Elsa - Ricard, Umberto, Harald and Melike; as well as previous batches especially Nils, Jack, Dani, Hannah, Nadia and Claudia; and following batches, including Rachel, Martina, Conor, Jose. Special mention also to Pauline and Joanna, from EMBL Grenoble.

On a more personal note, I would like to thank my friends and family, who despite the distance are always there for me, and are my biggest cheerleaders. I would like to thank Livia, Teresa, Camilla, Sofia and Martina for having been by my side for the last 22+ years. Giulia for always believing in me and for being with me on the most important day. Nagiua and Michele, and the rest of the (extended) family, for everything. I would like to thank my parents, Daniela and Renato, for always encouraging my curiosity and for making me who I am today. “Grazie mamma, grazie pa!”, cit. Last but not least, I would like to thank Kendall, for always believing in me and for being the best partner anyone could ask for.

ABSTRACT

Over the last fifteen years, genome-wide association studies (GWAS) have been used to identify thousands of DNA variants associated with complex traits and diseases, by exploiting naturally occurring genetic variation in large populations of individuals. More recently, similar approaches have been applied to RNA sequencing (RNA-seq) data to find variants associated with expression level, called expression quantitative trait loci (eQTL). Recent advances in experimental techniques have provided an unprecedented opportunity to measure gene expression at the single cell level, and the chance to study cellular heterogeneity. This represents a remarkable advance over traditional bulk sequencing methods, particularly to study cell fate commitment events in development. The challenge of studying early human development is partially overcome by advances in stem cell technologies. In particular, induced pluripotent stem cells (iPSCs) and cells derived therefrom represent a fantastic system to study development *in vitro*. In this thesis, I investigate the computational challenges of using single cell expression profiles to perform expression quantitative trait locus (eQTL) mapping, and provide suitable approaches for the identification of cell type and context-specific eQTL using single cell expression profiles. I further explore the application of such methods across a range of human iPSC-derived cell types, using data from the human induced pluripotent stem cell initiative (HipSci) project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ASE allele-specific expression

CPM counts per million (reads)

DNA deoxyribonucleic acid

eQTL expression quantitative trait locus
ES embryonic stem
ESC embryonic stem cell

FACS fluorescence-activated cell sorting
FDR false discovery rate

GTEx genotype-tissue expression
GWAS genome-wide association study
GxE genotype-environment

hESC human embryonic stem cell
HGP Human Genome Project
HipSci human induced pluripotent stem cell initiative
HVGs highly variable genes

iPS induced pluripotent stem

iPSC induced pluripotent stem cell
IVF *in vitro* fertilisation

LD linkage disequilibrium
LMM linear mixed model
LRT likelihood ratio test

MAF minor allele frequency
MLE maximum likelihood estimator
mRNA messenger RNA

OLS ordinary least squares

PBMC peripheral blood mononuclear cell
PC principal component
PCA principal component analysis

QC quality control
QQ quantile-quantile
QTL quantitative trait locus

RNA ribonucleic acid
RNA-seq RNA sequencing

SCNT somatic-cell nuclear transfer
scRNA-seq single-cell RNA sequencing
SNP single nucleotide polymorphism

TF transcription factor
tRNA transfer RNA
TSS transcription start site

UMI unique molecular identifier

1

Introduction

[We might have just found] "The secret of life."

Francis Crick, 1953

1.1 From peas to GWAS

The observation that human traits are heritable is evident, often visible by eye. Every one of us has been told that they have their mother's eyes, their father's height, their grandfather's nose, etc. Similarly, many diseases "run in the family": for example diabetes and some types of breast cancer are recurrent from generation to generation. In fact, for some conditions, family history can be one of the most reliable diagnostic tools. Describing the mechanisms by which we acquire traits, and the extent to which traits are heritable is at the core of the science we call genetics, and is a question that has occupied scientists for years.

I use this section to provide a brief historical overview and highlight the key scientific discoveries and technological advances in the field of genetics - from Mendel's experiments on peas in the 19^{th} century to the announcement of the completion of the human genome sequence in 2000 (**sections 1.1.1-1.1.6**). Next, I use **section 1.1.7** to discuss genome-wide association studies (GWAS), a statistical method aimed at identifying associations between common genetic variants and complex traits and diseases. Finally, in **section 1.1.8**, I describe expression quantitative trait loci (eQTL) mapping, where GWAS-like approaches are applied to find variants associated with expression level in order to gain insight into the molecular and regulatory role of trait- and disease-associated variants.

1.1.1 Principles of (Mendelian) Inheritance

In 'On the Origin of Species' [1], Charles Darwin proposed the theory of evolution, which is based on the assumption that natural variation between individuals provides differential reproductive advantages, and that this variation can be inherited from one generation to the next. This theory explains the adaptation of a species to its environment and the consequent development of new species, yet the mechanisms by which such variation occurs and the modes of inheritance were not described. In the words of Darwin himself in 'On the Origin of Species': "our ignorance of the laws of variation is profound" and "the laws governing inheritance are quite unknown".

Around the same time, the man who is now universally considered to be the father of genetics began conducting experiments to tackle exactly this problem. He was actually not a scientist but a friar, by the name of Gregor Mendel, and in 1853 he performed the now famous experiments on inheritance in peas. At St. Thomas' abbey, in Brno, Moravia (then part of the Austro-Hungarian empire), Mendel meticulously studied seven different traits of the plants (plant height, pod shape and colour, seed shape and colour, and flower position and colour), each of which segregated on one of the plant's seven chromosomes. For example, seeds were either yellow or green, wrinkled or round. For seven years, Mendel followed generations and generations of pea plants and noted that some traits occurred far more often than others. For example, when crossing a plant with round seeds and one with wrinkled seeds, the offspring (F_1) always had round seeds: Mendel called the round seed trait the 'dominant' trait. However, the wrinkled seed trait which had seemingly vanished in the first filial generation would appear again, in the second generation (F_2), in a 1:3 ratio between wrinkle-seed to round-seed plants. Somehow, this 'recessive' trait was being passed on, remaining hidden when overpowered by the dominant trait, but not forgotten. In 1866, Mendel published his experiments and results in 'Versuche über Pflanzenhybriden' (Experiments in Plant Hybridisation, [2]). In it, he proposes what will be called the Mendelian Laws of Inheritance: i) the Law of Independent Segregation (every individual contains two factors for each trait, one of which is passed on to its offspring at random), ii) the Law of Independent Assortment (traits are inherited independently of each other) and iii) the Law of Dominance (recessive alleles will be masked by dominant alleles and the trait corresponding to the dominant allele will be observed) [2]. The publication received almost no attention, but the Laws of Inheritance described therein built the foundations of modern genetics.

Mendel and Darwin never met, and died remaining unaware of each other's theories. Mendel's research remained completely unknown for decades, until at the turn of the century, in 1900,

four botanists (the Austrian Erich von Tschermak, the Dutchman Hugo de Vries, the German Carl Correns and the American William Jasper Spillman) independently rediscovered his work and validated his findings, officially beginning the modern age of genetics. Around the same time, the British geneticist William Bateson set out to make Mendel's work accessible to scientists that were not proficient in Mendel's native language, German. It was in a 1901 lecture at the Royal Society's evolution committee that Bateson described Mendel's principles and introduced some fundamental terms of genetics that we still use today, including 'allelomorph' (allele), 'zygote', 'homozygous', 'heterozygous' and, even the word 'genetics' itself (**Box 1**). Bateson translated Mendel's original papers on the Laws of Inheritance into English and published them, allowing Mendel's work to become known in the greater scientific world, more than 40 years after their original publication [3].

An important step towards reconciling Darwin's theory of evolution with Mendel's laws of inheritance was made in 1902, when Theodor Boveri showed, in sea urchin, that different chromosomes contained different hereditary material and that organisms required a full set of chromosomes to function. In 1903, Walter Sutton published a paper proposing how these principles, together with the random segregation of paternal and maternal chromosomes during gamete formation (which he studied in grasshoppers) could form the molecular basis for Mendel's Laws of Inheritance [4]. Importantly, he also noted how the number of traits was much larger than that of chromosomes, which meant that some traits had to be located on the same chromosome and be transmitted together.

1.1.2 Genetic Linkage and the birth of modern genetics

In 1908, the American geneticist Thomas Hunt Morgan set out to confirm (or better, disprove) Mendel's theories using a model organism that generates new offspring much quicker than pea plants. It was *Drosophila melanogaster*, the fruit fly. In his famous 'fly room' at Columbia University, thousands of experiments were performed on flies. Flies with known phenotypes (for example red or white eyes) were put in jars to mate, and the traits of the progeny were recorded. Through the key observations that some traits appeared to be sex-linked and that some other traits were co-occurring more often than expected by chance, Morgan theorised that 'markers' responsible for particular traits were positioned on chromosomes, like beads on a string. These markers (or genes, **Box 1**), when close together on a chromosome, were more likely to be passed on to the next generation. Morgan had described the concept of genetic linkage and essentially hypothesised the phenomenon of crossing over (exchange of paternal and maternal chromosomal material during meiosis [5]). In 1913, his student, Alfred Sturtevant, gathered all the data collected and developed the first genetic map, showing the

position of the fruit fly's known markers relative to each other in terms of recombination frequency [6]. Sturtevant would go on to call the unit of genetic linkage a centimorgan (cM), in honour of his mentor.

Box 1: Genetic terms & their origin

- **Alleles** (originally allelomorphs) were defined by Bateson as the units of inheritance described by Mendel [3].
- **Homozygote** and **heterozygote** were also used by Bateson to describe individuals carrying the same or different alleles.
- The word **gene** as a term for the Mendelian factors or units of inheritance was introduced by Danish botanist Johannsen [7].
- Johannsen also introduced the terms **phenotype**, as the outward appearance of an individual, and **genotype**, as their genetic traits.
- The terms **polygenetic** (today more often simply polygenic), for traits that are governed by multiple genes, and **pleiotropic**, for genes that affect multiple, seemingly unrelated, phenotypes also made their first appearance at that time.
- A **pedigree**, from the French *pied de grue* (crane's foot), is a diagram that depicts the biological relationships between related individuals. It is often used to look at the transmission of genetic disorders.

The Mendelian-chromosome theory, first proposed by Boveri and Sutton [4] and then elaborated and expanded by Morgan and his students [8], described chromosomes as the (paired) units of heredity that Mendel had described in his laws, and was widely accepted by scientists by the 1930s. In 1933, Morgan received the Nobel Prize in Physiology or Medicine "for his discoveries concerning the role played by the chromosome in heredity" [9].

However, the mechanisms of heredity and the physical molecule responsible for it were still unknown. The concept of the 'gene' existed, but it was an abstract entity. Most people in fact believed that proteins were the carriers of genetic material. In 1944, Erwin Schrödinger, an Austrian-Irish physicist perhaps better known for his contributions to quantum mechanics, published 'What is Life?' where he introduced the idea that genetic material may be stored as some sort of a 'code', a concept borrowed from Information Theory [10]. He had provided a theoretical physical description of the mechanism of 'storage' of genetic material [11]. Still,

as did most scientists at the time, Schrödinger bet on proteins as the responsible molecule. The other candidate, deoxyribonucleic acid (DNA), had been referred to as the 'stupid molecule', a molecule with a chemical structure far too simple to be able to explain the complexity of life. In fact, DNA consists of only four building blocks, often referred to by their initials: adenine (A), thymine (T), cytosine (C) and guanine (G) [12]. Proteins remained the most likely molecule responsible for carrying genetic information until 1944, when Oswald Avery and colleagues at the Rockefeller Institute in New York demonstrated experimentally that it had to be DNA. Avery, along with his co-workers Colin MacLeod and Maclyn McCarthy, performed an experiment in *Streptococcus pneumoniae*, where he removed various organic compounds from the bacteria, and observed whether it could still transform. Only upon treating the bacteria with an enzyme that removed DNA, did the bacteria stop transforming [13].

1.1.3 The double helix

The question of the physical structure of DNA remained unsolved until 1953, when a team of scientists at last proposed its structure. Key members of this team were Francis Crick, James Watson, Rosalind Franklin, Maurice Wilkins and Erwin Chargaff. Jim Watson had had a fascination for the structure of DNA, and had been studying it for years, starting in his native Chicago, then during his PhD in Indiana (under the supervision of future Italian Nobel Prize laureate Salvador Luria). Eventually, he ended up in Cambridge, UK, at the Cavendish laboratory, which was then directed by Australian-born British X-ray crystallographer Sir Lawrence Bragg. There, he met Francis Crick, 12 years his senior, a British physicist turned biologist. Crick had taught himself the mathematical theory of X-ray crystallography and had worked on determining the most stable helical conformation of amino acid chains in proteins, the alpha helix, only to be beaten to its solution by American chemist Linus Pauling [14]. Watson and Crick set out to obtain a model for the structure of DNA, building on Crick's experience and rigor, and Watson's intuition. Friend and collaborator of the pair was Maurice Wilkins, New Zealand-born British physicist at King's College London, who had extensively studied X-ray diffraction patterns. His colleague, Rosalind Franklin, had perfected the technique to produce X-ray crystallography images of the DNA and instructed her assistant, Raymond Gosling, to take the most precise image to date, the now famous 'Photo 51'. To Watson and Crick's eyes, Photo 51 (which Wilkins had shared without Franklin's knowledge), looked without a doubt like the footprint of a helix. The last piece of the puzzle came from a discovery that Austro-Hungarian Erwin Chargaff made, at Columbia University. He observed that globally the numbers of As and Ts in DNA were roughly the same, as were the numbers of Cs and Gs. This provided the idea that bases would be paired up and facing inwards in the

double helix, As with Ts, Cs with Gs, ensuring that the covalent bonds would be always of the same length, which would keep the helix stable. In 1953, Watson and Crick published 'Molecular structure of nucleic acids' [15]. Their work showed how the four nucleotide bases (A, T, C, G) formed "two helical chains each coiled round the same axis" [15] spelling out what Crick called "the secret of life". For this discovery, Crick, Watson and Wilkins won the 1962 Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material" [16]. Rosalind Franklin, who had played a critical role in the discovery, had died four years prior of ovarian cancer, and her contribution went largely unrecognised at the time.

1.1.4 Biometrics

While Mendel was studying the inheritance of traits in peas, Boveri studied sea urchins, Morgan fruit flies, Avery bacteria, and long before Crick, Watson and Franklin proposed a structure for DNA (working on squid), ever since Darwin's theories others were trying to quantify inheritance in the context of human traits. One such investigator was Francis Galton, a half-cousin of Darwin's, who was interested in mathematically describing and analysing Darwin's evolutionary concepts. In particular, Galton was interested in the effects that evolution had had on humans and how its effects could be used to better the human race. Galton was one of the founders of the study of biometrics, which attempted to measure and estimate the heritability of human traits such as height and intelligence. Linked to these efforts, and on a less honourable note, Galton was also the founder of eugenics, a theory for which genetics should be used as a tool to force evolution's hand by encouraging mating of individuals considered to have especially desirable qualities and by eliminating or preventing reproduction of individuals considered faulty. Eugenics theories are linked to one of the most horrifying pages of human history, motivating forced sterilisations of the 'unfit' in the United States in the 1920s and 1930s and of course having been used as justification for the racial policies of Nazi Germany. Nevertheless, some of the concepts and methods Galton developed during these studies are still fundamental to genetics today [17]. Among others, Galton introduced the concepts of correlation, regression toward the mean and the regression line, which he used to compare the heights of children to those of their parents. Galton and his student, mathematician Karl Pearson, worked together to make several more important contributions to statistics. For example, Pearson described the concepts of the p value and the chi squared (χ^2) test [18] and proposed principal component analysis (PCA) [19][1].

[1] later independently developed and named by the American statistician and economist Harold Hotelling.

1.1.5 Towards quantitative genetics

By cross-breeding Drosophila lines and performing genetic mapping, Morgan and his students had effectively conducted the first genotype-phenotype studies. Similar to Mendel and Bateson, the phenotypes they observed were predominantly categorical, such as the colour and shape of seeds in pea plants or the red or white-eyed phenotype in *Drosophila*. In contrast, biometricians like Galton and Pearson had mostly looked at continuous traits in humans, such as height, and believed that those could not be explained by Mendelian genetics. This controversy has been referred to as the 'Biometric-Mendelian debate'.

This debate was resolved by British statistician Ronald Fisher, who in a seminal 1918 paper showed that, if many genes affect a trait, then "the random sampling of alleles at each gene produces a continuous, normally distributed phenotype in the population" [20]. As the number of genes grows very large, the contribution of each gene becomes correspondingly smaller, leading in the limit to Fisher's famous 'infinitesimal model' [21].

In addition to showing that biometrics and Mendelianism are not contradictory but complementary, Fisher made several other contributions to the field, outlining statistical ideas and tools still used today. In particular, he published papers outlining an exact test for two-by-two contingency tables with small expectations (Fisher's exact test) [22], partial correlation coefficients [23] and the variance ratio, later named after him as the F statistic [24]. He also introduced the concepts of variance (as "the square root of the mean squared error") and analysis of variance (ANOVA).

Already as an undergraduate student at the University of Cambridge, Fisher published his first paper 'On an absolute criterion for fitting frequency curves' where he outlined the fundamental ideas of maximum likelihood estimator (MLE). He later extended on this work and by 1922, he had established the properties of the MLE such as consistency and minimum variability [25] that are still used today [26]. He demonstrated the utility of maximum likelihood estimation in genetics by solving a number of equations to elucidate a genetic map of eight *Drosophila melanogaster* genes based on their crossing-over frequencies [27].

Three decades later, building on Fisher's work, the American statistician Charles Henderson derived the solution of the mixed model equation [28]. Today, linear mixed models are very popular tools for several genetic analyses and are the models I use throughout this thesis (see **Chapter 2** for more details).

1.1.6 Molecular biology and technological advances

Whilst established as an official branch of science already in the 1930s, the resolution of the DNA structure, and all the discoveries that led to it, truly jump-started research in the new field of molecular biology. In the decades that followed, a critical combination of scientific discoveries and technological advances eventually led to the completion of the human genome sequence at the turn of the 21^{st} century [29] (**page 13**).

Cracking the code

The initial description of the structure of DNA in the early 1950s was a major advance for the field of molecular biology. At that time, technology did not exist for isolating a gene, determining its nucleotide sequence, or relating that sequence to the amino acid sequence of the corresponding protein [30]. Messenger ribonucleic acid (RNA) had not yet been discovered, and very little was known about protein synthesis. In 1941, George Beadle and Edward Tatum in their pioneering work with the fungus *Neurospora* had suggested a one-to-one correspondence between genes and enzymes [31]; yet how the nucleotide sequence of each gene was related to the amino acid sequence of its encoded protein was not yet understood.

In the years that followed, many fundamental discoveries driven by technological evolution helped to solve this riddle and greatly contributed to our general understanding of the function and structure of our genome, and the role of genomic variation. In 1955, Romanian-American cell biologist George Emil Palade first observed ribosomes [32], and in 1958, Francis Crick first postulated the central dogma of biology, stating that information is transmitted from nucleic acids (DNA and RNA) to proteins, but not vice versa [33]. In the following years, Sydney Brenner, François Jacob and Matthew Meselson discovered the role of messenger RNA (mRNA) in transporting genetic information from DNA in the nucleus to the protein-making ribosomes in the cytoplasm [34]. Finally, in 1961 Nirenberg deciphered the genetic code, discovering that combinations of three base pairs (called codons) code for one of 20 amino acids [35–37, 30]. Today, we know that DNA gets transcribed into RNA aided by the DNA polymerase molecule, and RNA gets translated to amino-acids making up proteins in the ribosome with the help of transfer RNA (tRNA) [12].

In addition to cracking the genetic code, many other advances were made around this time and in the following years, contributing to the knowledge of the molecular machinery we have today (**Fig. 1.1**).

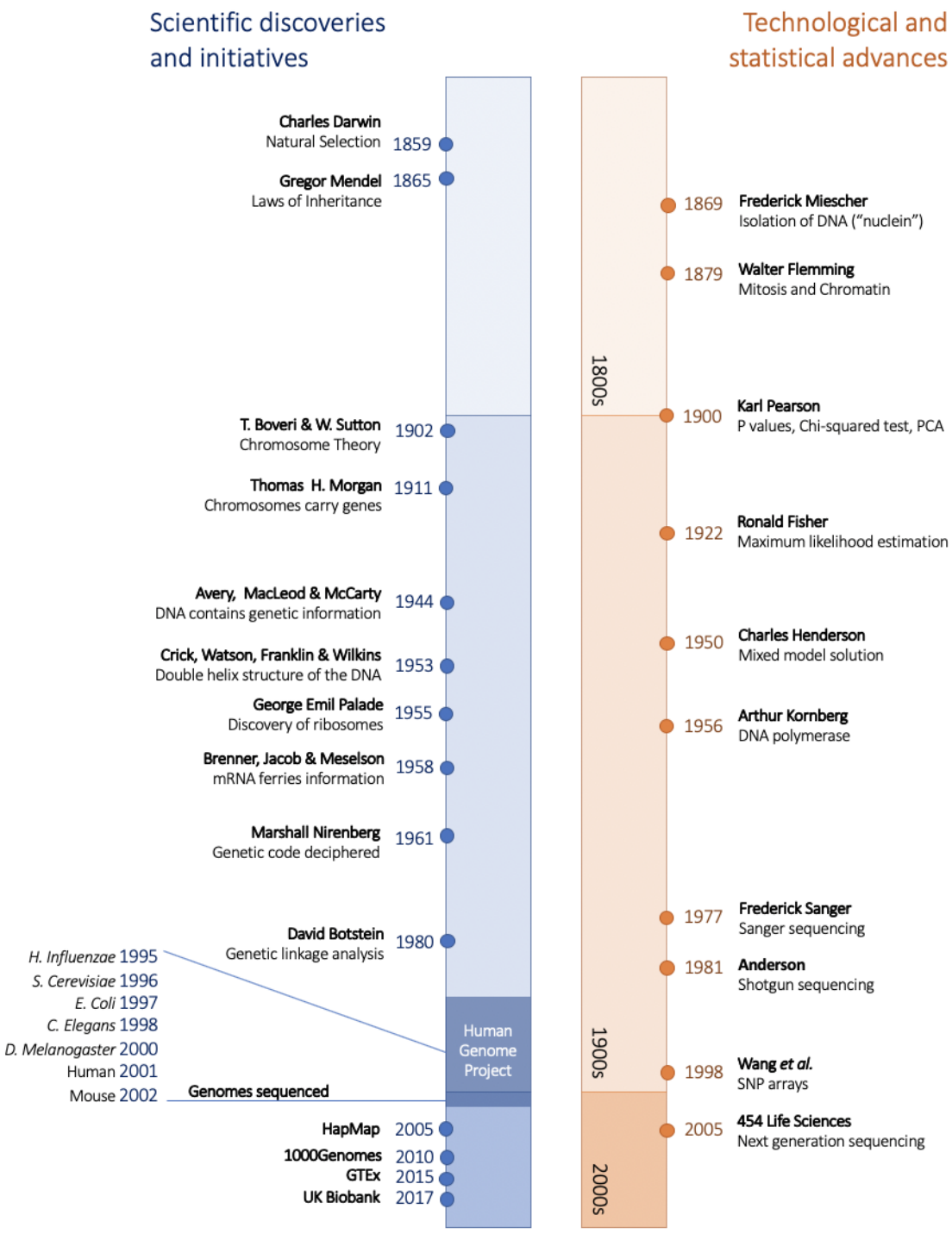


Fig. 1.1: 150 years of genetics.
A number of scientific discoveries, in combination with key advances in technology and statistical modelling, have led to the identification of thousands of genetic variants which are associated to complex and molecular traits [29]. Several fundamental contributions have been made, from Mendel's peas to the structure of DNA, to large databases cataloguing genetic variation of hundreds of thousands of individuals. Here, I have attempted to highlight the key events that led to today's field of quantitative genetics in the GWAS and post-GWAS era.

Sequencing DNA

In terms of technology, one major leap in our understanding of the biological basis of genetic variation was the development of DNA sequencing, which allowed the nucleotide sequence of a DNA segment to be determined. In the mid 1970s, two different DNA sequencing techniques were independently developed: the well known and still used Sanger chain-termination method [38, 39] and a chemical sequencing method that has been almost forgotten, known as Maxam–Gilbert sequencing [40]. The former, named after its developer Frederick Sanger, relies on *in vitro* DNA replication by DNA polymerase, which had been isolated some twenty years prior by Arthur Kornberg and colleagues [41]. The selective incorporation of 'chain-terminating' dideoxynucleotides (ddNTPs)[2] alongside regular nucleotides during DNA synthesis, and the use of polyacrylamide gel electrophoresis to separate the resulting fragments, allows the sequence of an input DNA fragment to be determined.

Sanger sequencing eventually became the standard for DNA sequencing, and subsequent innovations led to the development of automatic sequencing machines able to sequence DNA fragments of about one kilobase (kb) in length. For sequencing longer stretches of DNA, a novel strategy named 'shotgun sequencing' was developed [42, 43]. In shotgun sequencing, the long DNA of interest is randomly broken up into shorter DNA fragments which are cloned and sequenced separately. The occurrence of overlapping DNA fragments allows for the *in silico* reconstruction of longer DNA fragments.

In 1995, the first genome of a living organism (the bacteria *Haemophilus influenzae*) was sequenced and assembled by shotgun sequencing [44], shortly followed by that of another bacterium, *Mycoplasma genitalium* [45]. The genomes of other model organisms were to follow in subsequent years - the budding yeast *S. cerevisiae* [46], the bacterium *E. coli* [47], the worm *C. elegans* [48], the bacterium *M. Tuberculosis* [49], the fruit fly *D. melanogaster* [50] and the flowering plant *Arabidopsis taliana* [51]. The first two human chromosomes (chromosome 22 and chromosome 21) were sequenced in 1999 and 2000, respectively [52, 53], and the first draft of the entire human genome was published in 2001 [54, 55] (see **page 13**). The mouse genome was published in 2002 [56].

[2] either ddATP, ddCTP, ddGTP or ddTTP. These nucleotides are missing a 3' -OH group, which is required for the formation of a bond between two nucleotides, causing DNA polymerase to stop the extension of DNA when a modified ddNTP is incorporated - thus the name chain-terminating sequencing.

Understanding the genetic basis of disease

In parallel, along with the increasing understanding of the molecular structure of genes and the genome, scientists began to investigate the molecular basis of disease. One of the first conditions to be linked to a genetic cause was alkaptonuria[3], which in 1902 Garrod had identified to follow the Mendelian rules of inheritance [57]. In 1956, Vernon Ingram traced the cause of a disease to a genomic alteration for the first time: he discovered that sickle cell disease is caused by a chemical change in a hemoglobin protein [58].

Until the 21st century, success in identifying genetic factors responsible for human traits and diseases was predominantly through the use of genetic linkage studies, first proposed in 1980 by David Botstein [59]. These rely on the concept of linkage, first introduced by Morgan and his students: during meiosis, physically close genes remain 'in linkage', whereas genes that are further apart are less likely to co-segregate due to recombination events. When analysing a disease segregating[4] in a family, linkage analysis involves i) identifying a genetic marker (with a known location) that is linked to the (unknown) disease gene and then ii) testing each of the nearby genes to determine the one that causes disease [60]. Using this method, the first disease gene was mapped in 1983, when Gusella and colleagues linked a gene on chromosome 4 to Huntington's disease [61]. Shortly after, in 1989, there followed a study from Francis Collins and others, identifying the genetic basis of cystic fibrosis as a single base deletion on chromosome 7 [62]. As of 2003, about 1,200 genes were linked to Mendelian traits [63].

Mendelian disorders, as the name suggests, are those that follow the Mendelian laws of inheritance. They are often called monogenic, as they are typically driven by mutations within a single gene with high penetrance (defined here as the probability of having a trait/disease given the predisposition score). Examples of such traits include X-linked muscular dystrophies, cystic fibrosis, Fanconi anaemia and phenylketonuria. Mendelian traits are typically rare in the general population but often cluster in families. Genetic linkage analyses are best powered to discover these highly penetrant monogenic disease genes [64].

In comparison, complex traits are typically common in the general population[5]. They are driven by a combination of multiple genetic risk factors across the genome (hence the name

[3] a rare recessive condition that affects mainly the joints and causes unusual pigmentation.

[4] recurring within a family but without affecting all of its members.

[5] The genetic architecture of human disease is captured on a spectrum [65], but can be broadly categorised as i) Mendelian (monogenic), ii) complex (polygenic) or iii) chromosomal (e.g. Down's syndrome).

polygenic), environmental risk factors, as well as interaction effects between genetic variants and environmental exposures (GxE). These common, multifactorial diseases have a genetic architecture that is highly distinct from that of monogenic disease. They are known to be heritable (heritability estimates for most common diseases are estimated at 40-60% [65–69]), but their full etiology contains both a genetic and an environmental component. Consequently, it is the cumulative effect of many subtle genetic variants and environmental risk factors working in concert, that trigger disease onset. Complex diseases include, for example, heart disease, type 2 diabetes, asthma, cancer, schizophrenia and Parkinson's disease.

Whilst there was some success in using linkage studies to identify genetic regions involved in common diseases[6], genetic linkage studies in general proved underpowered to detect regions with significant linkage for complex traits and diseases [72, 73]. This was in line with the 'common disease/common variant' (CD/CV) hypothesis [72, 74], which postulated that common traits are driven by genetic variation that is common in the population, i.e. multiple variants, each with low penetrance. This hypothesis was described as early as 1996, and had already suggested that family-based linkage studies would be underpowered to detect variants with modest effects [75].

Instead, it was proposed that the combined use of linkage disequilibrium (LD)[7] and population- (rather than family-) based studies would be more suitable [75, 77] - thus practically proposing the design for genome-wide association studies (GWAS) [75]. However, at the time, implementation of GWAS was not possible, for two primary reasons. First of all, the technology required to genotype thousands to millions of markers in a single experiment for the larger required sample sizes was not available [75, 78]. Secondly, the distribution and density of genetic polymorphisms across the genome, and the LD between genetic variants across different populations, were unknown.

In some sense, population-based association studies can be viewed as an extension of family-based linkage studies, in which the population studied (derived from common ancestors) acts as an extended pedigree and a much greater number of meiotic recombinations will have occurred between the analysed samples. As a consequence, LD regions are much smaller than within pedigrees of close relatives, thus requiring a more dense panel of genetic markers to be examined [79].

[6] For example BRCA1 for breast and ovarian cancer [70, 71].

[7] LD: the nonrandom allocation of alleles at nearby variants to individual chromosomes as a result of recent mutation, genetic drift or selection, manifest as correlations between genotypes at closely linked markers [76].

The Human Genome Project

In order to study genomic variation, and therefore its role in disease, it was necessary to generate a reference genome. This was the goal of the Human Genome Project (HGP), which aimed at sequencing the entire human genome. The HGP was a major breakthrough that dramatically changed the landscape of genetics, and was described by United States President Clinton as "an epic-making triumph of science and reason" [80] at the announcement of its completion. Driven by and a driver of technological breakthroughs in DNA sequencing and genotyping, the HGP was a massive international undertaking and a truly collaborative effort; sequencing and analysis took place across twenty centers in six different countries (USA, UK, France, Germany, Japan, China) and took 13 years to complete, costing approximately $2.7 billion [81]. Led in the US by then NIH director Francis Collins and by founder of Celera Genomics[8] Craig Venter, and with large contributions from the UK, in particular from the Sanger Institute directed by John Sulston (who had first sequenced the genome of *C. Elegans*), the HGP was announced as a joint US-UK statement on June 26th, 2000. After a first publication in 2001, the project was truly completed on April 25th, 2003, on the 50^{th} anniversary of the Watson and Crick paper describing the helical structure of DNA.

The HGP provided the first map (obtained from the genomes of a small number of individuals) of the $\sim$3 billion bases in the human genome [54, 83, 84], and revealed that human DNA consists of surprisingly few exons (1.1% of the genome), whereas introns cover 24% of the genome [55, 54]. Additionally, the number of genes was found to be smaller than expected, with around 30,000 being identified in 2001, and circa 21,000 genes being the latest (still debated) estimate at the time that this thesis is written [85]. Additional breakthroughs in sequencing technologies have expanded and refined the reference genome, which now captures more than 92% of the genome and provides a landscape of its genes [81].

With a complete map of the human genome in place, genetic variants could now be identified as those bases discovered in an individual that did not match the (reference) base annotated in the human genome map. Common variants, i.e. variants with minor allele frequency (MAF)[9] larger than 5%, were called single nucleotide polymorphisms (SNPs). Previous studies had estimated that approximately 0.1% (1 base per 1,000) of an individual's genome was a polymorphism [87–89].

[8] The company Celera Genomics was formed in May 1998, with the objective of sequencing much of the human genome in three years [82].

[9] The frequency of an allele at a genetic locus is the proportion of chromosomes in the study sample that carry that allele [86]. For a biallelic variant (a variant for which only two possible alleles are observed in the population), the frequency of the less common (minor) allele is called the minor allele frequency (MAF).

These SNPs were scattered across the genome and it was now time to describe what type of variants they were, where in the genome they were located, and what (if any) effect they had on human phenotypes.

The International HapMap Project

The International HapMap Project was the first effort of its kind to systematically catalogue genomic variation. Additionally, it aimed to characterise the LD structure of the human genome, which would make GWA studies feasible. The HapMap was officially started in October 2002 as a collaboration between research groups and private companies in Canada, China, Japan, Nigeria, the United Kingdom and the United States with the goal of developing a haplotype map (HapMap) of the human genome [90]. By genotyping individuals of African (YRI), European (CEU) and East Asian (JPT, CHB) descent, in Phase I HapMap assembled a publicly available database of common variants (MAF>5%) in global samples [91]. The HapMap expanded rapidly. By Phase II (2007), the database contained 2.1 million SNPs from the four original populations [92]. Phase III (2010) added genotyping from seven additional populations, for a total of over 3 million SNPs in 11 global ancestry groups[10] [93].

The growing popularity of genetic association studies in parallel with the expansion of the HapMap effort paved the way for the last needed technological breakthrough: microarrays. By knowing the genetic location of thousands of variants, commercial companies were able to develop so-called 'SNP chips' [94, 95], which allowed for genotyping at specific locations across the genome. In parallel, the data generated by the HapMap project enabled calculation of the LD[11] between SNPs within the genome, effectively describing the chance that two SNPs will be inherited together [72]. This enabled the identification of haplotypes and thus a minimal set of SNPs that capture the majority of the haplotype diversity within a population, called 'tag SNPs' [90]. The collective LD information gathered by academics in those years [96–98] allowed companies such as Affymetrix and Illumina to develop SNP arrays that contained these tag SNPs, effectively capturing information about common variation across a large percentage of the genome while only directly genotyping a few thousand SNPs.

[10] ASW (African ancestry in Southwest USA); CEU (Utah residents with Northern and Western European ancestry from the CEPH collection); CHB (Han Chinese in Beijing, China); CHD (Chinese in Metropolitan Denver, Colorado); GIH (Gujarati Indians in Houston, Texas); JPT (Japanese in Tokyo, Japan); LWK (Luhya in Webuye, Kenya); MEX (Mexican ancestry in Los Angeles, California); MKK (Maasai in Kinyawa, Kenya); TSI (Tuscans in Italy); YRI (Yoruba in Ibadan, Nigeria).

[11] Pearson's correlation coefficient squared, r^2, is commonly used

1.1.7 Genome-wide association studies

The data generated by the International HapMap Project combined with development of appropriate chip-based microarray technology, which enabled simultaneous genotyping of more than one million SNPs, led to the first wave of GWAS [78]. GWAS are a hypothesis-free[12] approach to test for statistical association between the genotype frequency of common genetic variants (considered one by one across the genome) and a phenotype of interest [76]. The development of GWAS was accompanied by great enthusiasm, and the hope that these studies could better our understanding of the genetic underpinning of human disease, leading to improvement of prognosis and acceleration of drug and diagnostics development.

Initially, GWAS focused on complex phenotypes with binary outcomes, using a case-control design (i.e. diseased vs healthy). Then, for each SNP and binary trait, the association was evaluated using a Cochran–Armitage (trend) test, a χ^2 test or a Fisher's exact test comparing the numbers of cases and controls when stratified by their alleles at the locus of interest. The first successful GWAS was published in 2002 on myocardial infarction [99]. The same design was then applied in a landmark GWA study conducted in 2005 for age-related macular degeneration (AMD), using 96 cases and 50 healthy controls and testing for associations at $\sim$100,000 SNPs [100]. In 2007, the Wellcome Trust Case-Control Consortium (WTCCC) published a study where they performed GWAS on seven different common diseases, using 2,000 cases for each of bipolar disorder (BD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type I and II diabetes (T1D, T2D) and a common set of 3,000 healthy controls, demonstrating the feasibility of the use of a shared set of controls across several traits [101].

A year later, in 2008, the GWAS Catalog was founded to keep a record of all published GWAS and identified associations [102]. As of September 2020, when I am writing this thesis, the GWAS Catalog includes 4,694 publications, describing 197,708 SNP-trait associations [103].

Over time, quantitative traits have become increasingly popular to use as phenotypes, in addition to binary traits. These include continuous traits such as height, weight and blood pressure. Furthermore, linear regressions and their derivations have become more popular methods to assess association, due to their flexibility to include covariates [76]. I describe these models in detail in the next chapter (**Chapter 2**).

[12] bar the selection of SNPs on the chip. More recently, shallow DNA-seq has been increasingly used as an alternative, making the approach truly hypothesis-free.

GWAS results are often visualised using a Manhattan plot [76], where the negative log p value (as a measure of significance) is plotted on the y axis, against the corresponding genomic position (ordered by chromosome and position) on the x axis (**Fig. 1.2**). Peaks on these plots represent loci (multiple variants in LD) that display evidence of association with the analysed phenotype. Variants are deemed to be significantly associated with a trait if they exceed an appropriately chosen p value threshold.

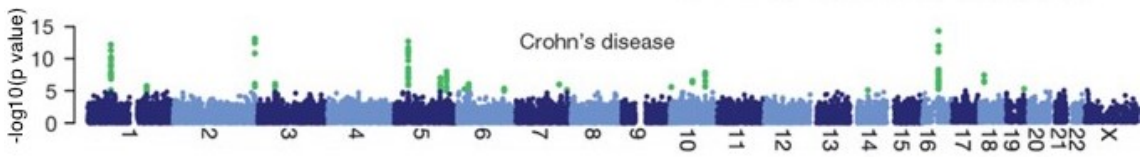


Fig. 1.2: Manhattan plot.
Manhattan plot for Crohn's disease (CD) from the WTCCC study [101]. On the x axis are plotted the genomic positions of the SNPs tested, one chromosome after the next. On the y axis are the association significance values. Alternating colours are used to distinguish chromosomes (odd numbered chromosomes -and chromosome X- are dark, even numbered chromosomes are light). Highlighted in green are statistically significant SNPs (p value $< 5 \times 10^{-8}$).

From global traits to molecular traits

Despite the success, the path from GWAS to biology is not straightforward, because an association between a genetic variant at a genomic locus and a trait is not directly informative of the causal mechanisms whereby the variant is associated with phenotypic differences [104]. GWAS have robustly associated thousands of genomic loci with complex traits. However, the interpretation of the identified variants is often far from trivial: the true causal variants are often oscured by LD, and the genes mediating a variant's effect on the trait are rarely ascertainable from GWAS results alone, especially for non-coding variants, which account for the majority of GWAS-identified risk alleles [65, 105, 106]. Maps of regulatory annotations and connections in disease-relevant tissues, generated by projects such as ENCODE (ENCyclopdia Of DNA Elements, [107]) and Epigenome RoadMap [108] can help the interpretation of these (putative) regulatory variants. Additionally, molecular quantitative trait loci (QTL) mapping, where genomic variants are associated with changes in expression (eQTL [109]), chromatin accessibility (caQTL [110]), DNA methylation (mQTL [111]), histone modifications (hQTL [112]), or protein level (pQTL [113]), have added an interpretative layer to the molecular mechanisms associated with disease-associated variants.

In this thesis, I focus on gene expression, i.e., the transcriptome, as a molecular phenotype. I use the next section (**section 1.1.8**) to describe key aspects of eQTL mapping.

1.1.8 Expression quantitative trait loci

Mechanisms of the genetic regulation of gene expression

During gene expression (or transcription), the genetic information stored in restricted portions of the DNA (genes) is used to produce RNA molecules. Next, during translation, the majority of RNA molecules are translated into proteins. As we have seen, only approximately 1% of the human genome is made up of protein-coding genes [54] whereas a large portion of the remaining sequence is thought to play a role in the regulation of gene expression [107].

The impact of genetic variation on traits and disease is the consequence of perturbations to this complex molecular machinery. An easily interpretable mechanism through which genetic variants may affect a phenotype is the direct alteration of the sequence and therefore functionality of the coded protein [114]. For example, sickle cell anaemia is caused by a SNP in the *HBB* gene, which causes one amino acid substitution in the sequence of the corresponding protein [86]. In alternative, genetic variation may affect the regulation of gene expression. One possible such mechanism would be the disruption of a specific sequence that affects the binding of proteins regulating the expression of a gene, for example a transcription factor (TF). In this scenario, the regulatory variant is said to be acting in *cis*. An alternative mechanism is the alteration of the DNA structure, which in turn can affect the functionality of regulatory elements, and ultimately gene expression [107, 108]. In this case the regulatory genetic variant is *trans*-acting.

Mapping eQTL

In the early 2000s, the new field of 'genetic analysis of global gene expression' or 'genetical genomics' [115] emerged, which applied traditional techniques of linkage and association analysis to thousands of transcript levels measured by microarrays [116]. This was partly driven by the decrease in cost of high-throughput profiling of gene expression, which made it possible to measure gene expression levels in large numbers of individuals. Genomic variants that were in this way associated with gene expression were termed eQTL. The first genome-wide maps of eQTL were performed in the early 2000s, using genetic linkage analysis, first in yeast [117], then in mammals, including humans [109]. A few years later, in 2007, the first modern eQTL studies using a GWAS-like approach were conducted in humans [118, 119]. It soon became clear that gene expression levels are strongly heritable: for all human genes the average heritability (portion of phenotypic variation due to genetic variation) was estimated to be around 0.25 [120–122, 114], making eQTL studies extremely popular.

RNA-sequencing[13], or RNA-seq, was a major breakthrough in the late 2000s and has been widely used since, substituting microarrays as the go-to technique to measure gene expression [123]. RNA-seq allows the measurement of the average expression abundance for each gene across a large population of input cells (**section 3.1.1**). The quantitative genetics field adapted quickly, and the first studies to perform eQTL mapping using RNA-seq data to measure expression level were published back-to-back in 2010, by groups in Chicago and Geneva [124, 125]. Today, (bulk) RNA-seq is the standard method to measure gene expression for eQTL mapping [126–128]. One of the advantages of RNA-seq compared to microarrays is that, in addition to the quantification of protein-coding genes, it also gives insights about other RNA traits, including splicing, exon level and increasingly also transcript level (see also **section 3.1.1**). As a consequence, splicing QTL (sQTL), can be mapped alongside standard (protein-coding) gene-level eQTL [125, 124], as well as exon eQTL (eeQTL), transcript ratio QTL (trQTL) and miRNA eQTL, among others [126, 129].

Cis and *trans* eQTL

An eQTL is a genomic locus that explains a fraction of the genetic variance of a gene expression phenotype (**Fig. 1.3**). As we have seen, regulatory effects on gene expression can usually be divided into those that act in *cis* (on the same molecule of DNA) and those that act in *trans* (on a different molecule of DNA, often through an intermediate). To identify true *cis* and *trans* effects, differences in gene expression abundance between a pair of individuals can be compared to the level of allele-specific expression (ASE) observed in their F_1 hybrid, similar to a classical *cis-trans* complementation test [130, 131]. In this test, a real *cis* effect would result in differential expression between the parents and corresponding ASE in the offspring. In contrast, a real *trans* effect would result in no ASE in the offspring.

On the other hand, standard eQTL analysis involves a direct association test between markers of genetic variation with gene expression levels, without requiring any previous knowledge about specific *cis* or *trans* regulatory regions. This association analysis can be performed proximally or distally with respect to the gene. Conventionally, variants within 1 Mb (megabase) on either side of a gene's transcription start site (TSS) are called *cis* eQTL, while those at least 5 Mb downstream or upstream of the TSS or on a different chromosome were considered *trans* acting [132, 114]. Although this terminology is technically not accurate, it has been widely embraced by the eQTL mapping community, and I adopt this convention in this thesis.

[13]Today we would call this technique 'bulk' RNA-sequencing as opposed to the more recent single cell RNA-sequencing; yet back then it was simply called RNA-sequencing, or 'RNA-seq'.

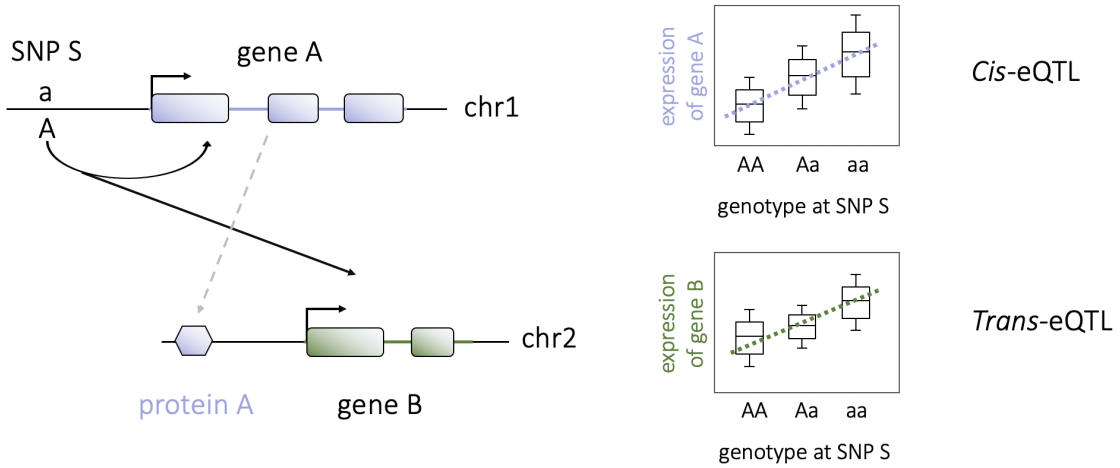


Fig. 1.3: *Cis* and *trans* eQTL.
Cis eQTL affect the expression of genes directly. *Trans* eQTL, in contrast, affect the expression of typically more distant genes, often by first having a *cis* effect on the expression of intermediate regulatory genes. Figure adapted from [114].

Typically, *cis*-eQTL have a large effect size [133], thus relatively modest sample sizes (from 80-100 samples) permit the detection of *cis*-eQTL for thousands of genes [118, 134]. *cis*-eQTL effects appear to be mostly additive [135], and *cis*-eQTL SNPs are often located close to the TSS of genes or within gene bodies [136]. An eQTL effect sizes generally increases as the distance between the eQTL SNP and the gene's TSS decreases. In contrast, the effect sizes of *trans*-eQTL are generally small [137, 138]. In addition, many more SNP variants need to be assessed, across all chromosomes, resulting in a much more severe multiple testing burden (see **section 2.2.2**). Combined, these two aspects mean that much larger sample sizes are required to detect *trans* eQTL. As a result, the number of reported *trans*-eQTL has remained small [138] in comparison to the number of reported *cis*-eQTL [114]. In this thesis, I worked on datasets with relatively small sample sizes, and so only performed proximal (*cis*) eQTL mapping.

From tissue-specific to cell type-specific eQTL

Early studies mainly performed eQTL mapping in whole blood or blood-derived cells, due to sample accessibility [118, 125]. However, more recently, several studies have published eQTL derived from a number of normal human tissues [134, 139–147], motivated by the observation that gene expression levels vary considerably between different tissues and cell types [148]. Perhaps most notable are those from the genotype-tissue expression (GTEx) project, which set out to collect and analyse gene expression profiles across several primary

tissues collected from post-mortem donors [149]. Since the publication of their pilot study in 2015 [127], several versions have been published [150, 151]. The most recent version (v8) was released in 2019 and it includes data from 53 tissue types[14] from 948 donors. These studies have collectively identified *cis* eQTL for the vast majority of human genes, emphasising their complex and cell type–specific nature, finding that 29%–80% of eQTL are cell type-specific [127, 141–143].

While eQTL from normal tissues provide valuable insights, tissues are constituted of multiple distinct cell types, each with specific gene regulatory profiles, as exemplified by the seminal work by Fairfax *et al.*, mapping eQTL in different blood-isolated cell types (monocytes and B cells [143]). Yet, apart from a few studies using purified cell types [143, 152, 153] or deconvolution methods [154, 155], eQTL datasets representing single primary cell types have been lacking [156]. Additionally, these methods are of limited use for less abundant cell types, and are dependent on accurately defined marker genes [157].

With the advent of single cell RNA-sequencing (scRNA-seq), eQTL mapping studies are now possible at the level of individual cells. scRNA-seq can be used to investigate rare cell types [158], and thus enables identification of cell type-specific eQTL in an unbiased manner. A first proof-of-concept study was published in 2013 on 15 individuals, where 92 genes were studied in 1,440 cells [159]. In 2018, the first genome-wide single cell eQTL mapping was performed in blood, on 45 individuals [160]. For further discussion on performing eQTL mapping using single cell RNA-seq as opposed to bulk RNA-seq, see **Chapter 3**.

Context-specific eQTL

In addition to cell type-specific eQTL, some effects of genetic variants on gene expression levels have been shown to manifest themselves only within cells that have been activated by a certain external stimulus (often called response eQTL [161–163]) or only in cells under certain conditions [164] - both are sometimes referred to as ‘context-specific eQTL’ [114].

[14] Adipose - Subcutaneous, Adipose - Visceral (Omentum), Adrenal Gland, Artery - Aorta, Artery - Coronary, Artery - Tibial, Brain - Amygdala, Brain - Anterior cingulate cortex (BA24), Brain - Caudate (basal ganglia), Brain - Cerebellar Hemisphere, Brain - Cerebellum, Brain - Cortex, Brain - Frontal Cortex (BA9), Brain - Hippocampus, Brain - Hypothalamus, Brain - Nucleus accumbens (basal ganglia), Brain - Putamen (basal ganglia), Brain - Spinal cord (cervical c-1), Brain - Substantia Nigra, Breast - Mammary Tissue, Cells - Cultured fibroblasts, Cells - EBV-transfortmed lymphocytes, Cervix - Ectocervix, Cervix - Endocervix, Colon - Sigmoid, Colon - Transverse, Esophagus - Gastroesophageal Junction, Esophagus - Mucosa, Esophagus - Muscularis, Fallopian Tube, Heart - Atrial Appendage, Heart - Left Ventricle, Kidney - Cortex, Kidney - Medulla, Liver, Lung, Minor Salivary Gland, Muscle - Skeletal, Nerve - Tibial, Ovary, Pancreas, Pituitary, Prostate, Skin - Not Sun Exposed (Suprapubic), Skin - Sun Exposed (Lower leg), Small Intestine - Terminal Ileum, Spleen, Stomach, Testis, Thyroid, Uterus, Vagina and Whole Blood.

Using eQTL to link genes to disease

Whilst useful in their own right to understand genetic regulation of gene expression across cell types and states, eQTL mapping can also be used for annotating variants associated with complex traits, as such variants are likely enriched for eQTL [165]. A recent study suggested that two thirds of candidate complex trait mediating genes identified as eQTL are not the nearest genes to the GWAS lead variants, highlighting the utility of this approach in annotating GWAS loci [166]. Importantly, GWAS variants are enriched in eQTL in a tissue-specific manner. For instance, whole blood eQTL are enriched with autoimmune disorder-associated SNPs but not with GWAS SNPs for bipolar disorder, or type 2 diabetes [127]. Thus, it is critical to use eQTL data from relevant tissues and cell types when following up GWAS loci for different diseases.

GWAS-eQTL colocalisation

Integrating eQTL maps with GWAS can identify potential molecular mechanisms underlying disease associations. One such integration method consists in testing for 'colocalisation' of a GWAS trait and a gene's expression trait - i.e. test for whether the same variant is causal to both traits [167]. Intuitively, if it can be established that a causal variant for a GWAS trait and one for a gene's expression are the same, this may suggest a regulatory role of the eQTL SNP in the pathway to the GWAS trait [168, 169].

However, simple overlap of GWAS and eQTL signals does not guarantee mechanism. First, the two variants may be two independent causal SNPs in LD with each other. Second, eQTL are abundant [126], with 48% of common genetic variants estimated to act as eQTL for at least one gene [170], making the overlap between GWAS and eQTL signals possible by chance. This motivated the development of formal statistical tests that estimate the probability of the overlaps between the two signals being due to chance - these are called colocalisation tests.

Early models [171–173] required full individual-level genotype data, which are seldom available [174]. More recently, Giambartolomei *et al*. proposed a colocalisation test (COLOC) which computes the odds of colocalisation compared to the null hypothesis using GWAS summary statistics [175]. Briefly, COLOC is a Bayesian statistical approach that tests for pairwise colocalisation of GWAS variants with eQTL, and generates posterior probabilities for each locus weighting the evidence for competing hypotheses of either no colocalisation or sharing of a distinct SNP at each locus [176]. Since its release, COLOC has become a reference method for colocalisation testing. Alternative models for colocalisation include

MOLOC (an expansion of COLOC to include multiple traits [177]). JLIM [178], HEIDI [166], Sherlock [168], eCAVIAR [179] and, most recently, 'jointsum' [180].

Transcriptome-wide association studies

Colocalisation analyses effectively use genome-wide significant SNPs to nominate causal genes for complex traits. However, the majority of variants contributing to complex traits and diseases have not yet been identified, arguably because their effect sizes are too small to be detected at current GWAS sample sizes [104]. Theoretically, one may want to use tissue-specific gene expression, rather than genotypes, for association testing. However, carrying out such studies is currently unfeasible, as it would require profiling gene expression across hundreds of thousands of individuals in both cases and controls, and across dozens of tissues [174]. Instead, transcriptome-wide association studies (TWAS) leverage eQTL information to predict (impute) the gene expression of the cases and controls from a GWAS, and then perform direct association of traits and genes - without directly profiling gene expression in every individual included in the GWAS [106]. Several methods have been proposed, such as TWAS [181], PrediXcan [182] and summary statistics-based Mendelian randomisation (SMR) [166].

In summary, both colocalisation and TWAS combine eQTL and GWAS catalogues with the aim to prioritise genes causally involved in complex diseases. In particular, colocalisation analysis integrates association signals from GWAS and eQTL on a locus-by-locus basis to identify instances in which both traits share a causal variant. In contrast, TWAS leverages information from eQTL data to impute gene expression values for all individuals in a GWAS, and then associates genes directly to traits. The availability of eQTL catalogues from a wider variety of cell types, as well as of larger sample sizes, will improve gene prioritisation and translate GWAS results to refined sets of disease-causal genes [174].

eQTL mapping in iPSC and iPSC-derived cells

In addition to human primary tissues, eQTL have been described in human induced pluripotent stem cells (iPSCs) and iPSC-derived cells (see **section 1.2.6**). In the second part of this introduction (**section 1.2**), I describe the use of human iPSCs as an *in vitro* model for early human development and as a resource to generate donor-matched cell types and tissues that are impossible to access *in vivo*. Population-scale iPSC derivation and differentiation provides an outstanding resource to investigate the role of genetic variants on expression in disease-relevant tissues and during development.

"It is not birth, marriage, or death, but gastrulation which is truly the most important time in your life."

Lewis Wolpert, 1986

1.2 Human iPSCs to study cell differentiation

Human iPSCs and cells derived therefrom are the biological system I use throughout this thesis to study the effect of common genetic variants on expression during cell differentiation. I use this section to provide some rudiments of human embryology and to describe the role of human stem cells in scientific research. I focus especially on the iPSC technology, which allows reprogramming of cells from easily accessible somatic tissues such as skin and blood to regain stem-cell like (pluripotent) properties and to be subsequently differentiated into virtually any desired cell type, when provided with the right stimuli.

In particular, in **section 1.2.1**, I provide a brief historical overview of the study of early development in humans and in **section 1.2.2** I describe the key stages of human embryology, which we attempt to mimic using iPSC-derived differentiation protocols. Then, in **sections 1.2.3** and **1.2.4**, I introduce the two key concepts of embryonic stem cells (ESCs) and somatic-cell nuclear transfer (SCNT), or somatic cloning. ESCs are cells collected at a very early stage of the embryo's development that can be grown *in vitro* to model development. Somatic cloning, on the other hand, involves introducing the nucleus of a donor's somatic cell into an enucleated ovum which can either be implanted into a host and let develop, creating a 'clone' for the donor; or, alternatively, grown *in vitro* for research purposes. Next, in **section 1.2.5**, I describe induced pluripotent stem cells (iPSCs), a technology that allows somatic cells to be reprogrammed to a pluripotent state, which can, in turn, be differentiated into virtually any differentiated cell types. This allows the generation of ESC-like cells directly from a(ny) donor, bypassing the need for cloning. In this section, I will describe the technology and highlight advantages and challenges in the use of iPSCs in biological research. Finally, in the last section (**section 1.2.6**), I describe applications of human iPSCs generated from many individuals to perform population-scale genetic analyses across a variety of iPSC-derived cell types.

1.2.1 From homunculi to developmental biology

In the previous section, we learnt that DNA encodes all instructions of life, and that each one of us contains a mixture of genetic information from both parents, which is identically copied in every one of our cells. But how do we go from one single cell containing our genetic fingerprint to a whole organism made of some 37 trillion highly specialised cells which differ in function and morphology and work together in harmony as a single organism?

Interest in the study of human development and the embryo has ancient origin. We define an embryo as the first stages of development of a fertilised ovum in the uterus. The word derives from the Greek εμβρυον (embryon, literally 'young one'). Early embryology (from embryon and -λογια, -logia, 'study of') was first proposed by an Italian, Marcello Malpighi, who was a promoter of preformationism. The theory of preformation believes that an embryo is contained in the semen (thus only derives from the father) and that it is essentially a preformed miniature infant ('homunculus') which just gets larger during development (**Fig. 1.4**).

The alternative explanation for embryonic development was epigenesis, originally proposed by Aristotle approximately 2,000 years earlier. Much early embryology and support for epigenesis came from the work of Italian anatomists such as Aldrovandi and Leonardo da Vinci (**Fig. 1.4**) in the Renaissance. According to the theory of epigenesis, the shape of an animal emerges gradually from a relatively formless egg. Yet for centuries, most people believed in preformationism over epigenesis. Only as imaging techniques improved during the 19th century, could biologists see that embryos take shape in a series of progressive steps, leading to epigenesis replacing preformation as the preferred explanation among embryologists.

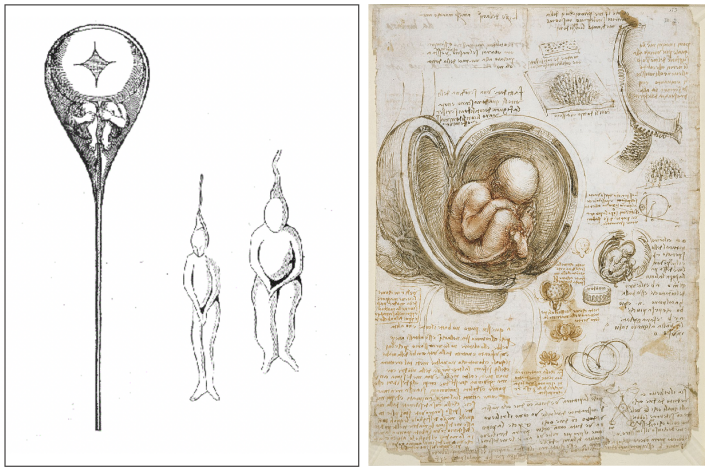


Fig. 1.4: Early theories of development.
Illustration of the two leading theories on human embryology before the 19^{th} century. Left, Illustration of Preformation: A tiny person (homunculus) growing inside a sperm, as drawn by Nicolaas Hartsoeker in 1695. Right, Epigenesis: study of foetus deveopment by Leonardo Da Vinci, c. 1510.

Modern embryology developed from the work of Estonian-born Karl Ernst von Baer. Von Baer is credited with discovering the mammalian ovum in 1827 and with becoming the first person to observe human ova. Only in 1876 did Oscar Hertwig prove that fertilisation is due to the fusion of an egg and a sperm cell [183]. Additionally, von Baer laid the foundation for the field of comparative embryogenesis in his 'On the developmental history of animals', published in 1828 [184]. In it, he formulated what became known as Baer's law of embryology. In particular, he observed that large changes in embryo development precede more specific changes and that at a time in development embryos from different species look very similar, despite looking very different as adults. Some of his ideas were later used by Darwin in his theory of evolution. In his study of embryology, Von Baer discovered the blastula and formulated the germ layer theory (see next section, **section 1.2.2**).

Human embryology as a scientific discipline is closely linked to the creation of human embryo collections [185–187]. The seminal work of Franklin Mall led to the creation of the Carnegie collection in 1887, which includes more than 10,000 human embryo specimens, and established the basic staging criteria for the developmental classification of human embryos [188]. Other collections were later created, such as the Kyoto collection, which today holds around 44,000 specimens [189]. Indeed, much of our current textbook knowledge of human development is derived from the early studies describing these samples.

The development of *in vitro* fertilisation (IVF) of human eggs [190–192], followed by the development of conditions to culture these fertilised eggs for up to 5-6 days [193, 194], allowed scientists to describe the dynamics of key morphological and genetic changes during early human development (see **section 1.2.2**).

In parallel, ever since the 1950s, with the helical structure of DNA being unraveled (see **section 1.1.3**), and with increasing knowledge in the field of molecular biology, developmental biology had emerged and was growing rapidly as a field of study that attempts to determine the interplay between morphological and gene expression changes during embryogenesis. Indeed, development requires both mechanical changes in cell and tissue shape, which drive morphogenesis, as well as gene expression changes, which regulate cell fate decisions and tissue patterning [195, 196]. Mechano-chemical feedback at the molecular, cellular and tissue level coordinates this crosstalk and instructs tissue self-organisation [197].

1.2.2 Human Embryogenesis

I use this section to briefly outline the key steps of human embryogenesis as this is the system we aim to mimic *in vitro* using differentiating human iPSCs. Embryogenesis describes the first eight weeks of development[15] after fertilisation, after which we generally refer to fetal development [201].

Thanks to technological advances in microscopy and *in vivo* experiments in model organisms such as fruit flies and mice, we have learnt a lot about this process since the pioneering experiments of von Baer. In particular, embryogenesis starts with a zygote, which is the single cell resulting from the fusion of gametes, an egg and a sperm cell; the fusion is known as fertilisation. Then, the first 12-to-24 hours post-fertilisation are spent in cleavage, which consists of very rapid cell (mitotic) division and no growth [202].

At around day 3, cells start to get clumped together in a process called compaction [203]. At this point (day 3/4), the 16-32 cell embryo is called the morula (which is the Greek word for mulberry), which will undergo blastulation [204] (**Fig. 1.5**). At around day 4, cells are still dividing, but they also begin to differentiate and develop specific forms and functions. Two layers develop: the cells on the outer layer are called trophoblasts, and the cells inside are embryoblasts [196, 195]. Next, at around day 5, the embryoblasts clump even further into an inner collection of cells called the inner cell mass, which is pushed to a side of the sphere formed by the trophoblast. The rest of the fluid-filled cavity is called the blastocoel and this conformation is called blastocyst. This is also the time when the zona pellucida (a protective membrane that surrounded the egg cell and that was limiting the embryo's growth) begins to disappear, allowing the blastocyst to grow, change shape and start moving [205].

The ability of the embryo to move allows the beginning of implantation: at around day 7, the embryo has now left the Fallopian tube and reached the uterus and attaches to its wall, the endometrium. The trophoblasts divide and then fuse with the endometrium - initiating the process that will eventually build the placenta.

During the second week, these cells form another cavity called the amniotic cavity. At the same time, they start differentiating further into two layers: the epiblast (closer to the amnitioc cavity) and the hypoblast (closer to the blastocoel) [202].

[15] covering the 23 Carnegie stages. The Carnegie system was based on work by Streeter [198] and O'Rahilly and Müller [199, 200] at the Carnegie Institution of Washington (based on the Carnegie collection, **page 25**). The system is standardised and provides a unified developmental chronology of the vertebrate embryo.

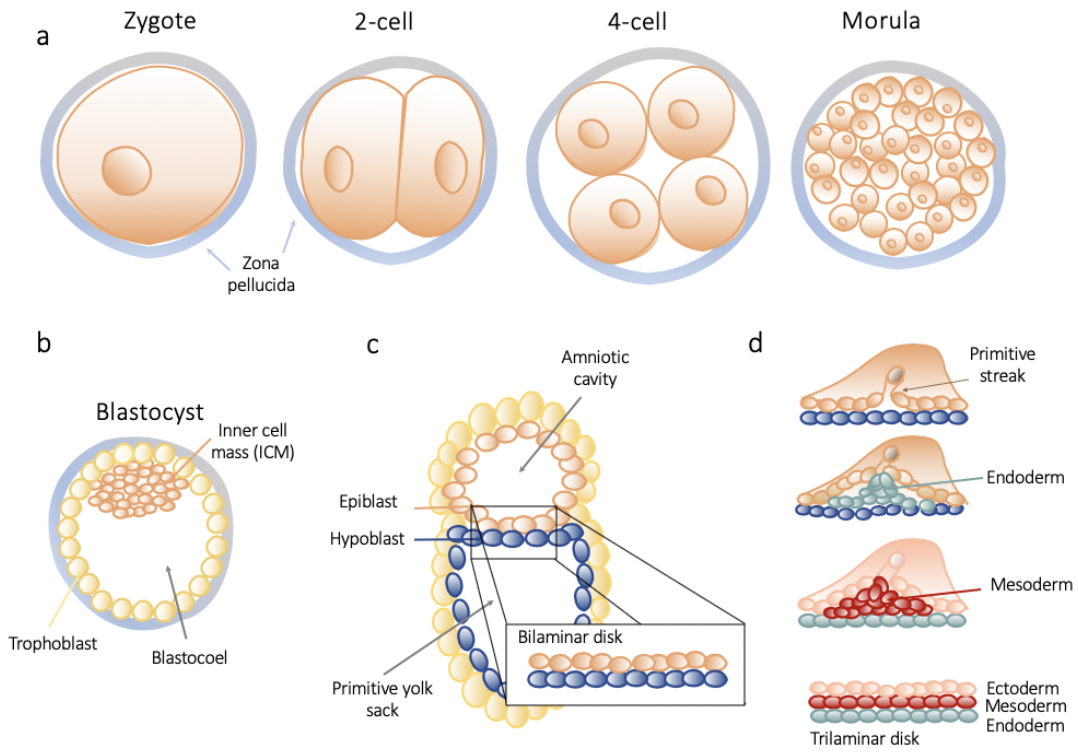


Fig. 1.5: Human Embryogenesis.
Early stages of human embryogenesis. (a) From zygote to morula: stages of clevage. (b) The blastocyst. Cells divide and form an outer layer, the trophoblasts. Inner cells also called embryoblasts get compacted and move to a side forming the inner cell mass (ICM), leaving a fluid filled cavity called the blastocoel. (c) The formation of the bilaminar disk. The ICM splits into the epiblast and hypoblast which distribute along the surface and result into two more cavities, the amniotic cavity and the primitive yolk sac, respectively. (d) Gastrulation: from bilaminar to trilaminar disk. The primitive streak forms and cells start to migrate moving down in between the epiblast and the hypoblast. The endoderm, mesoderm and ectoderm form and are called the trilaminar disk.

Together, the epiblast and the hypoblast form the bilaminar disc [206]. Only the epiblast contributes to the embryo, thus I do not discuss the hypoblast further (**Fig. 1.5**).

The next stage of early embryogenesis is gastrulation [207]. Famously called “the most important moment in your life” by Lewis Wolpert [208], gastrulation is the process during which the three germ layers (endoderm, mesoderm and ectoderm) form. At this stage, we call the cell mass a ‘gastrula’. The first step of gastrulation is the formation of the primitive streak (∼day 16). The primitive streak determines the midline of the body, separating the embryo’s left and right sides. At this point, cells are moving down from the epiblast, ending up between the original epiblast layer and the hypoblast. The first layer to invaginate descends the deepest and ends up closest to the hypoblast - this is the endoderm. The next layer forms

the mesoderm, and the remaining epiblast cells that continue to border the amniotic cavity are the ectoderm (**Fig. 1.5**).

The next stage is called neurulation (week 3-5). Directly underneath the primitive streak, mesoderm cells form a thin rod, known as the notochord. The notochord induces a change within the ectoderm above it, leading to the formation of the neural plate which then dives into the mesoderm to form the neural tube. This is the end of what is called 'early embryogenesis'.

Next, the germ layers start forming the various organs in a process called organogenesis. Briefly, the endoderm forms the gastro-intestial tract, from which upper tract the lungs, the liver and the pancreas form. The tube itself forms the esophagus, the stomach, and the small and large intestines. Second, the mesoderm forms some inner layers of the skin (endothelial cells), the muscles (including the heart), the bones, the kidneys, the bladder, ovaries and/or testis and blood cells. Finally, the ectoderm forms the outer layer of the skin (epithelial cells), sweat glands and hair, and importantly our nervous system. After eight weeks since fertilisation, we call the embryo a foetus, and fetal development begins.

1.2.3 Human Stem Cells

In mammals, roughly 50–150 cells make up the inner cell mass during the blastocyst stage of embryonic development, at around days 5–14 (**Fig. 1.5**). These have stem cell[16] capability, meaning that they can eventually differentiate into all of the body's cell types (making them 'pluripotent'). Only the zygote is considered to be truly 'totipotent' as it is able to form not only cells of the body (derived from the epiblast) but extra-embryonic tissues as well (from the hypoblast), which are necessary for the formation of a living organism. Cells with stem-cell properties are still present in the adult body, and are called somatic or adult stem cells. These are less potent than inner mass cells. In particular, some adult stem cells have the ability to differentiate into a whole suite of cell types, and are called 'multipotent'. For example, this is the case for hematopoietic stem cells, which give rise to all the cell types of the blood and the immune system. Other stem cells are more specialised, like epidermal stem cells, which are only able to differentiate into epidermal cells, or fibroblasts. These are called 'unipotent' (**Fig. 1.6**).

[16] stem cells are characterised by their self-renewing abilities and the 'potency' to differentiate into more specialised cells.

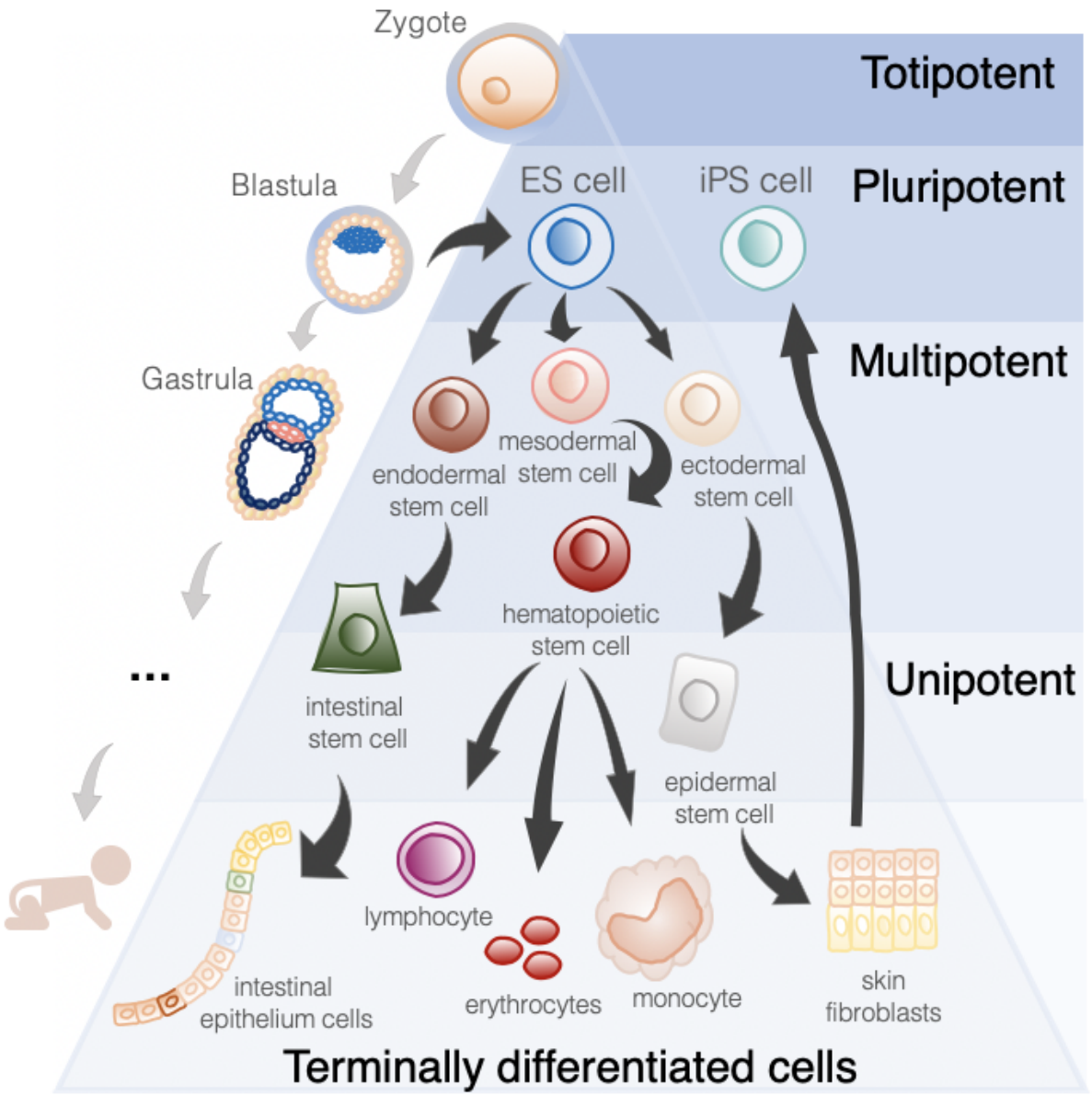


Fig. 1.6: The potency tree of stem cells.
The spectrum of potency along human development: from the totipotent zygote to terminally differentiated cells.

The existence of stem cells was first demonstrated by Canadian biologists Ernest McCulloch and James Till in the early 1960s. Together with graduate student Andy Becker and senior scientist Lou Siminovitch their work was published in *Nature* in 1963 [209].

Embryonic stem cells

In 1981, embryonic stem (ES) cells were first isolated and successfully cultured using mouse blastocysts by British biologists Martin Evans and Matthew Kaufman [210, 211]. In the 1990s, ES cell lines from two non-human primates, the rhesus monkey [212] and the common marmoset [213], were derived, and these offered closer models for the derivation of human ES cells. The first human embryonic stem cells (hESCs) were finally isolated in 1998 by the American developmental biologist James Thomson [214]. Thomson and colleagues derived hESC lines by culturing human blastocysts in a cocktail of growth factors and supporting mouse feeder cells. Indeed as they first demonstrated, human inner mass cells, when isolated and cultured *in vitro*, can be kept in the stem-cell stage and are known as hESCs.

Because they can proliferate without limit and can contribute to any cell type, human ES cells offer unprecedented access to tissues from the human body [215]. As a consequence, hESC lines held great promise for research purposes, as an *in vitro* model for both human development and disease [216]. First, embryonic stem cells have been used widely to improve our understanding of embryogenesis and development in general. Both in human and mouse, ESCs have been differentiated into a plethora of cell types. Additionally, ESCs can be grown not only in a 2D culture but also in 3D structures, to better mimic the formation of organs (i.e. organoids [217, 218]) or even entire (early) embryos (the so called ‘gastruloids’ [219]). Additionally, they can be used as disease models and for *in vivo* drug testing.

Second, ESCs could be used for cell replacement therapy, and, down the line, regenerative medicine [220]. Indeed, not only do ESCs have the ability to self-renew indefinitely, but in theory they can also be differentiated into any cell type in the body, thus providing functional replacement or support to worn-out or dysfunctional cells and tissues and, in a future, replacing organ transplants entirely.

However, in order to isolate ESCs the embryo needs to be destroyed which, naturally, caused a big ethical debate. Such ethical concerns caused the generation of ESCs to be reduced substantially or even made illegal in some countries. Today, a number of lines have been isolated and frozen and are available for research. Stem cell banks from a few countries provide (managed) access to several hundred hESC lines including the UK Stem Cell Bank

(139 lines) and the US NIH Human Stem Cell Registry (484 lines). In total, it is estimated that there are between 800 and 1,000 hESC lines available in the world which are, in the vast majority, derived from discarded IVF embryos [221].

In addition to the ethical controversies that the use of human embryos faces, which hinder the applications of human ES cells, it is difficult to generate patient- or disease-specific ES cells, which may be required for their effective clinical application [222].

1.2.4 Nuclear cloning of somatic cells

One early solution for the latter problem involved transplanting the nucleus of an adult donor cell in an enucleated oocyte in a process called nuclear cloning.

Nuclear cloning[17], also referred to as nuclear transfer or nuclear transplantation, denotes the introduction of a nucleus from an adult donor cell into an enucleated oocyte to generate a cloned embryo [223]. This embryo has the potential, when implanted into the uterus of a female host, to grow into an infant that is a clone of the adult that provided the donor cell, in a process termed 'reproductive cloning'. In alternative, the embryo can be explanted in culture, and give rise to embryonic stem cells that can differentiate into any adult cell type.

The first to perform nuclear cloning in animals was the British developmental biologist Sir John Gurdon. During his PhD in the Zoology department at Oxford, Gurdon worked on nuclear transplantation in a frog species of the genus *Xenopus*. In 1958, he successfully cloned a frog using intact nuclei from the somatic cells of a *Xenopus* tadpole [224]. This critical study proved that eggs receiving transplanted nuclei from a more mature cell type could be differentiated, directly contradicting what was believed at the time [225].

The first cloned mammal was Dolly the sheep, born on July 5, 1996. She was cloned by Keith Campbell, Ian Wilmut and colleagues at the Roslin Institute at the University of Edinburgh, using the process of somatic-cell nuclear transfer (SCNT) [226]. Dolly had three mothers: one provided an unfertilised oocyte (cytoplasmic donor), another provided the nucleus (from a mammary gland cell[18], nuclear donor) and finally a surrogate ewe hosted the embryo until its birth.

[17] The word 'clone' comes from the Greek word for twig, as it was first applied in plants.

[18] It is the use of a mammary gland as somatic cell source that motivated the name's choice: Dolly was named after country singer Dolly Parton, who is apparently famous (also) for her breasts.

Following the successful cloning of Dolly the sheep and the pioneering work of Tada *et al.*, who demonstrated somatic cloning by fusion of the somatic differentiated cells with ES cells [227], many other species have been cloned, including pigs, cats, dogs, horses, camels and even macaques. Combined, this work proved that somatic (differentiated) cells could be reprogrammed to a pluripotent state [228].

Until the early 2000s, the prospect of human cloned embryos explanted in culture was essentially the only envisoned way to generate patient- or disease-specific stem cells [222]. However, application of SCNT in human cells proved extremely challenging [229][19]. In parallel, funding restrictions and ethical concerns around hESCs provided the impetus to find alternative approaches for generating stem cells with the same degree of pluripotency.

1.2.5 Induced pluripotent stem cells

The need to find both more ethical solutions and more effective ways of generating donor-specific stem cells led to the generation of the first induced pluripotent stem cells (iPSCs). Discovered in 2006 by Japanese stem cell researcher Shinya Yamanaka, iPSCs are somatic cells (originally mouse fibroblasts) which are reprogrammed into acquiring a stem-cell identity [231]. Yamanaka has stated that it was the success of the cloning of Dolly the sheep that proved to him that reprogramming of somatic cells (in mammals) was possible. The next year, the labs of Yamanaka and Thomson[20] successfully generated iPSCs from human somatic cells [231–233], surpassing nuclear transplantation as the first step toward effective regenerative medicine and becoming the leading alternative to hESCs for developmental research.

In 2012 Sir John Gurdon and Shinya Yamanaka were jointly awarded the Nobel Prize in Physiology or Medicine for their combined efforts in discovering that “mature cells can be reprogrammed to become pluripotent” [234] (**Fig. 1.7**).

[19] Only more recently has somatic cell nuclear transfer been successfully performed to generate human ESCs (NT-ESC), providing an alternative method to convert human somatic cells to a pluripotent state [230].

[20] The Thomson group had also isolated human ESCs for the first time in 1998, see **page 30**.

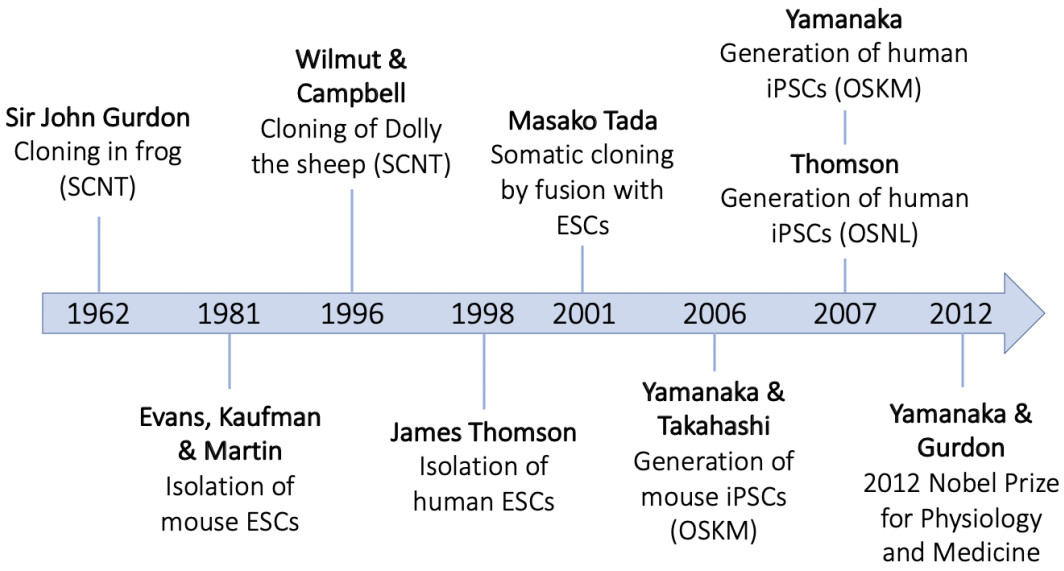


Fig. 1.7: Historical timeline of key events leading to the development of iPS cells.
Key events including somatic cloning and isolation of ES cells that eventually led Shinya Yamanaka and colleagues to generate iPSCs in 2006 and Yamanaka to win the 2012 Nobel Prize for Physiology and Medicine alongside Sir John Gurdon.

Inducing pluripotency

The generation of iPSCs involves the reprogramming of differentiated somatic cells from readily accessible tissues (such as skin or blood) into a pluripotent state by the introduction of a cocktail of factors that are able to 'reset' the transcriptional programme of the cells to an embryonic state [216].

The induction of iPSCs (in mice) was first described in a seminal paper published in *Cell*. In it, Takahashi and Yamanaka first demonstrated induction of pluripotent stem cells from mouse fibroblasts (both embryonic and adult) by induction of four transcription factors, Oct3/4[21], Sox2, c-Myc and Klf4, under ES cell culture conditions [231]. In the paper, the authors demonstrate that iPSCs exhibited similar morphology, proliferation properties and doubling times compared to ESCs. Additionally, iPSCs expressed major ES cell marker genes like *SSEA-1* and *Nanog* and formed teratomas upon injection into immunocompromised mice[22].

[21] also called Pou5f1.

[22] A teratoma is a non-malignant tumor comprised of a disorganised mixture of cells from all three germ layers. In a teratoma assay, putative pluripotent stem cells are implanted into an immune-compromised mouse where they may proliferate and differentiate to form a teratoma.

However, these 'first generation' iPSCs failed to generate adult chimerae[23] or contribute to the germline [231], suggesting that the iPSCs were only partially reprogrammed [236]. In 2007, Yamanaka and other laboratories modified the induction protocols to generate fully reprogrammed 'second generation' iPSCs that were competent for adult chimera and germline transmission [237, 238, 235].

A year later, the Yamanaka and the Thomson labs demonstrated, at around the same time, successful induction of pluripotent stem cells in human. In the Yamanaka paper [232] human iPSCs were derived from adult human dermal fibroblasts using a retroviral system and using the same 4 factors used in mice, which later became known as the Yamanaka factors: Oct3/4, Sox2, Klf4 and c-Myc (or OSKM). In their paper, published on the same day, the Thomson group also showed successful generation of human iPSCs, using a slightly different technique: they used a lentiviral system to express the factors, and changed two of the inducing factors used: Oct4 and Sox2 remained, but the other two were replaced by Nanog and Lin28: this set of factors is sometimes called OSNL [233].

In the next two years (2007-2009), the same technology was successfully applied by several groups and across a range of human cell types, including fibroblasts [239] and other somatic cell types, such as pancreatic β cells [240], neural stem cells [241, 242], stomach and liver cells [243], mature B lymphocytes [244], melanocytes [245], adipose stem cells [246] and keratinocytes [247], demonstrating the universality of cellular reprogramming [236].

Challenges in the use of iPSCs

Yet for iPSCs to fulfil their promise (that they are viable and possibly superior substitutes for ESCs in disease modeling, drug discovery, and regenerative medicine), several challenges on the road to their clinical application needed (and some still need) to be overcome. First, very low efficiency was recorded: in the original Yamanaka paper, only 0.01–0.1% [231] of the starting cells effectively exhibited pluripotency, and other initial reports did not exceed 1% [232, 235, 248].

[23] A chimera is a single organism composed of cells with more than one distinct genotype. The name derives from Greek mythology, where a chimera was a fire-breathing monster that was part lion, part goat, and part dragon. To generate a chimera, induced pluripotent stem (iPS) cells (e.g. with a gene-specific mutation) are injected into blastocysts, which are then transplanted into the uteri of pseudo-pregnant mice. By breeding a chimeric mouse with a wildtype mouse and observing the corresponding phenotype, e.g. coat colour, one can assess the contributions of the iPSCs to the germline [235].

Second, the over-expression of oncogenes (genes that have the potential to cause cancer) such as *c-Myc* and *Klf4* during the generation of iPSCs raised safety concerns. Indeed, in the original report of germline-competent iPSCs, ∼20% of the offspring developed tumors that could be traced back to the reactivation of the *c-Myc* transgene [235]. Furthermore, there is the risk of insertional mutagenesis due to virus-based delivery methods [231–233]. Finally, several studies have reported incomplete reprogramming, with cells maintaining some degree of 'epigenetic[24] memory' from their somatic cell of origin, which can lead to their biased differentiation potential into certain cell types depending on the donor cell source [249, 250].

Much progress has been made in the past decade to address these limitations and to improve the reprogramming technique, including the development of new methods to induce reprogramming. In the following sections I present an overview of the advances made to improve the reprogramming technique, and discuss methods for characterising iPSCs in general.

Reprogramming factors

Generating iPSCs requires the introduction of pluripotency related factors into the somatic cell. The generation of iPSCs by the Yamanaka and Thomson groups using different cocktails of transcription factors may suggest that different factors activate the same reprogramming pathway by reinforcing each other's synthesis. As a consequence, apart from the 'fantastic four' OSKM transcription factors, Oct4, Sox2, Klf4, c-Myc and the alternative OSNL combination described by the Thomson group containing Oct4, Sox2, Nanog and Lin28 [233], other factors, sometimes referred to as 'reprogramming enhancers', have been found to increase reprogramming efficiency and iPSC quality [251]. Those include other transcription factors, small molecules, microRNA's (miR) and different culturing conditions.

Mechanisms of iPSC induction

Several studies have described how the ectopic expression of OSKM in somatic cells induces the transition to a pluripotent state [222, 252–256]. Based on these studies, we can now describe the order of events during the reprogramming process, which can be divided broadly into two waves or phases: an initial, stochastic early phase (phase 1) and a more deterministic and hierarchical late phase (phase 2) [236, 251, 257].

[24] Epigenetics, from the Greek prefix epi- (meaning on top, around, in addition), refers to heritable changes of gene expression that do not involve variation in the genetic sequence of DNA. For example, DNA methylation and histone modifications are common epigenetic changes.

During induction using the OSKM factors, the first transcriptional wave (phase 1) is mostly mediated by c-Myc and occurs in all cells, whereas the second wave (phase 2) is mostly targeted to 'reprogrammable' cells, and involves a gradual increase in the expression of the Oct4 and Sox2 targets, leading to the activation of other pluripotency genes that aid in the activation of the pluripotency network. Klf4 seems to support both phases by repressing somatic genes during the first phase and facilitating the expression of pluripotency genes in the second phase [258].

The two phases describe the mechanisms of reprogramming in the cells that successfully become pluripotent iPSCs. However, as we have seen, these are a rather small percentage of all transduced cells. Three models, the elite, stochastic and deterministic models have been proposed to explain the reasons behind such low reprogramming efficiency [236]. For more detail into these three models, I refer the reader to Takahashi & Yamanaka [251].

Delivery methods

A number of different methods have been used to deliver the reprogramming factors into somatic cells. These delivery methods can be categorised into two groups: integrative systems (involving the integration of exogenous genetic material into the host genome) and non-integrative systems (involving no integration of genetic material into the host genome).

Early methods to deliver the reprogramming factors were integrative. These were mainly viral vectors, including retrovirus [231, 232, 238, 235, 222, 237], lentivirus [233, 259], and inducible and excisable retrovirus [260] or lentivirus [247] systems. Overall, integrative delivery methods have higher reprogramming efficiency as compared to non-integrative methods, but they are less safe, due to the risk of insertional mutagenesis.

Indeed, the introduction of efficient, integration-free methods for cell reprogramming was crucial to reduce safety risks. These alternative non-integrative induction methods have been developed in recent years, and involve the transient expression of reprogramming factors, including the use of viral vectors (adenovirus [240] and Sendai virus [261, 262]) and non-viral vectors, such as plasmids [263–266], transposons [267, 268, 263] synthetic mRNAs [269] and recombinant proteins [270].

Because they are safer, the use of non-integrative methods is overall more appealing for iPSC generation and use in clinical settings. Currently, episomal vectors, Sendai viruses and synthetic mRNAs are the most popular methods used for generating integration-free iPSCs.

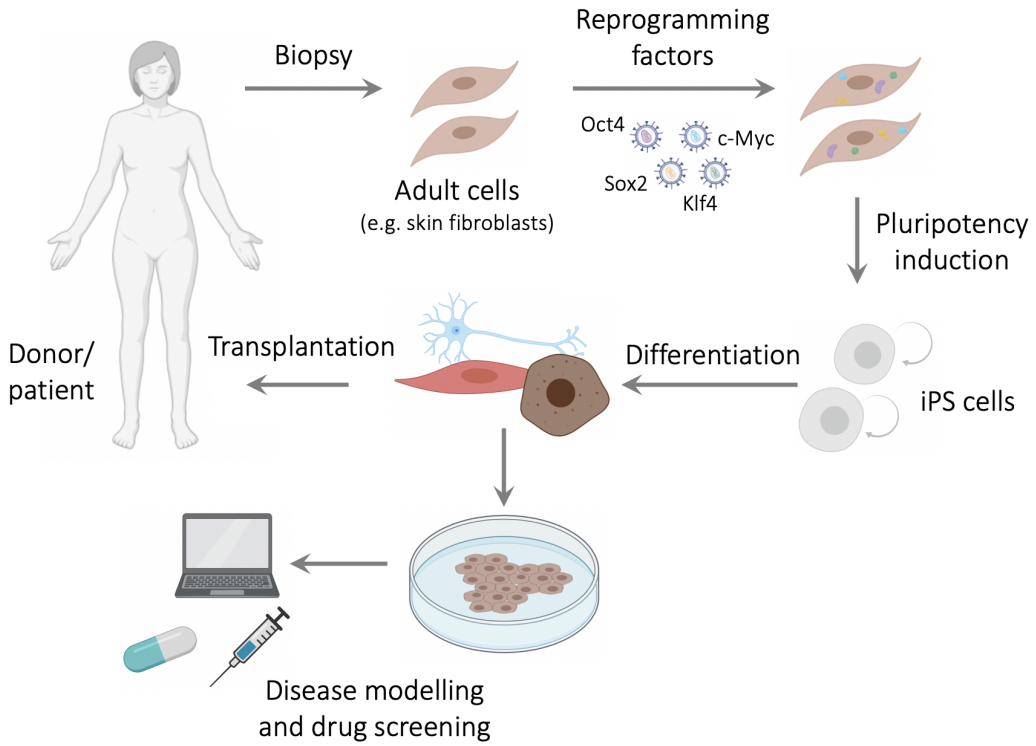


Fig. 1.8: iPS cells - derivation and applications.
The generation of iPSCs starts with a biopsy from, for example, the skin of an adult donor. Adult cells, in the example fibroblasts, are then isolated and pluripotency is induced through four reprogramming factors, for example the Yamanaka factors: Oct4, Sox2, Klf4 and c-Myc (OSKM). The factors can be delivered using a number of systems, see **page 36**. The resulting induced pluripotent stem cells (iPSCs) are self-renewing and are virtually indistinguishable from ESCs. They can be subsequently differentiated towards other cell types, including disease relevant cell types, for example dopaminergic neurons for Parkinson's disease, or pancreatic beta cells for diabetes. In the future, iPSC-derived cells might be used for cell therapy, i.e. they could be transplanted into the patient, with no risk of allogenic rejection. In the meantime, iPSCs and iPSC-derived differentiations can be used to model development and disease, and to do compound screening and drug testing. Figure created with BioRender.com.

Somatic cell of origin

As for the cell source for reprogramming, somatic cells should preferentially be easily accessible, susceptible for reprogramming and the reprogramming process should ideally be highly efficient [257]. Several different human somatic cell types have been successfully reprogrammed; however, reprogramming efficiencies and kinetics vary between somatic cell types. Keratinocytes for example showed a 100 times higher reprogramming efficiency ($\sim$0.8%) and were reprogrammed two times faster than skin fibroblasts under the same conditions [271].

Today, human iPSCs have been derived from a multitude of cell types [272], including dermal skin fibroblasts [232, 233], adipocytes [273], nucleated blood cells from peripheral blood [274, 275], dental pulp [276], keratinocytes from hair follicles [271] and renal tubular cells from urine samples [277]. However, approximately 80% of human iPSCs used in published studies are generated from fibroblasts [251].

As mentioned above, the choice of cell source may also have consequences in terms of the epigenetic makeup of the iPSCs generated. Indeed, during the reprogramming process the somatic cells' epigenetic signature must be erased in order to adopt a stem cell-like epigenome. These changes include chromatin reorganisation, DNA demethylation of promoter regions of pluripotency genes like *NANOG*, *SOX2* and *OCT4*, reactivation of the somatically silenced X chromosome, and genome-wide resetting of histone post-translational modifications [232, 237, 238, 258]. If the reset is incomplete, cells may maintain some degree of 'epigenetic memory' from their somatic cell of origin, which can lead to biased differential potential [249, 250]. However, it has been shown that their residual epigenetic memory diminishes as the cells are passaged in culture over a period of time [278].

Characterising iPSCs

As we have seen, reprogramming iPSC is a complex and not particularly efficient method. As a consequence, it is important to carefully characterise the iPSCs obtained after reprogramming [257] and establish criteria to evaluate the 'quality' of iPSCs. Different methods have been and should be used in combination to deeply characterise iPSCs. First, the characteristic ESC-like morphology of iPSCs is often used as an indication of their correct formation. iPSCs can be observed as small cells with large nucleus/cytoplasm ratios that form tightly packed colonies with clear, sharp edges.

Second, iPSCs are characterised by the expression of pluripotency markers including Oct4, Nanog, SSEA-3, SSEA-4, TRA-1-60 and TRA-1-81 [279]. Additionally, bioinformatics tools, such as the PluriTest [280], have been developed which use gene expression to assess the level of pluripotency.

In addition to morphological and gene expression considerations, iPSCs can also be evaluated functionally by their differentiation potential. Indeed, iPSCs should be able to terminally differentiate into cells of all three germ layers (endoderm, mesoderm, ectoderm) - which can be evaluated through *in vivo* teratoma formation assays or *in vitro* differentiation through embryoid body (EB) formation. Finally, since reprogramming influences the genetic and epigenetic make-up of the cells, iPSCs should be carefully characterised for their genetic and epigenetic profiles. Specifically, karyotyping is commonly used to evaluate genetic abnormalities in iPSCs, i.e. to verify that cells are diploid. Additionally, if transgenes are used for reprogramming, it is important to verify that the expression levels of the transgenes are properly down regulated once the iPSCs are formed. As for the evaluation of the epigenetic profile of the iPSCs, DNA methylation patterns can be assessed. Since DNA methylation contributes to silencing of genes, it is important that the generated iPSCs show DNA demethylation at key pluripotency genes (e.g., *Oct4, Nanog, Sox2*), while genes specific to the donor somatic cell type become methylated and silenced [257, 236].

Heterogeneity of iPSCs and between cell line variation

An important aspect of iPSC biology is the large variability observed between different iPSC lines and clones (even when derived from the same donor). This includes differences in terms of their differentiation capacity, epigenetic status, immunogenic and tumorigenic potential, maturation level, batch variability and co-occurrence of heterogenous populations of lineage subtypes and/or non-relevant cells as contaminating cell populations [258]. This observed diversity, which is greater than what has been described in ESCs, can be explained by the residual epigenetic memory, genetic background and other characteristics acquired during reprogramming and differentiation [249, 250, 281]. Naturally, understanding these sources of variable differentiation, particularly in terms of efficacy and safety, will be critical for the successful use of cell replacement therapies in the clinical setting [258].

Thus, it is important to investigate sources of variable differentiation potential across lines, including lines derived from genetically distinct individuals.

1.2.6 Applications of human iPSCs in genetics

In contrast to ES cells, the use of which, as we have seen, raises important ethical concerns[25], iPSCs can be derived from easily accessible tissues such as blood or skin, bypassing all such issues. Because they are fairly easy to derive, they can be generated for individuals from all genetic backgrounds, including individuals carrying a genetic disorder of interest. In the future, the iPSC technology opens the way to regenerative medicine where tissues can be re-generated with one's own cells thus avoiding the risk of immune rejection (when the body rejects a donor's allogenic organ as a foreign object).

In addition to the study of genetic disorders and the use for tissue regeneration, the iPSC technology can be applicable to basic biological research of human development and disease modeling. Indeed, human iPSCs have already been differentiated into a plethora of differentiated cell types, including neural stem cells [282], cortical, dopaminergic and motor neurons [283–285], astrocytes [286] and oligodendrocytes [287] as well as cardiomyocytes [288], skeletal muscle cells [289], vascular endothelial/smooth muscle cells [290], hepatocytes (liver cells) [291], pancreatic beta cells [292] and lung epithelial cells [293].

In recent years, as the iPSC technology became more established, iPSC lines have been derived from large cohorts, allowing deeper characterisation of these cells across multiple individuals. This paved the way to the study of how common genetic variants affect gene expression in iPSCs. These large cohorts represent a resource which allows differentiation of a number of these lines and the interrogation of eQTL in derived differentiated cell types.

HipSci and other human iPSC consortia

The Human Induced Pluripotent Stem Cell Initiative (HipSci) [294] is the largest human iPSC cohort to date, and the results described in this thesis are all obtained using HipSci lines. The original goal of HipSci was to generate a large, high quality reference panel of human iPSC lines for the research community to help deeply characterise these cells. All HipSci donors are volunteers from the Cambridge area in the UK. Donors are for the vast majority of European descent, both male and female and across a range of ages (for healthy donors, [50-54]-[65-69]). The majority of lines are derived from healthy donors, with small groups of diseased[26] samples. For sample collection, primary fibroblasts from skin biopsies (from the inner upper arm to minimise somatic mutations due to sun exposure)

[25] as it involves the destruction of the embryo

[26] Diseases included in HipSci are monogenic diabetes, Bardet-Biedl syndrome, Usher syndrome, Hypertrophic cardiomyopathy and a few others, see http://www.hipsci.org.

were collected from consented research volunteers recruited from the NIHR Cambridge BioResource and iPS cell derivation was performed either using the Sendai reprogramming kit or episomal plasmids expressing the (human) Yamanaka factors: human (h)OCT3/4, hSOX2, hKLF4 and hMYC [263]. Following transfer to feeder free culture and expansion, each line was submitted to quality control and the criteria for line selection were: (i) level of pluripotency, as determined by the PluriTest assay [280]; (ii) number of copy number abnormalities; and (iii) ability to differentiate into each of the three germ layers [294].

Additional iPSC consortia and large studies include iPSCORE (a panel of fibroblast-derived iPSC lines from 222 ethnically diverse and sometimes related individuals [295]), GENESiPS (a collection of 317 peripheral blood mononuclear cell (PBMC)-derived iPSC lines from 101 individuals [296]), PHLiPS (containing PBMC-derived iPSC lines from 91 individuals of European and African-American ethnicity [297]), and Banovich *et al* (a cohort of 58 lymphoblastoid cell line (LCL)-derived Yoruban lines, [298]).

eQTL mapping using iPSCs and iPSC-derived cells

In 2017, the HipSci consortium published a first paper, where they assessed iPSC lines through various assays, and mapped eQTL in 711 cell lines from 301 unique donors [294]. At around the same time, another map of iPSC eQTL was published in *Cell Stem Cell* [299]. Most recently, a meta-study was published as the result of an effort to combine iPSC resources into the i2QTL consortium [129]. Collectively, these studies have identified iPSC-specific eQTL, i.e. eQTL that function primarily in pluripotent cells. A subset of these tagged disease-associated loci, suggesting that they are capturing molecular changes early in development which are not well captured by studies of differentiated primary tissues from adult individuals. Alternatively, iPSC eQTL may be capturing stem-cell like molecular mechanisms, which are similar to mechanisms active in cancer [294].

In the last few years, eQTL have also been mapped in several iPSC-derived cell types. These include iPSC derived- macrophages [300], hepatocytes [297], neurons [301] and cardiomyocytes [302, 298]. These eQTL promise to greatly improve our understanding of genetic regulation in cell types that we do not typically have access to, including cells at early developmental stages and cells that are difficult to isolate, such as neuronal cell types. Combined with GWAS results (i.e. **page 21**), they can contribute to our understanding of the genetic architecture of complex diseases by facilitating analysis in the relevant cell types, for example dopaminergic neurons for Parkinson's disease (**Chapter 5**) or endoderm-derived (**Chapter 4**) pancreatic beta cells for diabetes.

1.3 Thesis outline

The overall aim of this thesis is to provide suitable computational methods for the identification of cell type and context-specific eQTL using single cell expression profiles, and explore their application across a range of human iPSC-derived cell types, using data from the HipSci project.

Specifically, in **Chapter 2**, I provide an overview of the use of linear and linear mixed models (LMMs) for genetic association analyses, focusing on their application in eQTL mapping.

In **Chapter 3**, I describe best-practice approaches to perform eQTL mapping using single-cell RNA sequencing (scRNA-seq) profiles and demonstrate these methods on matched bulk and single cell expression of around 100 human iPSC lines.

In **Chapter 4**, I present a dataset of almost 40,000 cells from 125 human iPSC lines differentiating to definitive endoderm, and demonstrate different approaches to eQTL mapping using scRNA-seq data, identifying genetic variants that affect gene expression dynamically along differentiation and across other cellular states.

In **Chapter 5**, I present a dataset of over one million cells from 215 human iPSC lines differentiating to midbrain dopaminergic neurons. We identify thousands of eQTL across a number of cell types and upon external stimulation. In addition, we identify hundreds of colocalisation events with variants that are known to be associated with neurological traits and diseases. Moreover, we investigate sources of variation in the capacity of individual cell lines to differentiate toward neurons.

Finally, in **Chapter 6**, I conclude and discuss future directions.

2

Linear mixed models for eQTL mapping

In **Chapters 3-5** I describe various models for eQTL mapping using single cell expression profiles. All of these models build on a linear mixed model (LMM) framework. I use this chapter to provide an overview of linear and linear mixed models and their application in quantitative genetics, with a focus on their use for eQTL mapping. I will also briefly introduce LMM-based models to test for genotype-environment (GxE) interactions, to provide the necessary theoretical foundations for some of the analyses in **Chapter 4**.

LMMs are a very popular framework for many genetic analyses. They are especially appealing because they provide robust control for confounding factors. One drawback of these methods is that inference using LMMs is typically computationally costly, yet efficient implementations of specific LMMs exist, enabling application to large cohorts. I use this chapter to provide an overview of the use of LMMs in genetic association studies: in **sections 2.1-2.2**, I discuss the linear model (also called linear regression) and basic applications for genome-wide association studies (GWAS) and quantitative trait locus (QTL) mapping. In **section 2.3**, I introduce the LMM and discuss applications in genetics, with a focus on the use of LMMs for eQTL mapping. Finally, in **section 2.4**, I briefly discuss extensions of the LMM framework to test for GxE interactions.

For mathematical model descriptions throughout this thesis, I use the following notation: bold, lower-case letters symbolise one-dimensional column vectors (e.g. $\mathbf{v}$) and bold capitalised letters matrices (e.g. $\mathbf{M}$). A normal distribution is specified by $\mathcal{N}(\mu, \sigma^2)$, where μ, σ are two scalars representing the mean and standard deviation parameters. For simplicity, I use the same notation for multivariate normal (e.g. $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$), noting that the specified parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) are an $N \times 1$ mean vector, and an $N \times N$ covariance matrix, respectively.

2.1 The linear regression model

A linear model, or regression, is a statistical approach to modelling a continuous output variable (dependent variable, or outcome) as a linear function of one or more input variables (features, or independent variables). For F features, the outcome variable for a single sample i can be specified as:

$$y_i = \sum_{f=1}^{F} x_{i,f}\beta_f + \psi_i, \tag{2.1}$$

where the noise term ψ_i ($\psi_i \sim \mathcal{N}(0, \sigma_n^2)$) accounts for measurement noise of y_i, reflecting the non-deterministic relationship between y_i and $x_{i,f}$, and is assumed to follow a normal distribution with 0 mean and constant variance σ_n^2. Furthermore, the noise term is assumed to be independent across samples, i.e. $cov(\psi_i, \psi_j) = 0$ for every $i \neq j$. For N samples, the model in eq. (2.1) can be expressed in matrix form as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi}, \tag{2.2}$$

where $\mathbf{y}$ is the $N \times 1$ outcome vector, $\mathbf{X}$ is the $N \times F$ feature matrix, and $\boldsymbol{\beta}$ is the $F \times 1$ corresponding weight vector. Finally, $\boldsymbol{\psi}$ is the $N \times 1$ noise vector such that: $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I_N})$, where $\mathbf{I_N}$ denotes the $N \times N$ identity matrix.

2.1.1 The maximum likelihood solution

Equation (2.2) is a realisation of the probability distribution of the data $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma_n^2)$ given the input variables $\mathbf{X}$ and the model parameters $\boldsymbol{\beta}$ and σ_n^2. This is known as the likelihood of the model and is considered as a function of the model parameters, denoted as $\mathcal{L}(\boldsymbol{\beta}, \sigma_n^2)$. We can then express the model in eq. (2.2) as:

$$\mathcal{L}(\boldsymbol{\beta}, \sigma_n^2) = p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma_n^2) = \mathcal{N}(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma_n^2\mathbf{I_N}), \tag{2.3}$$

or, equivalently, as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma_n^2 \mathbf{I_N}). \tag{2.4}$$

In parameter inference, the maximum likelihood estimator (MLE) of the model parameters is defined as the set of parameter values that maximise the likelihood. In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the log likelihood ($\ell = \log \mathcal{L}$), noting that both functions will be maximised by the same parameter values. The log likelihood of the model can be explicitly specified as:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma_n^2) &= -\frac{1}{2}\left\{ N \log(2\pi\sigma_n^2) + \log|\mathbf{I_N}| + \frac{1}{\sigma_n^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{I_N}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\ &= -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_n^2) - 0 - \frac{1}{2\sigma_n^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned} \tag{2.5}$$

Denoting with $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma_n^2}$ the MLEs of $\boldsymbol{\beta}$ and σ_n^2, we can write:

$$\hat{\boldsymbol{\beta}}, \hat{\sigma_n^2} = \text{argmax}_{\boldsymbol{\beta}, \sigma_n^2}\, \ell(\boldsymbol{\beta}, \sigma_n^2). \tag{2.6}$$

By setting the gradient of the log likelihood in eq. (2.5) with respect to both parameters to zero, and solving the joint system:

$$\begin{cases} \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma_n^2)}{\partial \boldsymbol{\beta}} = \mathbf{0} \\ \dfrac{\partial \ell(\boldsymbol{\beta}, \sigma_n^2)}{\partial \sigma_n^2} = 0 \end{cases}, \tag{2.7}$$

We find:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.8}$$

and

$$\hat{\sigma_n^2} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \frac{1}{N}(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}). \tag{2.9}$$

Note that the solution for $\hat{\boldsymbol{\beta}}$ in eq. (2.8) is equivalent to the ordinary least squares (OLS) solution [303].

2.1.2 The restricted maximum likelihood solution

In Gaussian models as in eq. (2.4) the MLE of the variance parameter $\hat{\sigma}_n^2$ suffers from downward bias[1] because the weights $\hat{\boldsymbol{\beta}}$ are estimated from the data, which involves a reduction of the effective number of degrees of freedom.

Patterson and Thompson [304] presented a $\boldsymbol{\beta}$-free estimation of σ_n^2 via the restricted (or residual) maximum likelihood (ReML). Given the model in eq. (2.2) the ReML can be obtained by projecting the output vector in a space that is orthogonal to $\mathbf{X}$ such that:

$$\mathbf{AX} = \mathbf{0}. \tag{2.10}$$

Using eq. (2.10) and rewriting eq. (2.2) in terms of the projection $\mathbf{w}$ we obtain:

$$\mathbf{w} = \mathbf{Ay} = \mathbf{A}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\psi}) = \mathbf{A}\boldsymbol{\psi}, \tag{2.11}$$

which provides an expression of $\mathbf{y}$ that is independent of $\boldsymbol{\beta}$.

By estimating $\ell(\sigma_n^2|\mathbf{Ay})$ for the model in eq. (2.4), we derive the log restricted maximum likelihood:

$$\ell(\sigma_n^2) = -\frac{N-F}{2}\log(2\pi\sigma_n^2) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{X}| - \frac{1}{2\sigma_n^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.12}$$

which is maximised by:

$$\hat{\sigma}_{n\,ReML}^2 = \frac{1}{N-F}(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T(\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}), \tag{2.13}$$

which is now an unbiased[2] estimator for σ_n^2.

Note that eq. (2.13) is identical to eq. (2.9) except for the fact that N is replaced by $(N-F)$, denoting the loss of F degrees of freedom.

[1] The bias of an estimator refers to the difference between this estimator's expectation (here, $E[\hat{\sigma}_n^2]$) and the true parameter value (σ_n^2). Downward bias indicates that $E[\hat{\sigma}_n^2] < \sigma_n^2$.

[2] i.e. $E[\hat{\sigma}_{n\,ReML}^2] = \sigma_n^2$.

2.2 Regression models for association studies

As we have seen (**section 1.1.7**), originally GWA studies used contingency table tests (such as the χ^2 test) to assess the significant effect of a variant on a dichotomous trait [76]. However, these tests are sub-optimal in the case of quantitative traits, such as height and weight, which would need to be arbitrarily discretised. Additionally, these tests are not equipped to account for confounding factors (these are described in more detail in **section 2.3**). As a consequence, regression-based models became the preferred approach, because they allow covariate adjustment, and can directly provide a measure of the effect size for each tested variant [72].

When applying regression models in genetic association studies, the phenotype of interest is modeled as the outcome variable ($\mathbf{y}$). In GWAS, we typically look at 'organismal phenotypes', including traits such as height and eye colour or disease status/risk for complex disorders [76]. In (molecular) QTL mapping, we consider 'molecular phenotypes', such as gene expression (i.e. eQTL), or protein level (pQTL).

The genotype at the SNP of interest is modelled as the independent variable. We assume all SNPs to be biallelic, that is that they can only assume two possible values - A (major allele) and a (minor allele)[3]. Three models can be considered for the minor allele a. First, a dominant model (i.e. AA = 0, Aa = 1, aa = 1; where one copy of the minor allele is sufficient to have a phenotypic effect); second, a recessive model, (AA = 0, Aa = 0, aa = 1; where two copies of the minor allele are necessary for observing a phenotypic effect); finally, an additive model (AA = 0, Aa = 1, aa = 2; where the phenotypic effect is proportional to the number of minor allele copies). Throughout this thesis, we will consider an additive genetic model, which is commonly used in both GWAS and eQTL mapping analyses [86].

The test, then, consists of assessing the effect of each individual SNP ($\mathbf{g}$) on the phenotype ($\mathbf{y}$), one at a time. For quantitative traits, including molecular traits such as gene expression and some complex traits such as height, blood pressure or BMI[4], a linear regression is typically used[5]:

$$\mathbf{y} = \mathbf{g}\beta + \boldsymbol{\psi}. \tag{2.14}$$

[3]The notation of major and minor alleles refers to the allele frequency in the population studied. Alternatively, it is possible to denote A as the reference allele and a as the alternative allele, based on the reference genome.

[4]BMI: body mass index, a measure of weight normalised by height.

[5]In contrast, case-control designs are better modeled by a logistic regression, i.e. $logit(E[\mathbf{y}]) = \mathbf{g}\beta + \boldsymbol{\psi}$ [305, 306].

2.2.1 Statistical hypothesis testing

To test for whether an association between a genetic variant and a trait is present, we can compare the hypothesis where the genetic variant has no effect on the trait (called null hypothesis, H_0) and the alternative hypothesis when the variant does have an effect (effect different from 0, H_1). Formally, the association hypothesis test for eq. (2.14) is:

$$H_0 : \beta = 0 \tag{2.15}$$

vs

$$H_1 : \beta \neq 0. \tag{2.16}$$

We are then comparing the following models:

$$H_0 : \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I_N}) \tag{2.17}$$

and

$$H_1 : \mathbf{y} \sim \mathcal{N}(\mathbf{g}\beta, \sigma_n^2 \mathbf{I_N}). \tag{2.18}$$

Statistical hypothesis testing comprises three fundamental steps: i) define a test statistic; ii) obtain a p value and iii) upon a threshold on the p value, reject or accept the null hypothesis. A test statistic is a random variable that measures the level of concordance between the observed sample and the null hypothesis H_0. Using this test statistic, we calculate a p value as the probability, under the assumption that the null hypothesis H_0 is correct, of sampling a test statistic at least as extreme as the one we observed. An extreme value of the test statistic generally indicates evidence against H_0. The p value is a function of the test statistic and, by definition, it is uniformly distributed under H_0. Finally, if this probability is low (under a defined threshold, e.g. p value < 0.05), H_0 is rejected.

Two types of errors can be made in statistical hypothesis testing. A false positive (type I error) occurs when we reject H_0 when H_0 is true; in contrast, a false negative (type II error) is generated when we reject H_1 when H_1 is true. Other key concepts in statistical hypothesis testing are summarised in **Box 2**).

Box 2: Key concepts of statistical testing

Confusion matrix for statistical classification:

		Test Result	
		reject H_0	accept H_0
Actual value	H_1	True Positive (TP)	False Negative (FN)
	H_0	False Positive (FP)	True Negative (TN)

Below, I list some key concepts in statistical testing and their relationship to the confusion matrix above:

- Type I error $= FP$
- Type II error $= FN$
- Sensitivity $=$ Recall $=$ True positive rate (TPR) $=$ Power $= \frac{TP}{TP+FN}$
- Specificity $=$ True negative rate (TNR) $= \frac{TN}{TN+FP}$
- False positive rate (FPR) $= 1-$Specificity $=$ Size $= \frac{FP}{TN+FP}$
- Precision $=$ Positive predictive value (PPV) $= \frac{TP}{TP+FP}$
- Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$
- F_1 score[a] $= 2\frac{PPV*TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$
- Family-wise error rate (FWER) $= P(FP \geq 1) = 1 - P(FP = 0)$
- False discovery rate (FDR) $= \frac{FP}{TP+FP} = 1-$Precision

[a] Harmonic mean of precision (PPV) and sensitivity (TPR).

Three approaches are commonly used for statistical testing in genetic association analyses: the Wald test, the likelihood ratio test (LRT), and the score test (**Fig. 2.1**). In the next paragraphs, I will describe briefly all three but note that only the LRT is used in this thesis.

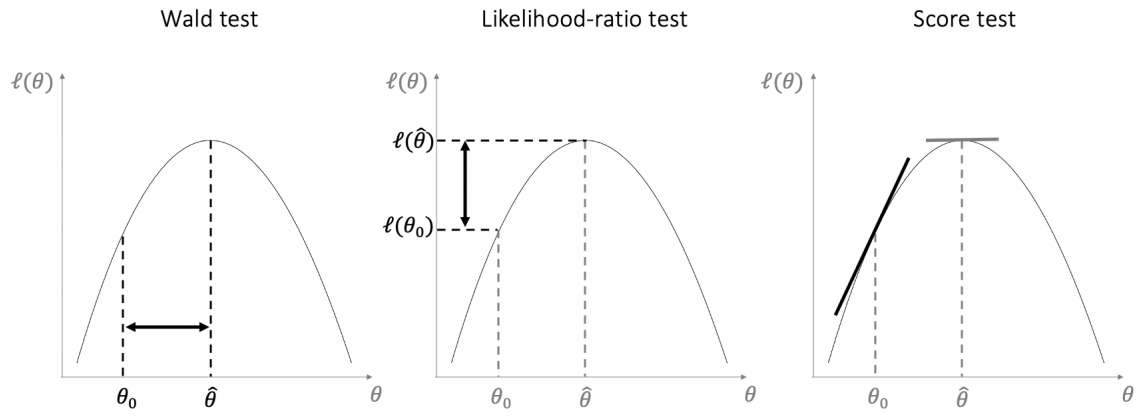


Fig. 2.1: The Wald test, the likelihood-ratio test and the score test.
The three most commonly used statistical testing approaches are illustrated here in the univariate case (single parameter θ). On the x axis is the parameter θ, and on the y axis the log likelihood $\ell(\theta)$. The Wald test essentially evaluates the difference between the MLE $\hat{\theta}$ and the parameter under H_0, θ_0. The likelihood-ratio test evaluates the difference between the log likelihoods evaluated at those values, i.e. $\ell(\hat{\theta})$ and $\ell(\theta_0)$. Finally, the score test evaluates the slope of $\ell(\theta)$ at θ_0. Note that the slope at $\hat{\theta}$ is $= 0$ by definition, as the MLE maximises $\ell(\theta)$.

Wald test

First, let us consider the Wald test and the generic null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ and the alternative $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}$ are the parameters in our model and $\boldsymbol{\theta}_0$ are their values under H_0. The Wald test statistic is defined as:

$$W = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0})^T [var(\hat{\boldsymbol{\theta}})]^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta_0}), \tag{2.19}$$

where $\hat{\boldsymbol{\theta}}$ are the MLEs for the $\boldsymbol{\theta}$ parameters (values of $\boldsymbol{\theta}$ that maximise the likelihood).

It can be shown that, under some assumptions [307], W follows a chi-squared (χ^2) distribution with number of degrees of freedom (dof) d equal to the number of the parameters tested ($W \sim \chi^2_d$).

In the univariate case (d = 1), eq. (2.19) can be expressed as:

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var(\hat{\theta})} \sim \chi^2_1. \tag{2.20}$$

Intuitively, our chance of rejecting the null hypothesis increases as the distance between $\hat{\theta}$ and θ_0 increases, and as our confidence in the estimation of the MLE increases (i.e. $var(\hat{\theta})$ decreases).

Likelihood ratio test

Second, let us consider the LRT. Here, the test statistic is the (log) likelihood ratio (LLR):

$$\text{LLR} = \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \ell(H_1) - \ell(H_0), \tag{2.21}$$

where we compare the value of the log-likelihood of the model under the null and alternative hypotheses, by evaluating eq. (2.5) (or eq. (2.12)) using MLE parameters estimated under H_0 ($\mathbf{0}$, σ_0^2) or H_1 ($\hat{\boldsymbol{\beta}}$, $\hat{\sigma_1^2}$).

The Wilks theorem [308], under some assumptions, guarantees that 2LLR, too, follows a χ^2 distribution with d dof (2LLR $\sim \chi_d^2$). The p value can then be calculated as:

$$P(\text{LLR}) = 1 - F_{\chi^2}(2\text{LLR}; d). \tag{2.22}$$

Score test

Finally the score test, also known as Lagrange multiplier test, is the last hypothesis test we consider. It was first developed by Rao in 1948 [309]. First, we define the score vector of Fisher as the gradient of the likelihood with respect to its parameters:

$$\mathbf{S} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}}. \tag{2.23}$$

The score test statistic is the Lagrange Multiplier (LM) and is defined as follows:

$$\text{LM} = \mathbf{S}(\boldsymbol{\theta_0})^T [var(\boldsymbol{\theta_0})]^{-1} \mathbf{S}(\boldsymbol{\theta_0}). \tag{2.24}$$

It can be shown that the LM, too, follows a χ^2 distribution with d dof (LM $\sim \chi_d^2$). To understand the intuition behind this test let us consider again the univariate case (d = 1):

$$\text{LM} = \frac{S(\theta_0)^2}{var(\theta_0)} \sim \chi_1^2. \tag{2.25}$$

At the MLE $\hat{\theta}$, the log likelihood is maximised and therefore its gradient $S(\hat{\theta})$ is equal to 0. On the contrary, in principle, $S(\theta_0) \neq 0$ (**Fig. 2.1**). Intuitively, the further $S(\theta_0)$ is away from 0, the more likely we are to reject the null hypothesis.

Intuition on differences between the three tests

In this thesis, I only apply the LRT. However, for different applications one might want to use alternative approaches. I use this paragraph to highlight the key differences between the three tests and provide an intuition as to when we should use one or the other. First of all, it can be shown that [310]:

$$W \geq \text{LLR} \geq \text{LM}, \tag{2.26}$$

which guarantees that the Wald tests statistic is always equal to or greater than the log-likelihood ratio, which in turn is always equal to or greater than the Lagrange multiplier. The Neyman-Pearson lemma [311] proves that, as a consequence, the power of the Wald test, defined as the probability of rejecting the null hypothesis when it is false: $P = P(\text{reject } H_0|H_1)$ (**Box 2**) is the same or higher than the power of the LRT, which is the same or higher than the power of the score test:

$$P_W \geq P_{\text{LLR}} \geq P_{\text{LM}}. \tag{2.27}$$

On the other hand, the false positive rate (FPR, **Box 2**) of the LRT, defined as the probability of rejecting the null hypothesis when it is true: $\text{FPR} = P(\text{reject } H_0|H_0)$ is also the same or higher than the FPR of the score test (but the same or lower than the Wald test). The LRT can be therefore considered as a good compromise between statistical power and accuracy. Moreover, one advantage of the LRT is that it is robust to re-parametrisation of the parameters, whereas the score and Wald tests are not. On the other hand, one main advantage of the score test is that it does not require the evaluation of the MLE of the parameters under the alternative hypothesis, but only under the null hypothesis. Additionally, while eq. (2.27) is generally true, close to the null the score test is considered "locally most powerful". Finally, the LLR follows a χ^2 distribution (asymptotically) only under the assumptions of the Wilks' theorem, which can be violated in certain applications. Namely, the value of parameters tested should be far away from the boundaries of the possible values the parameter can assume. For example, $-\infty < \beta < \infty$, so testing $H_1 : \beta \neq 0$ satisfies the assumption. On the other hand, $0 < \sigma^2 < \infty$ so testing $H_1 : \sigma^2 \neq 0$ violates the assumption, because the value 0 is at the boundary. Similarly, the score test statistic does not follow a χ^2_1 distribution in the presence of variance parameters, however an alternative test statistic can be defined which follows a linear combination of χ^2_1 distributions, and efficient methods to evaluate its significance have been proposed by Davies [312], Kuonen [313] and Liu [314, 315].

In the analyses in the following chapters (**Chapters 3-5**), I will use use tests that only evaluate the value of β, thus the LRT is well suited for these applications.

2.2.2 Correcting the multiple testing burden

In a typical GWAS one might test hundreds of thousands or millions of genetic variants. In eQTL mapping, tens of thousands of genes are tested, each essentially equivalent to a GWAS trait. Even when we only test for local eQTL (in *cis*, only SNPs within a window around the gene position, see **section 1.1.8**), we will still test hundreds of variants per gene, taking the average number of tests performed well into the millions. When performing so many tests, considering single test p values results in a high number of false positives (for example, for p value < 0.05 and 10^6 tests we expect 50,000 false positives under the null hypothesis). This is known as the 'multiple hypothesis testing problem'. I use the next section to provide a brief overview of commonly used approaches to correct for multiple hypothesis testing in the context of genetic analysis, with a focus on methods used in eQTL mapping.

Family-wise error rate correction

One strategy to perform multiple testing correction is to control the probability of having at least one false positive for a trait, which corresponds to a trait-wise p value known as the family-wise error rate (FWER, **Box 2**). The widely-used Bonferroni method follows this strategy assuming independence between tests [86]. Given a desired family-wise significance level α, the method consists of calculating 'adjusted p values' for each of the un-corrected p values (P) as $P_{adj} = P * n$, where n is the number of tests carried out. Next, setting $P_{adj} < \alpha$ ensures FWER < α. The Bonferroni correction strategy is conservative, because of the assumption of independence between tests, which ignores correlations between genotypes due to LD (**page 12**).

An alternative strategy, which accounts for the dependency of the statistical tests, is to use a permutation-based approach. The idea here is to build a background model by drawing from the empirical distribution maintaining the dependency structure of the underlying data but permuting the genotype data across individuals. This way, we disrupt a possible association between phenotype and genotype whilst maintaining the overall data dependencies. For each association test (resulting in an observed p value p_i), we can perform the same test M times, each time considering a different permutation of the genotype data across individuals. The resulting p values q_i, m represent the null distribution, i.e. the p values we might expect in the absence of any associations. One first approach is to use the p values from these M additional experiments to calculate a per-SNP adjusted p value, as the fraction of the M permutation-p values that are lower than the observed p value. For test i, the experimental adjusted p value after M permutations is calculated as:

$$P_{adj,i}^{perm} = \frac{1 + \sum_{m=1}^{M} q_{i,m} \geq p_i}{1 + M}, \tag{2.28}$$

where $q_{i,m}$ is the p value obtained at the mth permutation run equivalent to test i, and ones are added to avoid zero divisions. This strategy accounts for local LD, thereby increasing the statistical power, and has been widely used in *cis* molecular QTL mapping to estimate gene-level p values [127, 316]. However, as evident from eq. (2.28), the lower bound of $P_{adj,i}^{perm}$ depends on the number of selected permutations[6] for which the test statistic is calculated, and thus a higher number of permutations may be required to estimate a p value with high enough accuracy [317]. This can be improved by increasing the number of permutations, e.g. using M as large as 100,000, but that entails a great computational burden and can become unpractical in molecular analyses of large cohorts.

A second approach is to adjust for multiple testing when considering a more complex hypothesis, for example by pooling across variants, thus requiring fewer permutations. As an example, in the method described in [318], the minimal p values for each permutation across all SNPs are selected, and used to build a beta distribution of background p values, against which to compare the real p values. This method has been shown to robustly work for as little as (M=) 50-100 permutations, making use of the benefits of the assumption-free permutation approach without too much of the computational burden [318]. This is the method I will use in the following chapters (**Chapters 3**, **4** and **5**) of this thesis to control for FWER at gene level when performing large scale eQTL mapping.

False discovery rate correction

An alternative solution for multiple testing correction is to control the false discovery rate (FDR), i.e. the expected percentage of false discoveries (**Box 2**). The most widely used FDR-based correction method is the Benjamini-Hochberg (BH) procedure [319], which again assumes independence between tests[7]. Let us consider T tests with p values $p_1, p_2, ..., p_T$ and let $r_1, r_2, ..., r_T$ be their ranks (the smallest p value has rank 1, the largest has rank T), defining adjusted p values as $P_{adj,i} = \frac{T * p_i}{r_i}$ and setting $P_{adj,i} < \alpha$ ensures FDR < α [321]. Alternatively, the Storey procedure (proposed in 2002 [322, 323]) optimises the BH procedure by taking into account the distribution of the p values in the experiment. FDR-corrected p values are sometimes called 'q values'. For eQTL mapping, the assumption of independent tests holds

[6] For example, for M=1,000 the smallest adjusted p value we can obtain is only $P_{adj,i}^{perm}$=0.001 (when no permuted p value is smaller than the p value observed).

[7] Although, more recently, the BH procedure has actually been shown to hold under positive dependency [320].

when we consider a single SNP per gene. In this case, the number of tests T coincides with the numbers of genes tested. In practice, in most applications one is interested in the lead SNP for each gene, i.e. the SNP corresponding to the minimum p value. Conditional analyses (where we include the top SNP as a covariate in the model) can be used subsequently to detect secondary and tertiary effects for a gene.

Multiple testing correction for *cis* eQTL mapping

A typical strategy to correct for multiple hypothesis testing in *cis*-eQTL mapping is to use a two-step procedure [127]. First, for each gene an experiment-wise p value is obtained by correcting for multiple testing across variants using a FWER-based method. These gene-level p values are probability values for the hypothesis of a gene having at least one eQTL in the analysed region (i.e. of being an eGene). Second, the gene-level p values are corrected to control the FDR, for example using the Benjamini-Hochberg procedure.

In the analyses in **Chapters 3-5** of this thesis, I adopt this two-step approach. I use M=1,000 permutations and the method described in [318] to correct p values at the gene level (I will call the p values obtained this way 'empirical feature p values'). Then, I select the top SNP per gene and correct the corresponding empirical feature p values a second time, using the Storey procedure [322]. I will call the resulting p values: 'globally corrected p values'.

2.2.3 Calibration studies and distributions of p values

Under the assumption of no association between genetic variants and the analysed trait, an association model is expected to produce p values that approximately follow a uniform distribution. If that is the case, the model is said to be 'calibrated'. To verify that a model is calibrated we can disrupt the association between genotypes and phenotypes, by randomly permuting the genotypes. A representation that is often used to compare the observed (permuted) and the expected distributions of p values is the quantile-quantile (QQ) plot. In a QQ plot the observed negative $\log_{10}$(p value) is plotted against the expected negative $\log_{10}$(p value), where the expected value is obtained by drawing from a uniform distribution. I define as 'confounding', variables that are correlated with the genotype **g** (and sometimes also the phenotype **y**) and thus may create spurious associations when left unaccounted for (see also **section 2.3**). Inflated QQ plots are typically associated with the presence of such confounding factors, and can be used as a diagnostic tool (**Fig. 2.2**).

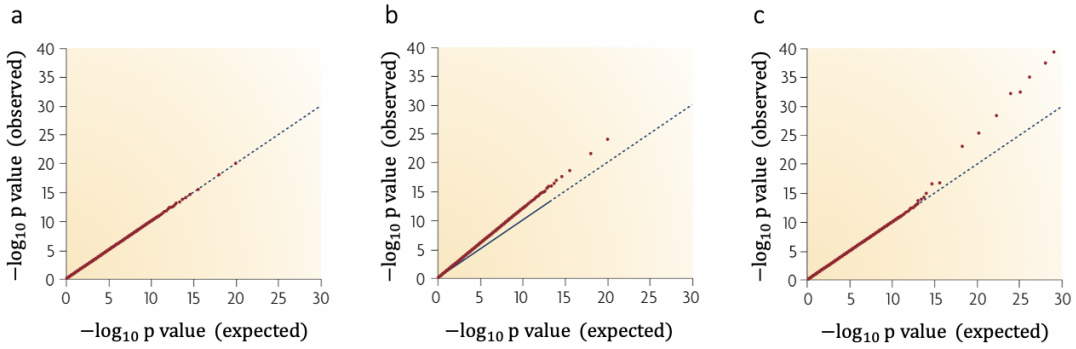


Fig. 2.2: Example QQ plots.
Examples of quantile-quantile (QQ) plots, displaying the expected negative log p values (x axis) versus the observed negative log p values (y axis) under the null hypothesis (diagonal blue line) and observed (red circles) in (a) when no associations are present, such that there is no departure from the null distribution, (b) in the presence of confounding factors, such that there is constant departure from the null distribution and (c) in the presence of a genuine genetic association, such that there is departure from the null in the tail of the distribution. Adapted from [76].

2.2.4 Including covariates in a linear model

The model in eq. (2.14) models the SNP tested as the only factor affecting the measured phenotype. However, if available, additional relevant information for the samples tested (such as the sex or the age of the individuals) can be added to the model as covariates ($\mathbf{W}$), and often improve discovery power by controlling for additional phenotypic variation. The model becomes:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{g}\beta + \boldsymbol{\psi}, \tag{2.29}$$

where $\mathbf{W}$ is an $N \times P$ matrix whose columns are known covariates, and $\boldsymbol{\alpha}$ is a $P \times 1$ vector of the corresponding weights. Here, covariates are implemented as fixed effects (FEs), and they only contribute to the mean value of $\mathbf{y}$, such that $E[\mathbf{y}] = \mathbf{W}\boldsymbol{\alpha} + \mathbf{g}\beta$, while $Var(\mathbf{y}) = Var(\boldsymbol{\psi}) = \sigma_n^2\mathbf{I_N}$, as before. In some cases, we can add FE covariates to adjust for technical batches or other factors that might affect $\mathbf{y}$, increasing the accuracy of our model. In eQTL mapping, we can often take advantage of the full transcriptome to identify factors affecting gene expression in a global manner, thus efficiently capturing known and potentially hidden covariates affecting expression across all genes. For example, we can perform principal component analysis (PCA) on the full expression matrix (genes x samples) and include such expression PCs as covariates in the model. Alternative more sophisticated methods to compute factors capturing global trends include, among others, SVA [324], PEER [325, 326], f-scLVM [327], pCMF [328], scVI [329, 330] and MOFA [331].

Different approaches can be used to determine the optimal number of PCs (or other factors) to include in the model. A popular approach is to choose the number of factors that maximises the number of eQTL discoveries [114, 150, 151]. The assumption is that global effects captured by PCs are orthogonal to the effects of a single variant on the expression of one gene, and thus there is no risk of generating false positives. There remains however a risk of over-correction, which can introduce synthetic associations as a result of collider bias [332]. Alternatively, to verify that specific covariates of interest (not modelled explicitly) are accounted for in the PCs, one may verify that PCs explain enough variance for those covariates (see for example the analysis I describe in **Chapter 5, section 5.5.1**).

2.3 Population structure and linear mixed models

One major source of confounding effects in genetic analysis - that I have purposefully not discussed thus far - is latent population substructure, which includes population stratification (i.e. the presence in the study sample of individuals with different ancestral and demographic histories) and relatedness between individuals, both known relatedness (e.g. known familial relations within a sample) and cryptic relatedness (i.e. evidence that individuals in the study sample have residual, non-trivial degrees of relatedness) [76]. It was acknowledged, even before the first GWAS was conducted, that there was a possibility of identifying false positives (or that true positives may be masked) when using population based association studies (instead of family based linkage studies), due to those confounding effects [333]. This is because both phenotypic prevalence (proportion of individuals exhibiting the phenotype) and allele frequencies (frequencies of a specific allele within a population) vary across different populations, which may result in the identification of spurious association between variants and the phenotype of interest due to ethnicity or population sub-structure (including relatedness) [333] (**Fig. 2.3**).

2.3.1 Early approaches to account for population structure

Various approaches have been proposed to account for population structure. An early solution was genomic control [334]. Genomic control correction adjusts for inflation due to confounding effects by dividing the test statistic of each marker by the genomic control parameter (λ_{GC}[8]).

[8] λ_{GC} compares the test-statistic value with the median value, such that $\lambda_{GC} \approx 1$ means there is no evidence for confounding, $\lambda_{GC} \gg 1$ indicates that the presence of confounding is likely.

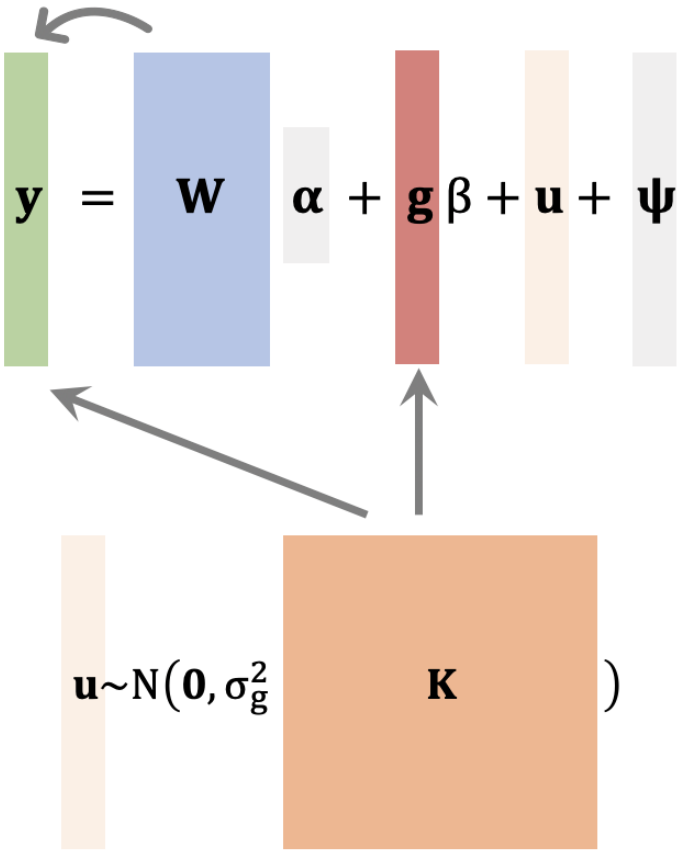


Fig. 2.3: Confounders and covariates.
Traditionally 'confounders' are variables that are correlated with both the genotype vector and the phenotype, i.e. hidden common causes of both **g** and **y**. For example, population structure (**K**) is a confounder in population genetics studies. On the other hand, other covariates such as sex (unless we are testing genes on the X or Y chromosomes) and batch (**W**), only have an effect on **y**. By including them as covariates in the model, we control for additional phenotypic variation and thereby increase association power. We note that in the illustration, a linear mixed model is used to illustrate how to account for both effects, with a random effect term used to correct for population structure, and covariates included as fixed effects. However, mathematically, both can be treated in the same way (i.e. they both can be included as fixed or random effects). In contrast the motivation for including them, and the effect their inclusion has, are different.

However, as different markers have different abilities to distinguish between populations, this uniform adjustment is far from optimal. Indeed, as a result, markers that strongly segregate across different population groups are only partially corrected, whereas markers that do not segregate tend to be over-corrected [335, 336]. Alternatively, a method that attempts to correct the underlying problem is STRUCTURE, which assigns individuals to discrete population subgroups and then combines the evidence for association across the different subgroups [337]. A limitation of this method is that only discrete subgroups can be considered. Additionally, it does not scale with sample size and is also highly sensitive to the number of defined population clusters [336].

The analysis of genotype data from large population studies, showed that genome-wide genetic variation could be used to accurately infer population structure [338–340]. In particular, it could be shown that the first principal components (PCs) calculated from the genotypic data were correlated with geographic axes [341]. As a result, one approach to account for population structure is to add the first genotypic PCs as covariates in the model (or to regress them out [336]):

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \sum_{i=1}^{p} PC_i b_i + \mathbf{g}\beta + \boldsymbol{\psi}. \tag{2.30}$$

The number of leading PCs p can be determined as the number of principal component (PC)s that cumulatively describe a certain proportion of the total genotypic variance. This approach (eq. (2.30)) was shown to perform relatively well in removing global population structure, but often failed to detect the more subtle relatedness effects. Therefore, when PCs are used to correct for population structure, closely related individuals have to be removed from the association analyses, prior to PCA calculations. Even then, cryptic relatedness might still be present and would not be properly accounted for by this method.

2.3.2 Linear mixed models for genetic analyses

Alternatively, linear mixed models (LMMs) can be used to successfully account for confounding effects linked to both population stratification and cryptic relatedness [342–347]. In an LMM, istead of being used to calculate principal components, the genotype data is used directly to estimate an $N \times N$ kinship matrix, $\mathbf{K}$, that describes the genetic similarity between pairs of individuals (see section below for a description of commonly used approaches to generate $\mathbf{K}$).

This genetic similarity is modelled through the use of an additional random effect term in the linear model described by eq. (2.29), as follows:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{g}\beta + \mathbf{u} + \boldsymbol{\psi}, \tag{2.31}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2\mathbf{K})$. The use of the notation $\mathbf{K}$ signifies the fact that this matrix reflects a degree of 'kinship' between individuals. It can be shown that the LMM approach is theoretically equivalent to the PC approach when all PCs are regressed or included as covariates (see Hoffman *et al.* [348] for details), explaining why LMMs are able to account for more subtle population structure than the PC approach. However, regressing all PCs or including all PCs as covariates is not feasible in practice (as the number of variables would exceed that of observations).

Kinship matrices

The covariance matrix $\mathbf{K}$ captures the latent population sub-structure of the samples considered, including population stratification and genetic relatedness. Several approaches have been proposed to compute this matrix, which is sometimes called the genetic relatedness matrix (GRM). Fisher's infinitesimal model (see **section 1.1.5**) [20] demonstrated that under an additive model with an infinite number of infinitesimal genetic effects, the phenotype follows a normal distribution, and the correlation of phenotypes between individuals is proportional to the fraction of genetic material that is 'identical-by-descent' (IBD)[9]. A consequence of this observation is that we can define the genetic relatedness between two individuals by using the predicted proportion of the genome that is IBD between them. Traditionally, an IBD (relatedness) matrix was estimated using known pedigrees (see **Box 1**) [349].

An alternative, increasingly popular solution is to estimate the relatedness matrix using genome-wide SNPs. The use of SNP-based relatedness matrices improves heritability estimates [350–352] and allows the user to better account for population structure [343, 353] compared to pedigree-based matrices. Different ways of estimating relatedness matrices from genotype data have been proposed [354–356]. Finally, a commonly-used approximation of the GRM is the 'realised' relatedness matrix (RRM) [352], which is defined as:

$$\text{RRM} = \frac{1}{M}\mathbf{G}\mathbf{G}^T, \tag{2.32}$$

[9] A genetic locus is IBD between two individuals if it has been inherited by a common ancestor.

where $\mathbf{G}$ is the $N \times M$ genotype matrix with standardised genotypes across individuals and M denotes the number of genome-wide variants. The RRM can be directly obtained from the polygenic model (where many genetic variants contribute to the trait):

$$\mathbf{y} \sim N(\mathbf{W}\boldsymbol{\alpha} + \mathbf{G}\mathbf{b}, \sigma_e^2\mathbf{I}), \tag{2.33}$$

where $\mathbf{b}$ is an $M \times 1$ vector containing the weights corresponding to $\mathbf{G}$ defined as above (standardised genotype matrix). If we assume that $\mathbf{b}$ is drawn from a normal distribution such that: $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \frac{\sigma_g^2}{M}\mathbf{I_M})$[10], and marginalising out the random effect we obtain the RRM in one of the terms of the covariance:

$$\mathbf{y} \sim N(\mathbf{W}\boldsymbol{\alpha}, \sigma_g^2\frac{1}{M}\mathbf{G}\mathbf{G}^T + \sigma_e^2\mathbf{I}), \tag{2.34}$$

The RRM can also be interpreted as an IBD relatedness matrix where the base population is the current population [357]. Throughout this thesis, we use the RRM described in [358] as implemented in PLINK [355].

2.3.3 Fast implementation of LMMs

The biggest limitation of the use of LMMs for association studies is that they are in general very computationally intensive. Computations associated with the parameter inference in LMMs scale cubically with the number of individuals in the dataset. Indeed, for the generic LMM the evaluation of the (restricted) marginal likelihood entails the computation of operations with $\mathcal{O}(N^3)$ complexity, specifically the inversion and the determinant of the total covariance. However, for the model in eq. (2.31), i.e. when the covariance matrix is known (does not depend on any parameters), and which can be alternatively expressed as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha} + \mathbf{g}\beta, \sigma_g^2\mathbf{K} + \sigma_n^2\mathbf{I}), \tag{2.35}$$

it is possible to speed up computations [343, 344, 359, 346], thus enabling application to large population studies. These stratagems reduce the computational complexity from $\mathcal{O}(N^3)$ per-SNP to a single $\mathcal{O}(N^3)$ cost upfront, and a per-SNP complexity of $\mathcal{O}(N^2)$. The complexity can be further reduced to $\mathcal{O}(N^2)$ for the upfront computation and a per-test complexity of $\mathcal{O}(N)$, provided the genetic relatedness matrix is low-rank.

[10] Note that each genetic marker explains on average variance $\frac{\sigma_g^2}{M}$, so that genome-wide variants jointly explain variance σ_g^2.

I use this section to briefly describe the efficient FaST-LMM algorithm proposed by Lippert *et al* [359]. This is the algorithm implemented within the LIMIX toolset [360, 361] which I use throughout this thesis.

The intuition is to project all data into a space where phenotypic variables ($\mathbf{y}$) and covariates ($\mathbf{W}, \mathbf{g}$) are uncorrelated so that in the rotated space the joint system to estimate the optimal model parameters can be solved in closed-form. To do so, we perform eigen decomposition of $\mathbf{K}$ from eq. (2.35), such that: $\mathbf{K} = \mathbf{QSQ}^T$, where $\mathbf{S}$ is a diagonal matrix containing the eigenvalues of $\mathbf{K}$ on the diagonal and zeroes elsewhere, and $\mathbf{Q}$ is orthonormal ($\mathbf{QQ}^T = \mathbf{I}$), with columns corresponding to the eigenvectors of $\mathbf{K}$. Then, if we define $\delta = \sigma_n^2/\sigma_g^2$ and consider it fixed, the full covariance matrix can be expressed as:

$$Var(\mathbf{y}) = \sigma_g^2\mathbf{K} + \sigma_n^2\mathbf{I} = \sigma_g^2(\mathbf{QSQ}^T + \delta\mathbf{I}) = \sigma_g^2\mathbf{Q}(\mathbf{S} + \delta\mathbf{I})\mathbf{Q}^T. \tag{2.36}$$

To simplify notation, we use $\mathbf{\Sigma} = \mathbf{K} + \delta\mathbf{I}$; $\mathbf{X} = [\mathbf{W}, \mathbf{g}]$ and $\boldsymbol{\beta} = [\boldsymbol{\alpha}, \beta]$, such that we can express the full log-likelihood as:

$$\ell(\boldsymbol{\beta}, \sigma_g^2, \delta) = -\frac{1}{2}\left\{N\log(2\pi\sigma_g^2) + \log|\mathbf{\Sigma}| + \frac{1}{\sigma_g^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{2.37}$$

Computationally, the two expensive operations are the calculation of the inverse of the covariance matrix $\mathbf{\Sigma}$ (i.e. $\mathbf{\Sigma}^{-1}$) and its determinant ($|\mathbf{\Sigma}|$). Both can be solved efficiently as follows:

$$\mathbf{\Sigma}^{-1} = [\mathbf{Q}(\mathbf{S} + \delta\mathbf{I})\mathbf{Q}^T]^{-1} = \mathbf{Q}(\mathbf{S} + \delta\mathbf{I})^{-1}\mathbf{Q}^T = \mathbf{QD}_\delta\mathbf{Q}^T, \tag{2.38}$$

where we define $\mathbf{D}_\delta = (\mathbf{S} + \delta\mathbf{I})^{-1}$ which is a diagonal matrix whose ith element is: $\frac{1}{S_{ii}+\delta}$ and use the property of orthonormality of $\mathbf{Q}$ (i.e. $\mathbf{Q}^{-1} = \mathbf{Q}^T$). As a result, this operation can be computed in linear time $\mathcal{O}(N)$. Next,

$$|\mathbf{\Sigma}| = |\mathbf{Q}(\mathbf{S} + \delta\mathbf{I})\mathbf{Q}^T| = |\mathbf{Q}||(\mathbf{S} + \delta\mathbf{I})||\mathbf{Q}^T| = |\mathbf{S} + \delta\mathbf{I}| = |\mathbf{D}_\delta{}^{-1}| = -|\mathbf{D}_\delta|, \tag{2.39}$$

where we used the property of orthonormality of $\mathbf{Q}$ ($|\mathbf{Q}| = 1$), and the determinant of a diagonal matrix ($|diag(\boldsymbol{\lambda})| = \prod_{i=1}^N \lambda_i$) as well as properties of the logarithm[11].

[11] $\log|\mathbf{D}_\delta{}^{-1}| = \log|\mathbf{S} + \delta\mathbf{I}| = \log(\prod_{i=1}^N(S_{ii} + \delta)) = \sum_{i=1}^N(\log(S_{ii} + \delta)) = -\sum_{i=1}^N \frac{1}{\log(S_{ii}+\delta)} = -\log\prod_{i=1}^N \frac{1}{(S_{ii}+\delta)} = -\log|\frac{1}{\mathbf{S}+\delta\mathbf{I}}| = -\log|\mathbf{D}_\delta|$

We can now re-write the log-likelihood as:

$$\begin{aligned}\ell(\boldsymbol{\beta},\sigma_g^2,\delta) &= -\frac{N}{2}\log(2\pi)-\frac{N}{2}\log(\sigma_g^2)+\frac{1}{2}\log|\mathbf{D}_\delta|-\frac{1}{2\sigma_g^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{Q}\mathbf{D}_\delta\mathbf{Q}^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{N}{2}\log(2\pi)-\frac{N}{2}\log(\sigma_g^2)+\frac{1}{2}\log|\mathbf{D}_\delta|-\frac{1}{2\sigma_g^2}(\mathbf{Q}^T\mathbf{y}-\mathbf{Q}^T\mathbf{X}\boldsymbol{\beta})^T\mathbf{D}_\delta(\mathbf{Q}^T\mathbf{y}-\mathbf{Q}^T\mathbf{X}\boldsymbol{\beta}) \\ &= -\frac{N}{2}\log(2\pi)-\frac{N}{2}\log(\sigma_g^2)+\frac{1}{2}\log|\mathbf{D}_\delta|-\frac{1}{2\sigma_g^2}(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T\mathbf{D}_\delta(\tilde{\mathbf{y}}-\tilde{\mathbf{X}}\boldsymbol{\beta}),\end{aligned} \tag{2.40}$$

where $\tilde{\mathbf{y}}=\mathbf{Q}^T\mathbf{y}$ and $\tilde{\mathbf{X}}=\mathbf{Q}^T\mathbf{X}$ are the rotated phenotype vector and covariate matrix (including the genotype vector $\mathbf{g}$).

If we consider δ fixed, the first and third terms are constant, so we can rewrite:

$$\ell = const-\frac{N}{2}\log(\sigma_g^2)-\frac{1}{2\sigma_g^2}[\mathbf{D}_\delta{}^{\frac{1}{2}}\tilde{\mathbf{y}}-\mathbf{D}_\delta{}^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}]^T[\mathbf{D}_\delta{}^{\frac{1}{2}}\tilde{\mathbf{y}}-\mathbf{D}_\delta{}^{\frac{1}{2}}\tilde{\mathbf{X}}\boldsymbol{\beta}]. \tag{2.41}$$

Then, we can compute $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma_g^2}$ using eq. (2.8) and eq. (2.9) for the linear model:

$$\mathbf{D}_\delta^{\frac{1}{2}}\mathbf{Q}^T\mathbf{y}\sim\mathcal{N}(\mathbf{D}_\delta^{\frac{1}{2}}\mathbf{Q}^T\mathbf{X}\boldsymbol{\beta},\sigma_g^2\mathbf{I}). \tag{2.42}$$

To further speed up computations, $\boldsymbol{\alpha}$ and δ are only optimised once, under H_0, which is the same for all SNPs for a given trait. Since there is no closed-form for δ, we use the Brent search numerical procedure [362] to find the optimal $\hat{\delta}$. For every SNP, we find the MLEs for β and σ_g^2 using the closed-form from equations (2.8),(2.9).

When the rank of $\mathbf{K}$ is much smaller than the number of individuals $rank(\mathbf{K}) << N$, we can further speed up the algorithm by computing the 'economical' eigen decomposition [359], where we use the fact that most eigenvalues of $\mathbf{K}$ are 0:

$$\overbrace{\begin{bmatrix}\mathbf{Q_0} & \mathbf{Q_1}\end{bmatrix}}^{\mathbf{Q}}\overbrace{\begin{bmatrix}\mathbf{S_0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}}^{\mathbf{S}}\begin{bmatrix}\mathbf{Q_0}^T\\ \mathbf{Q_1}^T\end{bmatrix}=\mathbf{K}. \tag{2.43}$$

For details on this implementation I refer the reader to the Supplementary note from Lippert *et al.* [359].

Whilst extremely efficient, I note that the described implementation is limited to one random effect term only (besides the identity, see eq. (2.36)). Indeed, the stratagem used to reduce the computational complexity and calculate the inverse (eq. (2.38)) and the determinant (eq. (2.39)) of the covariance matrix in quadratic (or linear, in case of low-rank kinship matrix) time only works for a single kinship matrix[12]. While I am aware of alternative implementations that allow the use of several kinship matrices (e.g. [363]), these are currently optimised for SNP-heritability estimates, and the application to association testing is not trivial. Thus, in this thesis, I only present LMMs with one single random effect term.

2.3.4 Modelling non-Gaussian data

Linear regressions and linear mixed models are established approaches for mapping associations between genotype and phenotype. One limitation of these models, however, is that they assume the residual noise to be normally distributed, a premise that rarely holds in practice. Deviation from normality can result in model mis-specification, which in turn can lead to false conclusions and reduced statistical power [364].

To alleviate this issue, it is common to preprocess the phenotype vector ($\mathbf{y}$) to 'gaussianise' it. For example, common approaches involve applying different kinds of non-linear transformations to $\mathbf{y}$. Log-transformations are universally used on gene expression values (measured using RNA-seq) to reduce the range of possible assumed values. In addition, several other 'variance-stabilising' transformations can be applied later, including Box-Cox transformations [365] or rank transformations [366]. More recently, warped LMM [367] was proposed, which learns from the data the optimal phenotype transformation to be applied prior to testing for associations using an LMM.

Note that these models assume that the noise level in the transformed phenotype space is constant, which may not be an appropriate assumption in some cases. In such instances, it will remain appropriate to use generalised LMMs (GLMMs[13]) with non-Gaussian likelihoods that enable to incorporate stronger assumptions about the nature of the data [367, 368].

[12] In case of two distinct kinship matrices, i.e. $\sigma_1^2\mathbf{K}_1 + \sigma_2^2\mathbf{K}_2 + \sigma_n^2\mathbf{I} = \sigma_1^2\mathbf{QSQ}^T + \sigma_2^2\mathbf{UDU}^T + \sigma_n^2\mathbf{I}$, which cannot be further simplified.

[13] Generalised linear mixed models (or GLMMs) are an extension of linear mixed models to allow response variables from different distributions, as long as those belong to the exponential family. Common example of exponential family distributions are the Bernoulli, Exponential, Gamma, Normal and Poisson distributions.

2.3.5 Linear mixed models for eQTL mapping

In this thesis, we perform different sets of eQTL mapping, adopting the FaST-LMM algorithm (**section 2.3.3**) as implemented in LIMIX. The expression of each gene is normalised and log transformed (more details are specified in the various applications from the next chapters).

Next, when running the FaST-LMM algorithm in the context of eQTL mapping, we run the model for each gene separately. For a given gene, the model under H_0 is fixed. In order to perform *cis* eQTL mapping, we define a *cis* window around the gene (typically ranging from 100 kb to 1 Mb) and a minimum MAF (e.g. 0.05) and test all SNPs within the window with MAF > 0.05 in our population. We test each SNP independently, with only the alternative model changing. This means that all weights $\boldsymbol{\alpha}$ excluding the SNP effect, as well as δ are only computed once per gene.

Typically, for each gene-SNP pair tested the reported values are i) the p value (e.g. obtained using eq. (2.22)), ii) the estimated effect size β, and iii) the effect size standard error ($se(\beta)$). Subsequently, multiple testing correction is applied, first at the gene level (using FWER) and then globally (using FDR, as described in **section 2.2.2**). The resulting adjusted p values are also reported. To summarise a typical eQTL map, a number frequently reported is the number of genes with at least one eQTL (from now on 'eGenes'), often reported together (or as a fraction of) the number of genes tested.

In the analyses in **Chapters 3-5** of this thesis, I report significant results (i.e. number of discoveries) at FDR<5% or 10%, which are commonly used in the field [294, 150, 151]. This allows me to put these numbers in context, particularly to assess differences in number of discoveries between results obtained using single cell data (from this thesis) with existing eQTL results (typically obtained using bulk RNA-seq data).

2.4 Linear Mixed Models for interaction tests

The statistical test described in **section 2.2.1** refers to an 'association test' where the aim is to test for an effect that the SNP of interest has on a trait directly[14]. However, the same LMMs and fast implementation approaches described so far can be applied, with care, to testing for genotype-environment (GxE) interactions instead ('interaction test' for simplicity).

[14]Note that this was described for a linear regression but can be equivalently applied to the linear mixed model, i.e. $H_0 : \mathbf{y} \sim N(\mathbf{0}, \sigma_g^2\mathbf{K} + \sigma_n^2\mathbf{I})$ vs $H_1 : \mathbf{y} \sim N(\mathbf{g}\beta, \sigma_g^2\mathbf{K} + \sigma_n^2\mathbf{I})$, where the test assesses $\beta = 0$ vs $\beta \neq 0$.

In particular, a significant statistical GxE interaction is defined when there is a significant difference in the genetic effect between groups of individuals with different environmental exposures (**Fig. 2.4**).

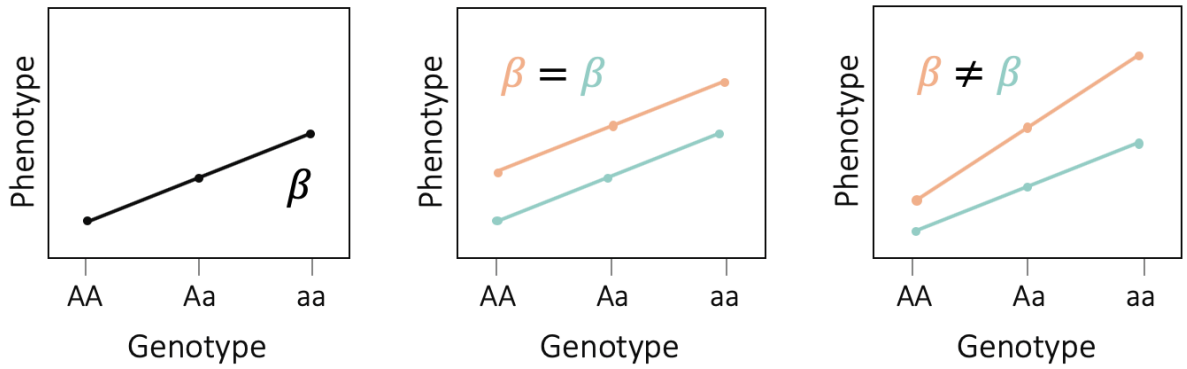


Fig. 2.4: Illustration of GxE effect for two environmental groups.
Illustrated are the mean phenotype values across three genotypic groups for a given locus (AA: homozygous major allele, aA: heterozygous individuals, aa: homozygous minor allele). The first plot (left) describes the case of a genetic effect of the variant on all individuals. β represents the (in this case positive) effect size of the locus tested. The second plot (middle) shows that two groups of individuals characterised by different environmental exposure (represented by the colours). The environments have an effect on phenotype (as represented by a shift upwards for the orange group) but the genetic effect remains constant. Finally, the third plot (right) shows a GxE effect. There is an interaction effect between the individuals' genotypes and their environmental exposure, such that for one group (orange) the genetic effect was exacerbated whilst for the other group (seagreen) the effect was dampened.

In the following, I describe various approaches to test for GxE interactions, which all build on the LMM framework. One possible way to detect interaction effects is to stratify samples into discrete subgroups based on their environmental exposure. In eQTL mapping, one might for example cluster cells into cell types, or separate samples into different condition groups. Then, an LM (eq. (2.29)) or LMM (eq. (2.31)) can be applied to each stratum and the marginal variant effects can be compared to assess whether there is a significant difference in these effects across the sub-populations. This method can be defined as a 'stratified interaction test'. However, as more detailed environmental data is collected, allowing for finer stratification of the population, these methods are no longer optimal as the sub-populations become too small to obtain stable estimates of the variant effects. For example, as more and more rare cell types are identified and as the definition of cell types becomes more blurred, joint analyses may be preferable.

Another commonly-used method to test for interaction effects is an extension of the LM or LMM to include two additional FE terms, an interaction term (GxE) and an environment term (E):

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{e}\gamma + \mathbf{g}\beta_G + \mathbf{e} \odot \mathbf{g}\beta_{GxE} + \mathbf{u} + \boldsymbol{\psi}, \tag{2.44}$$

where **e** represent the environment term, γ is its corresponding weight and where two genetic effect terms are present, β_G which is the 'persistent' genetic effect size whilst β_{GxE} is the effect of the interaction term (GxE). Finally, $\odot$ denotes element-wise multiplication (Hadamard product). The test is:

$$H_0 : \beta_{GxE} = 0 \tag{2.45}$$

vs

$$H_1 : \beta_{GxE} \neq 0. \tag{2.46}$$

This model allows to directly test for a GxE interaction effect with a certain environment/factor[15], beyond the additive effects of the SNPs and the environment themselves. Of course, the same model can also be extended to the case of multi-environment interaction by simply adding more environments as FE terms:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{e}_1\gamma_1 + \mathbf{e}_2\gamma_2 + ... + \mathbf{g}\beta_G + \mathbf{e_1} \odot \mathbf{g}\beta_{GxE,1} + \mathbf{e_2} \odot \mathbf{g}\beta_{GxE,2} + .. + \mathbf{u} + \boldsymbol{\psi}. \tag{2.47}$$

However, this becomes quickly infeasible for large numbers of environments, especially for relatively small sample sizes, because of the degrees of freedom loss when increasing the number of parameters to estimate (two additonal parameters to estimate for every environment included).

Finally, in the specific context of measuring genetic effects on gene expression, and testing for interactions between environmental exposures and such genetic effects, an alternative approach is to model the effect that these environments have on allele-specific expression (ASE, see **Box 3** in the next chapter). I describe this approach in more detail upon its application in **Chapter 4** (**section 4.5**).

[15] Or with another genetic variant, to test for epistasis [369].

2.5 Discussion

Linear mixed models are widely applied in genetic association analysis because they offer great control for confounding effects. Throughout this thesis, I will use different models for eQTL mapping using single cell expression profiles, which all build on this framework. I have used this chapter to lay the foundation and notation for which we build extensions in the coming chapters.

Specifically, in **Chapter 3**, I provide a best-practice pipeline for performing bulk-like eQTL mapping using single cell data, where I test various aggregation methods as well as design matrices and then use the standard linear mixed model from eq. (2.31) to map eQTL. I compare our results to an eQTL map obtained using bulk RNA-seq from the same samples and also compare results across scRNA-seq technologies for a subset of donors. Next, in **Chapters 4** and **5** I use different adaptations of linear and linear mixed models to test both for associations (eq. (2.31)) and interactions (eq. (2.44)) in two separate population-scale human scRNA-seq datasets.

3

Comparison of eQTL mapping using bulk and single cell RNA-seq readouts

As discussed in **section 1.1.8**, in traditional eQTL mapping individual-level gene expression is measured using bulk RNA-sequencing, where the quantified expression profiles represent several thousands of cells from each individual. As we have seen, recent technological advances have allowed molecular phenotypes, including gene expression, to be assayed at the level of single cells. In particular, scRNA-seq is now an established technique, and can be deployed at population-scale, across several individuals. Owing to their ability to identify cell types and cell states in a data-driven manner, scRNA-seq data from a single experiment can be used to define homogeneous cell populations, quantify expression levels within them, and then map eQTL in each of them separately. As a consequence, studies where single cell expression profiles (rather than bulk) are used to perform eQTL mapping have emerged recently, and promise to greatly improve our understanding of the genetic architecture of gene regulation across tissues, in both human disease and development. However, the use of single-cell RNA-seq to map eQTL maps as opposed to using bulk RNA-seq profiling has not been systematically benchmarked. To address this, here I select a very homogeneous cell type (iPSCs), where direct comparison is possible. In particular, I use matched bulk and single cell RNA-seq from over 100 human iPSC lines to assess our ability to detect eQTL using single cell RNA-seq data, and the extent to which we can replicate eQTL identified using bulk RNA-seq, when mapping eQTL using a common analytical framework based on LMMs (see **Chapter 2**). Additionally, for a subset of lines, I compare results using two different scRNA-seq technologies. As more and more single cell-eQTL (sc-eQTL) studies emerge, I believe that the insights presented here will help establishing a 'best practice' workflow, to optimise yield of sc-eQTL maps and to establish uniform methods across the field.

Contributions In this chapter, I present results from two main bodies of work.
First, I describe a set of results from analyses I have conducted in the context of a larger collaborative project between the Stegle, Marioni and Vallier labs. In particular, the data was generated by Ludovic Vallier's lab at the Sanger Institute, and the experiments were led by Mariya Chhatriwala, using cell lines from the HipSci project. Davis McCarthy and I processed the scRNA-seq data and performed quality control. Marc Jan Bonder processed the bulk RNA-seq data. I performed all analyses presented in the first part of this chapter, under the supervision of Oliver Stegle and John Marioni. The code for processing, analysing and plotting the data is open source and freely accessible here: https://github.com/single-cell-genetics/singlecell_endodiff_paper. The analyses presented here are part of the following paper, which is available at https://www.nature.com/articles/s41467-020-14457-z and has been published as:

Anna S.E. Cuomo*, Daniel D. Seaton*, Davis J. McCarthy*, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, Andrew Knights, Kedar Nath Natarajan, the Hipsci Consortium, Ludovic Vallier, John C. Marioni, Mariya Chhatriwala, Oliver Stegle. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, 2020, (* equal contribution).

In the second part of this chapter, I present more recent preliminary results from work in collaboration with Giordano Alvari and Marc Jan Bonder, from the Stegle lab[a]. This project was designed by Marc Jan Bonder and myself, to extend the results presented in the first part of this chapter. Giordano performed most of the analyses, under mine and Marc Jan Bonder's supervision. Marc Jan Bonder and I performed the remaining analyses and summarised the results.

I generated all figures included in this chapter.

[a]Note: an updated version of this work is now available as a preprint at https://www.biorxiv.org/content/10.1101/2021.01.20.427401v1.

3.1 Introduction

As outlined in **section 1.1.8**, since the publication of the first human eQTL map in 2003 [109], the field has adapted to adopt technological advances as they emerged, both in terms of statistical approaches (from linkage analysis to genome-wide scans), and of new technologies for the estimation of gene expression (from microarrays to RNA-seq). I use this section to give a brief overview of methods to estimate gene expression (**section 3.1.1**), the advent of single cell RNA-seq (**section 3.1.2**) and the consequent birth of the (very new) single cell-eQTL mapping field of research (**section 3.1.3**).

3.1.1 Measuring gene expression

Early methods for estimating the number of expressed mRNA molecules associated with a particular gene (hereafter defined as a gene's expression) were Northern blots [370] and quantitative reverse transcription polymerase chain reaction, qRT-PCR [371]. In Northern blots, electrophoresis is used to separate RNA, which is then visualised by hybridisation with labelled probes. A key limitation of Northern blots is that they require large amounts of input material and the results are mostly only qualitative. In qRT-PCR, RNA is reverse-transcribed into complementary DNA (cDNA) and then amplified using PCR, after each cycle of which the concentration of DNA is measured using a fluorescent dye. This required normalisation relative to a standard gene (e.g. a housekeeping gene, or a ribosomal gene) which was assumed to be 'stable', making this method also not very quantitative. Additionally, both of these methods were very low-throughput, typically used only to quantify one, or at most few genes - hence being referred to as 'single gene approaches'.

In 1995, DNA microarrays were introduced [372], which for the first time allowed the estimation of expression levels for many genes simultaneously. Like qRT-PCR, microarrays rely on the reverse transcription of RNA into cDNA. This cDNA is then labelled with a fluorescent dye and hybridised to a DNA microarray containing complementary DNA for thousands of known transcripts at specific locations. The RNA levels can then be estimated by measuring the intensity of fluorescence at each location and either normalising it using RNA transcripts of known concentration (called RNA spike-ins; 'one colour array') or directly comparing it to a second sample on the same microarray using two different fluorescent dyes ('two colour array'). Microarrays quickly became the most commonly used method for measuring gene expression levels. However, since microarrays only allow the measurement of RNA with a known sequence, they are not suitable for the detection of novel transcripts or alternative splice isoforms.

In the late 2000s, RNA sequencing (RNA-seq), based on next-generation sequencing (NGS), was introduced. RNA-seq allows for the direct sequencing and quantification of cDNA libraries [373] and has since been shown to be superior to microarrays in almost all regards [374], as it provides information about a gene's sequence, in addition to its expression level. In particular, in addition to the quantification of known transcripts, RNA-seq also enables the identification of completely new genes, previously unknown genetic variants in the genes, variation in alternative splicing, or post-transcriptional modifications (see also **section 1.1.8**).

In recent years, next generation sequencing approaches have been extended to quantify variation in RNA expression at single cell resolution, starting the 'single cell RNA-seq era'.

3.1.2 The 'resolution revolution'

The first single-cell RNA sequencing (scRNA-seq) experiment was published in 2009, and it involved profiling of only eight cells [375]. Seven years later, 10X Genomics released a dataset of 1.3 million cells [376]. All in all, in the last decade, over 1,000 scRNA-seq datasets have been published [377–379], using a number of different technologies [380–390] (**Fig. 3.1**).

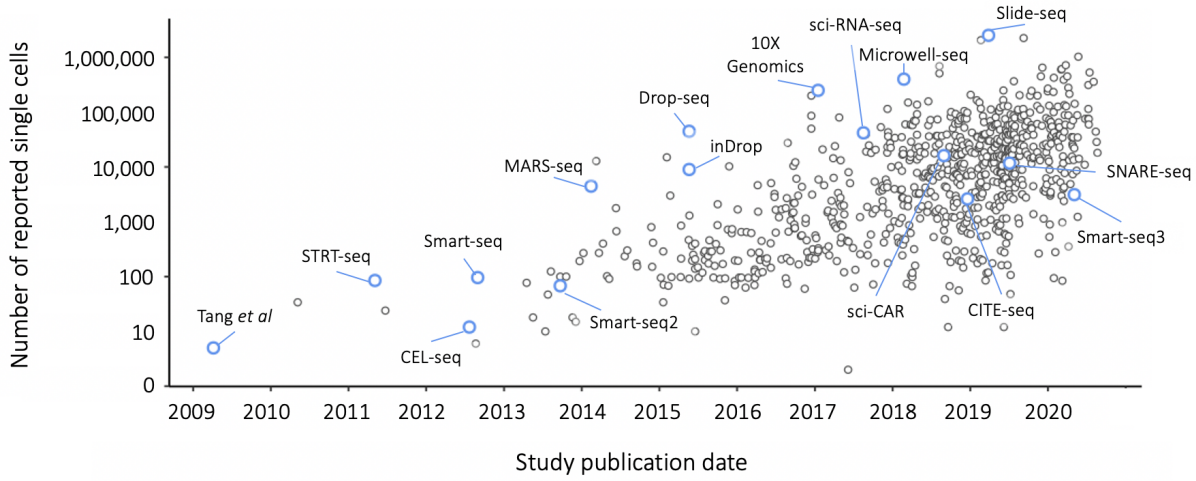


Fig. 3.1: Scale of scRNA-seq experiments.
Number of single cells reported in all scRNA-seq publications to date (as collected in [379], y axis), ordered by publication date (x axis). Key scRNA-seq methods are indicated. Similar to [377].

Single cell RNA-seq protocols differ extensively in terms of scalability, costs and sensitivity [391, 377]. However, they can be broadly categorised into methods that are 'plate-based' or

'droplet-based', based on the capture technology used (**Fig. 3.2**).

Initially, most studies used plate-based assays, where cells are isolated using micropipettes or flow cytometry into individual wells of a plate, where the library preparation is performed (**Fig. 3.2**). This class of methods include single-cell tagged reverse transcription sequencing (STRT-seq [380]), Cell Expression by Linear amplification and Sequencing (CEL-seq [381]), massively parallel single cell RNA-seq (MARS-seq [385]) and Smart-seq [382, 383, 390].

On the other hand, droplet-based methods employ microfluidics to capture individual cells in nanolitre-sized droplets, each loaded with all the necessary reagents for library preparation. The droplet suspension is later broken down for pooling of cell libraries prior to sequencing (**Fig. 3.2**). These methods have been developed by academic groups (InDrop [387] and Drop-seq [386]) and commercially, by 10X Genomics (Chromium [389]). These protocols share similar technologies, particularly the use of unique molecular identifiers (UMIs) to correct for biases in PCR amplifications [392].

Each approach has its own advantages and disadvantages. The main advantage of plate-based methods is the higher quality of libraries and, in the case of Smart-seq, the full length transcript information which enables the quantification of splice variants [393], allele-specific expression [394] and RNA velocity information [395]. However, this comes at the expense of lower cellular throughput, processing hundreds or thousands of cells compared to the hundreds of thousands that droplet-based methods can achieve. Indeed, by capturing cells in individual droplets, each containing all necessary reagents for library preparation, droplet-based protocols allow the profiling of thousands or even millions of cells in a single experiment. This, however, comes at the cost of reduced sensitivity. Additionally, current droplet methods capture gene information exclusively from the 3' or 5' end of each transcript, and are more likely to produce 'doublets', where two different cells become labelled with the same barcode (**Fig. 3.2**).

In the last 10 years, technological improvement (**Fig. 3.1**) has gone hand-in-hand with computational advances to analyse the resulting data, which require a new set of considerations that were not relevant for bulk RNA-seq data. Indeed, to complement the explosion of scRNA-seq studies published, an entire ecosystem of computational methods for analysing them has emerged. In some cases, those methods have been directly borrowed from bulk RNA-sequencing methods; other times, methods tailored specifically for single cell data were proposed [397–399].

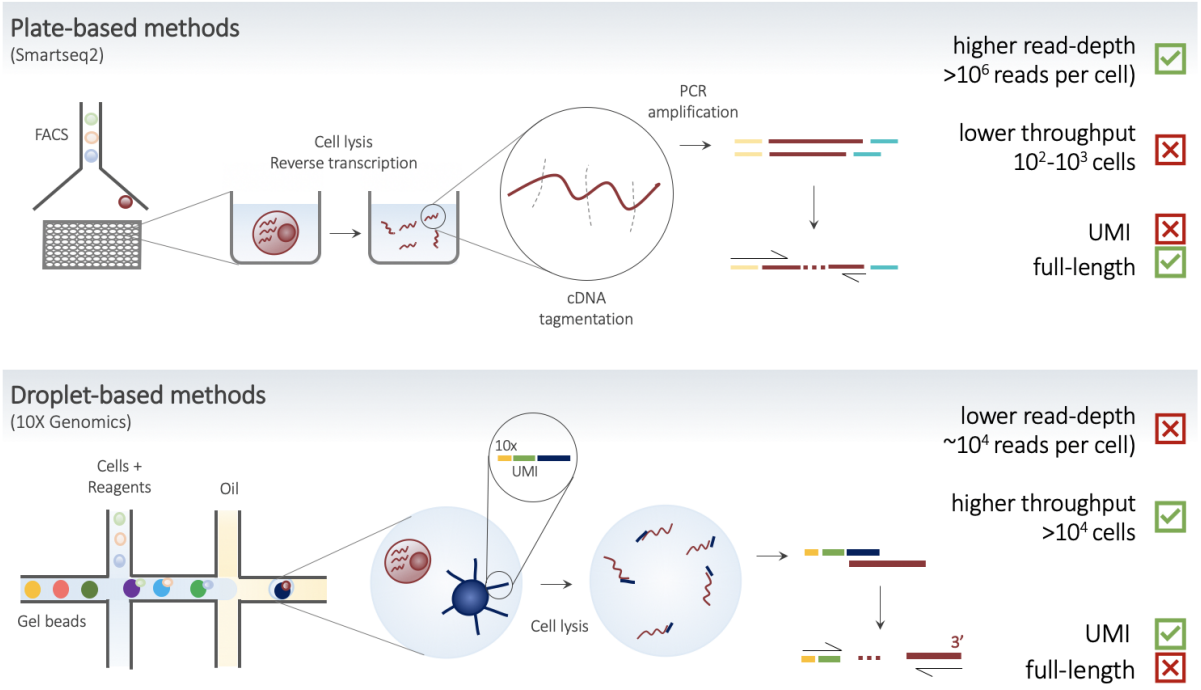


Fig. 3.2: Plate-based vs droplet-based methods for scRNA-seq.
An illustration of the key differences between plate-based methods (exemplified by SmartSeq2 [383]) and droplet-based methods (represented by 10X Genomics Chromium [389]). The key trade-off is between the cell-throughput (much higher for droplet-based methods) and the read-depth per cell (higher for plate-based methods). Additionally, the full-length transcripts obtained using SmartSeq2 allow quantification of allele-specific expression and splice variants, which are not possible with 3' tag 10X data. Finally, all droplet-based methods include unique molecular identifiers (UMIs), which allow the robust quantification of PCR duplicates. Note that these last two differences (in terms of UMIs and full length) specifically hold true for the two methods shown here (and used in this thesis: SmartSeq2 and 10X Genomics). Indeed, not all plate-based methods provide full-length transcript information (e.g. MARS-seq [385] and CEL-seq [381] do not). In contrast, the most recent SmartSeq3 [390] can include UMIs, despite being a plate-based method. Figure similar to [396].

Single cell-specific bioinformatics workflows such as Cell Ranger [389], indrops [387], SEQC [400], or zUMIs [401] have been developed to perform raw data processing tasks, i.e. read-level QC, assignment of reads to their cell barcodes and mRNA molecules of origin (i.e. 'demultiplexing'), alignment to the reference genome, and quantification. Additional methods allow the assignment of cells to their donor of origin, in case of multi-individual pooled designs [402, 403]. The data resulting from a scRNA-seq experiment are typically represented as an integer matrix of gene expression levels, with entries representing the number of sequenced reads (or molecules, if UMIs were used) assigned to a particular gene in a specific cell [396]. Starting from these count matrices, a common scRNA-seq analysis workflow may be divided into pre-processing steps and downstream analysis [399] - and scRNA-seq-specific tools have been implemented for several of the steps along the pipeline.

In particular, methods have been proposed to perform cell calling, i.e. to detect, and exclude, empty droplets [404], doublets [405–407], and ambient RNA [408]. Moreover, methods for normalisation have been described in [409–411]. After normalisation, data matrices are typically log(x+1)-transformed. Additionally, several novel methods allow to correct for confounding factors including batch effects [412–417] and cell cycle effects [418, 419]. To ease the computational burden on downstream analysis tools, reduce the noise in the data, and to visualise the data, one can use several approaches to reduce the dimensionality of the dataset. First, feature selection, for example by detecting highly variable genes (HVGs) [420, 421]. Next, dimensionality reduction is performed either using linear methods, such as PCA, or non-linear methods, with the latter being preferred for visualisation purposes. In particular, t-distributed stochastic neighbour embedding (tSNE) [422] and uniform manifold approximation and projection (UMAP) [423] are extremely popular (for a review of other methods, see [424]). Downstream analysis methods can be classified into cell-level and gene-level. The former include clustering [425, 426], often followed by cell type annotation [427], as well as pseudotime inference [428–431]. Finally, single cell-specific methods have been developed for gene-level analyses, including differential expression analysis [432], and gene regulatory networks identification [433–435].

A typical workflow for single cell RNA-seq data implemented in R can be found on Bioconductor[1] using scRNA-seq specific R packages scran [436, 437], scater [438], and SingleCellExperiment [439]. Other pipelines for scRNAseq data analysis include Seurat [413], Scanpy [440], and SINCERA [441].

[1] at https://bioconductor.org/packages/devel/bioc/vignettes/scran/inst/doc/scran.html and https://osca.bioconductor.org.

3.1.3 Single cell eQTL mapping

With the ability to identify cell types and states in an unbiased manner [442, 443], the use of scRNA-seq data, combined with genotype information, is uniquely positioned to provide an extra layer of information on the regulatory role of common genetic variants on gene expression, across a plethora of cell types and states. As a consequence, single cell eQTL mapping is increasingly feasible, and promises to improve our understanding of genetic regulation both in health and disease across tissues [159, 160, 444–447].

When performing eQTL mapping using scRNA-seq profiles, a first important step is to verify the feasibility of traditional 'mean level' eQTL mapping, i.e. to reproduce eQTL previously identified using bulk RNA-seq. Only then can we explore new avenues and alternative types of eQTL analyses, which are especially enabled by the single cell resolution (**Fig. 3.3**).

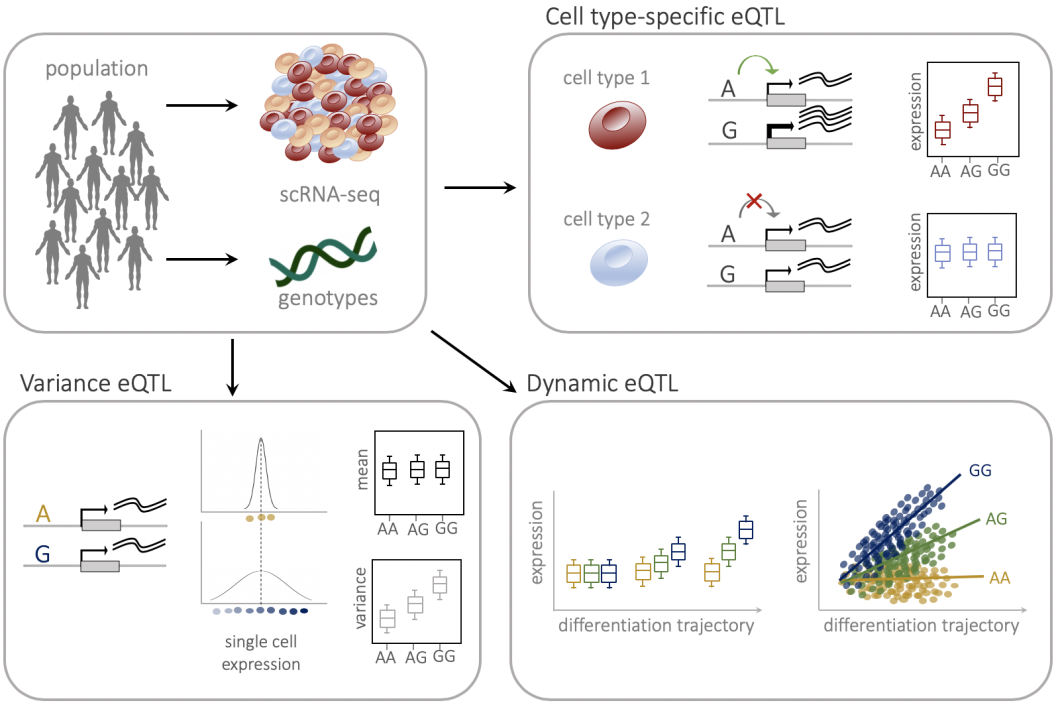


Fig. 3.3: Overview of single cell eQTL mapping methods.
Matched genotypes and scRNA-seq data from several individuals allow the detection of cell type-specific eQTL, variance eQTL (genetic effects on cell-to-cell transcriptional variability), and dynamic eQTL (dynamic genetic effects along cellular differentiation or other cellular states).

In this chapter, we address the first point (i.e. mapping mean-level sc-eQTL). To do so, we leverage bulk and single cell gene expression of matched human iPSC lines from around 100 donors to identify general guidelines for eQTL mapping using scRNA-seq data.

3.2 What is different in single cell data?

When we perform eQTL mapping, we are interested in finding differences in expression level between individuals, when stratified by their genotypes at a genomic locus of interest (**page 19**). Under the assumption that we are looking at an otherwise homogeneous population of cells (e.g. all cells are from the same cell type), it is reasonable to consider the total (or the average) expression for each individual and gene, across all cells. When we use bulk RNA sequencing expression profiles, that is essentially what happens: all cells from an individual are pooled, the mRNA is extracted, reverse-transcribed to cDNA, and then sequenced. The resulting reads are then mapped onto a reference genome, and the expression level of each gene is quantified as the number of reads (raw counts) obtained from one donor that uniquely map to that gene, after normalisation, e.g. transcripts per million (TPM)[2]. A bulk RNA-seq experiment, therefore, results in one individual measure of 'abundance' of each gene for each donor. Such a measure results in aggregating over hundreds of thousands of cells and, at least for expressed genes (e.g. average TPM > 1), the vector of gene expression across individuals follows a distribution that can be approximated as Gaussian [448]. On the other hand, whilst scRNA-seq data provides increased resolution and promises great insights into cellular function, the data are also much sparser, and the number of cells that can be assayed for an individual is limited compared to bulk (often as little as 10-100 cells). In addition, the number of cells that can be assessed often varies substantially from individual to individual. As a result, the distribution of total counts from a single cell experiment as opposed to its corresponding bulk experiment has lower mean (fewer cells, fewer reads, **Fig. 3.4**) and higher variance (due to the variable number of cells across donors, **Fig. 3.4** vs **Fig. B.1**).

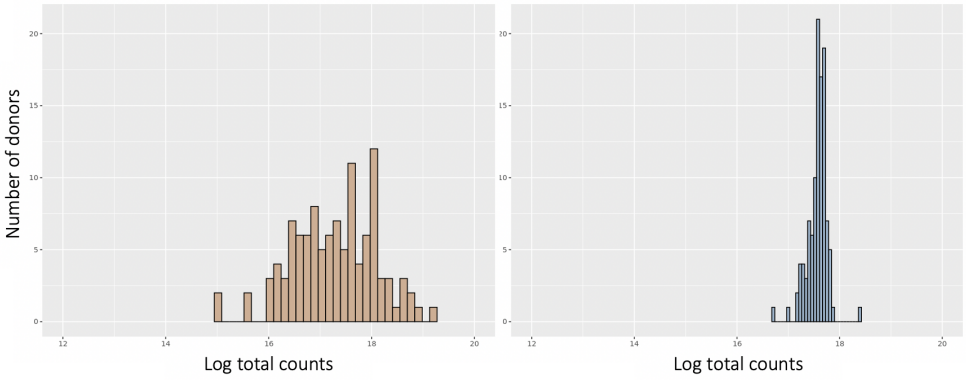


Fig. 3.4: Distribution of reads.
Distribution of total reads (across all genes, log10-transformed) per individual for matched iPSC data (108 individuals) using single cell (left, from [447]) and bulk (right, from [449]) RNA-seq data.

[2]i.e. for every 1,000,000 RNA molecules in the RNA-seq sample, x came from this gene/transcript.

3.3 Single cell and bulk RNA-seq profiling of iPSCs

The data I use in this chapter to benchmark methods for single cell eQTL mapping were generated as part of a larger study, where iPSC lines from over 100 donors (from HipSci) are differentiated towards definitive endoderm. A pooled design was adopted, where cells from 4-6 lines were differentiated together, to avoid for individual genetic differences to be confounded with batch variation, and to increase throughput. Cells were later assigned to their donor of origin using Cardelino [403]. Next, cells were collected at four time points (day0, day1, day2, day3) and sequenced using SmartSeq2 [383], a plate-based single cell technology (**Fig. 3.2**). This study was published earlier this year [447], and I discuss the key results from it in the next chapter (**Chapter 4**).

Here, I focus on the earliest time point (i.e. day0), where cells are still pluripotent, prior to cell differentiation. We expect iPS cells to be fairly homogeneous, so it is the ideal cell type to use to perform this kind of study. After QC[3], data was available from 9,661 iPS cells and 11,231 genes, from 112 unique unrelated donors, across 24 differentiation pools (**Fig. 3.5**).

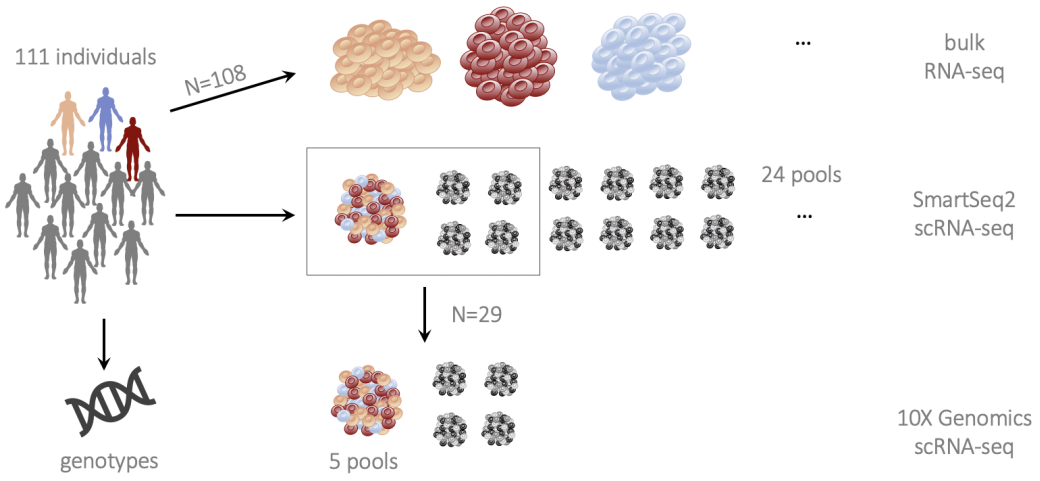


Fig. 3.5: Overview of iPSC data used in this chapter.
We use SmartSeq2 [355] data from 111 iPSC lines (from 111 individuals) from [447], across 24 experimental pools, each containing cells from 4-6 cell lines (middle). For 108 of those lines, we have bulk RNA-seq profiles from [449] (top). Finally, for five of the pools (corresponding to 29 individuals/lines), we also have scRNA-seq data sequenced using the 10X Genomics pipeline [389] (bottom).

[3]Some QC steps were performed for all time points jointly, therefore I refer the reader to the detailed QC pipeline described in the next chapter, at **page 105**.

To map eQTL, I considered only donors with at least 10 cells, which excluded one donor, bringing the sample size down to 111. The number of cells per individual varied widely, ranging from 10 to 383 iPS cells from individual to individual. Additionally, we have matched bulk RNA-sequencing profiles generated as part of the HipSci project [294] for the vast majority of cell lines used (108/111, 97%). Finally, we have scRNA-seq data measured using a droplet-based technology (10X Genomics [389]) for a subset of common lines (29 iPSC lines, from five of the experimental pools, **Fig. 3.5**).

3.4 eQTL mapping pipeline

To map eQTL, I use a pipeline which was originally written by Marc Jan Bonder, and which I have expanded to the use for single cell eQTL. It is a wrapper around LIMIX [360, 361], and it is publicly available at https://github.com/single-cell-genetics/limix_qtl. For all three eQTL maps (single cell SmartSeq2, bulk, single cell 10X), I used the same pipeline and tested the same set of genes (n=10,840). In particular, I performed *cis* eQTL mapping, considering common (minor allele frequency > 5%), in HWE (p value > 0.001)[4], variants within a *cis*-region spanning 250 kb upstream and downstream of the gene body. For each gene (**y**) SNP (**g**) pair, the association test was performed using an LMM (**section 2.3**):

$$\mathbf{y} = \sum_{i}^{P} \alpha_i \mathbf{PC}_i + \mathbf{g}\beta + \mathbf{u} + \boldsymbol{\psi}, \tag{3.1}$$

where $\mathbf{y}$ is the $N \times 1$ standardised[5] expression-level phenotype vector (i.e. bulk expression or mean single cell expression; details in next section); the first P=10 PCs calculated on the expression values (matrix with $\mathbf{y}$ as columns, before standardisation) are included as fixed effect covariates[6], and α_i are the corresponding weights; $\mathbf{g}$ is the $N \times 1$ vector of alleles for each sample at the locus tested (modelled as the number of minor alleles present - 0, 1 or 2), and β is the corresponding effect size; $\mathbf{u}$ is a random effect term used to account for the samples' population structure, i.e. $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{K})$, where $\mathbf{K}$ is an $N \times N$ kinship matrix estimated using PLINK [355], and $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I})$, where $\mathbf{I}$ is the $N \times N$ identity matrix, is the noise vector.

[4] HWE: Hardy-Weinberg equilibrium.

[5] Centered at 0 and scaled to have variance = 1.

[6] This is a common approach to correct for both known and unknown unwanted variation, including batch effects, which usually affect the expression of many genes, and therefore are detectable in the principal components of expression. Moreover, these global effects are orthogonal to the effects of a single variant on the expression of one gene (see **section 2.2.4**).

All models were fitted using LIMIX [360, 361]. Note that, in order to map eQTL efficiently using the fast implementation described in **section 2.3.3**, we are limited to a single random effect term (**u**; rather than being able to incorporate for example pool as a random effect). The significance was tested using a likelihood ratio test (i.e. $\beta \neq 0$, **section 2.2.1**). In order to adjust for multiple testing (**section 2.2.2**), we used a permutation scheme, analogous to the approach proposed in Ongen *et al.* [318]. Briefly, for each gene, we generated 1,000 permutations of the genotypes while keeping covariates, kinship, and expression values fixed. We then adjusted for multiple testing using this empirical null distribution. To control for multiple testing across genes, we then applied the Storey procedure [323]. Genes with significant eQTL were reported at FDR < 10%.

3.5 Single cell eQTL map of iPS cells

Using the method just described, we first tested for associations between common genetic variants and gene expression in iPSCs using our SmartSeq2 single cell data. To reproduce bulk-like abundance measurements, we considered a gene's average expression level for each sample, across cells. In particular, expression level is measured as log2(CPM+1)[7] using scater [438].

Since we did not use any batch correction method on the single cell expression data a priori, we cannot exclude differences across batches. As internal control we have, for a subset of donors (23/111), data from two (or three, in one case) distinct experimental batches. We therefore compute average expression levels not for each individual line, but for each line-experiment combination (i.e. cell_lineA-experiment1, cell_lineA-experiment2), which enables effective correction for batch-to-batch differences using PCs as covariates (see above and **section 2.2.4**). The linear mixed model described in eq. (3.1) can be readily adapted to include the resulting replicate measures from the same line: N will now be not the number of unique lines but that of line-experiment combinations. Additionally, both the genotype vector **g** and the kinship matrix **K** need to be adjusted[8], such that the latter is now in fact accounting at once for the population structure and the repeatedness of the samples tested (**Fig. 3.6**).

Using this approach, I identified 1,833 genes with at least one eQTL (from hereon 'eGenes'), at FDR <10%, out of 10,840 genes tested (17%).

[7] CPM: counts per million, mapped reads are count-scaled by the total number of fragments sequenced per cell, times one million.

[8] i.e. expanded, by duplicating genotype values across pool replicates.

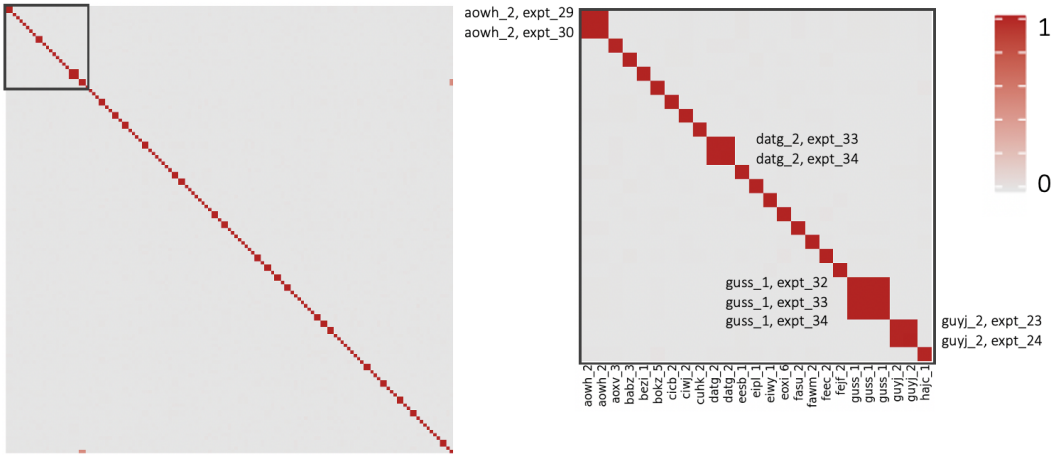


Fig. 3.6: Kinship matrix highlighting repeated structure of samples used.
Heatmap of the kinship matrix used to map eQTL using single cell data. Replicate observations for a line across experimental batches are genetically identical, which is captured by the kinship (maximum relatedness, red). Line-to-line relatedness is effectively 0 (grey), indicating unrelated individuals. On the right, zoom-in to only consider the first 20 iPSC lines, highlighting the repeated samples.

3.6 Replication of iPSC eQTL using bulk RNA-seq

For comparison, I performed *cis*-eQTL mapping using the matched bulk RNA-seq data. I used the same pipeline (eq. (3.1))[9] and tested the same set of genes. This yielded 2,908 significant genes at an FDR of 10% (27% of genes tested). Such difference in number of discoveries can be explained, at least in part, by the reduced noise in the gene expression estimates when using bulk RNA-seq, partly due to the more consistent number of cells, and as a consequence reads, across individuals (**Fig. 3.4**).

In terms of agreement between the sets of results, I found that over 70% of eQTL identified using scRNA-seq data were replicated in the bulk study, where a single-cell eQTL lead variant (top variant per gene) was replicated if it achieved nominal significance (p value < 0.05) and had consistent direction of effect in the full set of results from the bulk eQTL analysis (**Fig. 3.7**). On the other hand, only around 50% of the eQTL identified (at FDR $<$ 10%) using bulk RNA-seq could be replicated in our single cell eQTL map. However, when we subsetted to eQTL identified using bulk data at a more stringent FDR threshold (1%), the replication proportion was much larger (76%), and the more stringent the FDR threshold, the more bulk eQTL we could replicate using single cell data (**Fig. 3.7**).

[9]The only difference of course is that there were not multiple replicates from the same line in the bulk RNA-seq data, but the model in eq. (3.1) still holds.

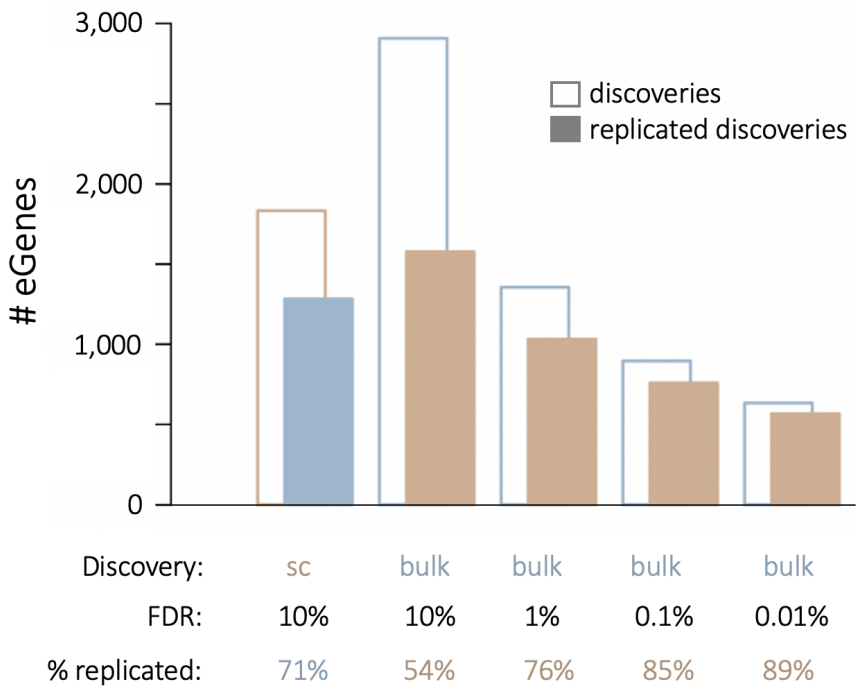


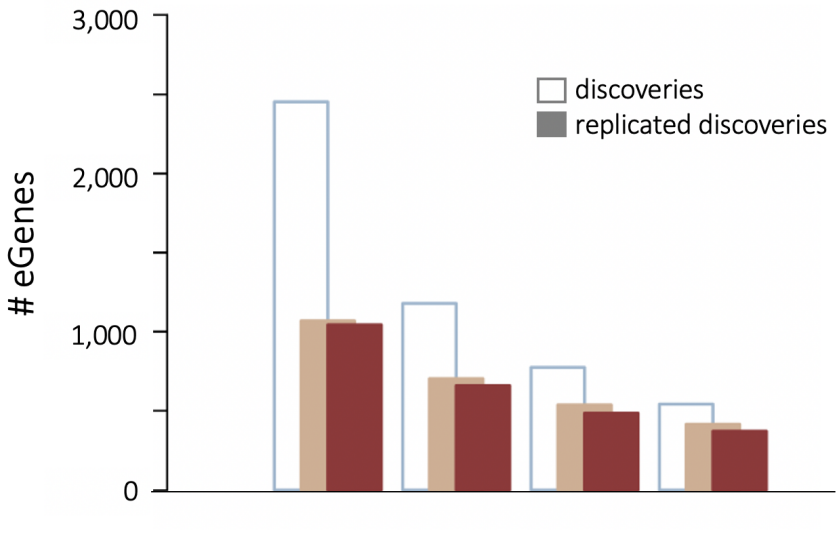
Fig. 3.7: Replication of iPSC bulk eQTL using single cell data and vice versa.
Replication of iPSC eQTL discovered with (matched-sample) bulk RNA-seq data using scRNA-seq data, and vice versa. The total number of genes with at least one eQTL (i.e. eGenes) discovered is shown, along with the number of discoveries replicated in the other dataset, at various FDR thresholds. FDR: false discovery rate; sc: single cell.

This result suggests that we are able to detect the stronger eQTL signals, but lack the statistical power (compared the corresponding test using bulk RNA-seq profiles) to identify smaller effects.

3.7 Replication of iPSC eQTL using 10X data

Next, to further confirm our iPSC eQTL map, we performed eQTL analysis (eq. (3.1)) using scRNA-seq data generated from a subset of 5 experiments (29 lines) using a droplet-based approach (10X Genomics [389], **Fig. 3.5**).

Similar to before, we assessed how many bulk-identified iPSC eQTL could be replicated using 10X samples. Since this study is fairly underpowered with only 29 samples, we did not consider the opposite analysis, i.e. 10X discoveries replicated in bulk. We did, however, compare results to an iPSC eQTL map using the SmartSeq2 data, when sub-setted to the same 5 experiments (and 29 lines, **Fig. 3.8**).



FDR:	10%	1%	0.1%	0.01%
% replicated with SmartSeq2 (N=29):	42%	60%	69%	77%
% replicated with 10X (N=29):	41%	56%	63%	68%

Fig. 3.8: Replication of iPSC bulk eQTL using single cell across technologies.
Replication of iPSC eQTL discovered with bulk RNA-seq (108 samples), using single cell RNA-seq data from a common set of 29 samples (SmartSeq2 in sand, 10X Genomics in red). The total number of bulk eGenes discovered is shown, along with the number of discoveries replicated using single cell profiles, at different FDR thresholds. As before, replication was defined as nominal significance, at p value < 0.05, and same direction of effect. FDR: false discovery rate; eGene: gene with at least one eQTL detected (at a given FDR threshold).

Overall, we observe that replication of bulk eQTL using scRNA-seq is reduced when we reduce sample size (for example, at FDR < 10% replication was 41% using 29 lines compared to 54% using all 111 lines in **Fig. 3.7**), but comparable across technologies (SmartSeq2, 10X Genomics), with SmartSeq2 slightly outperforming 10X (**Fig. 3.8**).

Once again, we can explain these differences in the number of discoveries at least partly as the result of differences in sequencing depth. If the variability between donors was the main difference between bulk and plate-based scRNA-seq (**Fig. 3.4**), here the key difference is the reduced number of reads obtained using the 10X technology, compared to SmartSeq2. Indeed, despite the higher number of cells (15,168 vs 2,275 for the same set of 29 lines), the total read count is significantly lower (median of 9 compared to 36 million reads per donor on average[10], **Fig. B.2**).

[10]For reference, the equivalent median reads per donor using bulk is 44.

Additionally, we found good agreement in terms of effect sizes between the eQTL maps obtained using the two different single cell technologies, highlighting the robustness of the approach (**Fig. 3.9**).

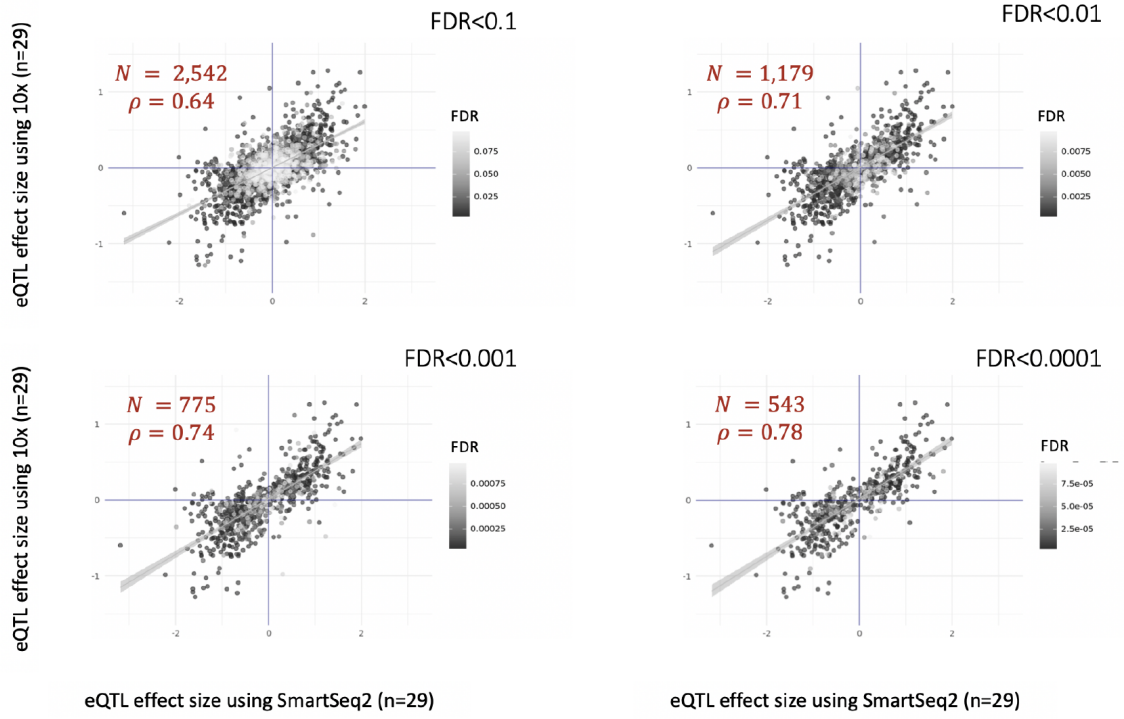


Fig. 3.9: Effect size agreement of bulk eQTL between single cell technologies.
Scatter plots of eQTL effect sizes obtained when testing association of iPSC eQTL discovered using bulk RNA-sequencing (108 cell lines), using SmartSeq2 ([383] x axis) and 10X Genomics ([389] y axis) on cells from 5 experimental batches (experiments 31, 40, 41, 43, 44; 29 cell lines in total). The number of eQTL examined and the correlation between effect sizes is indicated when we consider bulk iPSC eQTL discoveries at four different FDR thresholds (0.1, 0.01, 0.001 and 0.0001).

Since the vast majority of scRNA-seq datasets presented recently use droplet-based (rather than plate-based) technology, as it allows, as we have seen (**page 74**), the assessment of a much larger number of cells in a single experiment, it is important to show that this approach would work for such datasets as well. Whilst in this case we had data from too few individuals to make a very strong argument, the good concordance of results between 10X and SmartSeq2 suggests that this approach would work well for all single cell RNA-seq datasets, across technologies.

3.8 Preliminary steps towards a best-practice pipeline

The results presented so far are included in Cuomo *et al.* [447], the other results of which I discuss in detail in the next chapter (**Chapter 4**).

More recently, in collaboration with Marc Jan Bonder and Giordano Alvari from the Stegle group, I have worked on a best-practice pipeline which extends on the work presented so far, by testing the effect of different parameters of the model, to optimise yield of single cell eQTL mapping. In particular, using the same iPSC data described in this chapter so far (**Fig. 3.5**), we systematically compared results when mapping eQTL i) using various aggregation strategies to obtain 'pseudo-bulk' expression levels to use as phenotypes in the model, and ii) varying the type and number of 'global expression effect' covariates (see **section 2.2.4**) that are included in the model.

3.8.1 Overview of the iPSC data used

As I have mentioned, we broadly use the same iPSC data as before, i.e. scRNA-seq from [447] and (matched) bulk RNA-seq from the HipSci resource. However, we implemented some changes, to increase our confidence in this comparison. First, to make the data most comparable between the scRNA-seq data and the bulk RNA-seq data, we re-quantified single cell expression at the gene level using the 'featureCounts' tool [450], as was done for the bulk RNA-seq data (rather than relying on the quantification using the pseudo-aligner salmon [451], which was used to obtain the results described above and all results in **Chapter 4**). Additionally, to remove further possible confounding effects, a small group of lines from monogenic diabetes donors were excluded, as well as four lines which were slight outliers in the genotype space (**Fig. B.3**). In total, we map single cell eQTL for 88 cell lines (from 88 donors).

In addition to cells from 4-6 donors being multiplexed in 24 distinct pools (**Fig. 3.5**), cells from one pool were often sequenced in several runs ('sequencing run', or 'run'), which adds one layer of batch. As a cell-QC step, we calculated the average correlation of each cell with all other cells. Then, for each sequencing run, we calculated the median of the resulting cell-correlation values. If a run had median cell-correlation < 0.7, all cells from the run were discarded (**Fig. 3.10**, panel a). In a second step, cell-cell correlations were calculated again, between cells from the remaining runs only. This time, we considered line-run combinations, and discarded all combinations that had median cell-correlation < 0.5 (**Fig. 3.10**, panel b).

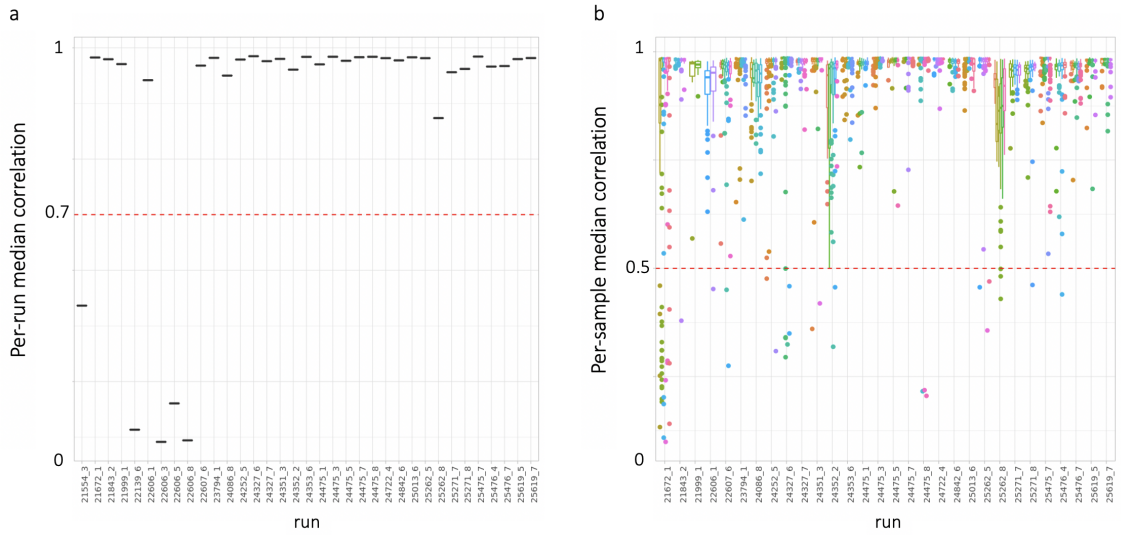


Fig. 3.10: Correlation-based cell QC
(a) Median cell-correlation calculated per run. Runs with median cell-correlation < 0.7 were discarded. (b) For the remaining runs, donor-run (i.e. sample) level median cell-correlations were calculated, and samples with median cell-correlation < 0.5 were discarded.

Moreover, in order to compare eQTL results across genes at different expression levels, we chose to be more lenient on the criteria for gene inclusion. Indeed, we considered all variable genes as obtained by ranking all genes (n=50,425[11]) by their squared coefficient of variation (CV^2)[12] across all cells, and selecting the upper two quartiles. As a result, 20,545 genes were included in the analysis. On the other hand, to reduce the multiple testing burden, we only tested SNPs with MAF >10% and within a smaller window around the gene (100kb on either side of the gene body).

Bulk RNA-seq data to assess replication

Finally, we compared the resulting single cell eQTL maps with results obtained using bulk RNA-seq both limited to the same set of samples (n=88, 'matched bulk', or 'm-bulk'), and using all samples that were available at the time (n=810 HipSci lines from 527 unique donors, 'all bulk', or 'a-bulk').

[11] This includes all genes annotated in the Ensembl reference [452], which include protein-coding genes, non-coding genes (e.g. transfer RNAs, ribosomal RNAs, long intergenic non-coding RNAs, etc.) and pseudogenes.
[12] $CV^2 = \sigma^2/\mu^2$.

3.8.2 Aggregation strategies

In order to use traditional bulk eQTL mapping methods for single-cell eQTL mapping, we first need to aggregate the multiple measurements from each donor to obtain bulk-like measurements. Here, we explore different aggregation methods (**Fig. 3.11, Table 3.1**). In particular, we consider the mean, the median, and the sum as aggregation strategies.

Initially, we performed aggregation at the donor ('d') level, i.e. taking all cells for a donor, to maximise the numbers of cells per donor. We call the resulting methods 'd-mean', 'd-median', and 'd-sum' (**Table 3.1**). Additionally, we consider aggregating not only at the donor level but also for each individual sequencing run (i.e. all cells from a given donor in a single sequencing run; designated 'dr', **Table 3.1**). While this approach better accounts for variation across technical batches, it also introduces multiple measurements from the same donor. We can account for these repeated measurements in our linear mixed model by including replicate and population structure information as covariates (eq. (3.1), **Fig. 3.6**). We call the corresponding methods 'dr-mean', 'dr-median', and 'dr-sum'.

The 'dr' aggregation is very similar to the approach used in the first part of this chapter, i.e. using the same principle of accounting for batches. The difference is that this is done at a deeper level of batch, noting that cells from the same experimental pool were sometimes sequenced in more than one run and thus some donors are present in multiple runs. Visually, the various aggregation methods show a similar picture across donors/samples and genes, with the median aggregations being most affected by the 0-inflated expression (**Fig. B.4, B.5**).

	aggregation method	normalisation	aggregation level
dr-mean	mean	single cell level (scran)	donor & run
dr-median	median	single cell level (scran)	donor & run
dr-sum	sum	pseudo-bulk level (TMM)	donor & run
d-mean	mean	single cell level (scran)	donor
d-median	median	single cell level (scran)	donor
d-sum	sum	pseudo-bulk level (TMM)	donor

Table 3.1: Types of aggregation methods tested.
Summary of the six key aggregation-normalisation strategies used in this study. In particular, for each approach we specify the aggregation method used, the type of normalisation adopted, and the level of aggregation selected. TMM: trimmed mean of M-values; normalisation method proposed in [453].

In all cases (i.e. using any of the aggregation methods) aggregated expression values were only calculated for samples (i.e. donors or donor-run combinations) with at least 5 cells.

3.8.3 Normalisation strategies

Importantly, normalisation of the scRNA-seq data was performed in different ways depending on the aggregation method used. For the mean and median aggregation (both at the donor and the donor-run level), we performed single cell-level normalisation using scran [409] implemented in scater [438], which is one of the standard methods used for single cell normalisation. The mean and the median were then calculated on the resulting normalised (logged) counts (**Fig. 3.11, Table 3.1**). As an alternative to scran, we also tested bayNorm [454], another recent single-cell normalisation approach. We found that the normalised counts are highly correlated (Pearson's R^2 mean: 0.88, median 0.93). On the other hand, summed count values (both dr-sum and d-sum) were obtained directly from the raw count data (i.e. non-normalised). Normalisation was then applied on the resulting pseudo-bulk counts, using methods typically used for bulk RNA-seq data. In particular, we perform trimmed-mean of M-values (TMM, [453]) normalisation on the aggregated counts, as implemented in edgeR [455], one of the best-established methods for bulk RNA-seq normalisation (**Fig. 3.11, Table 3.1**), followed by log transformation.

3.8.4 Phenotype transformation

Linear (mixed) models assume normality of the phenotype vector **y** (**Chapter 2**). However, when using gene expression, we are in the presence of count data. By log-transforming (log2) and normalising (previous section) such counts, the model-fit is improved, yet this approach still remains sub-optimal (see **section 2.3.4**).

In addition, two commonly-used phenotype transformations include standardising each phenotype vector (as I have done before) or quantile-normalising it, i.e. ranking the values and then making the data fit to a Gaussian distribution, forcibly. Here, we chose the latter strategy, which is more conservative, as it ensures a better fit to the (Gaussian) model. While common in the field (e.g. [151, 456]), this approach comes at the cost of some information loss on the real distribution of the expression vector, which can result in reduced power to identify eQTL. However, it is more robust to outliers, and thus better suited for this comparison. I note that it is important to be aware of the limitations and implications for the downstream analyses. For example, the additive genetic effect modelled in our LMMs (eq. (3.1)) is only strictly additive with respect to the sample analysed and the transformation that is performed on the phenotype vector. Consequently, it is important to confirm potential interesting effects using non-transformed expression values, as the raw data would be needed to estimate biologically meaningful effect sizes.

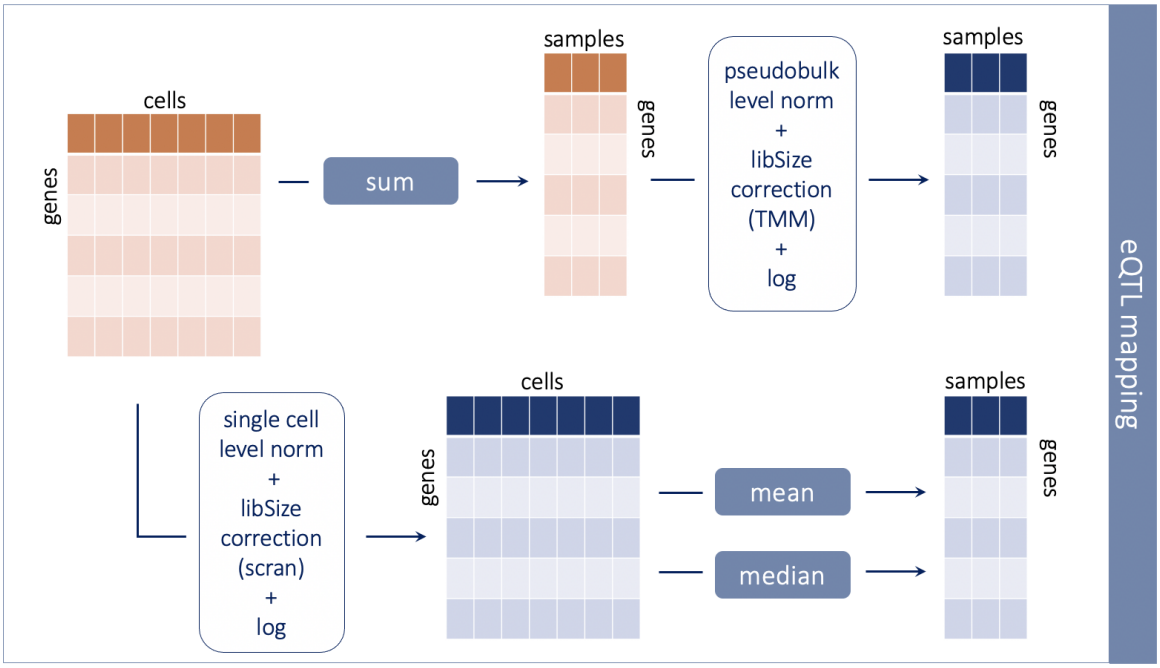


Fig. 3.11: sc-eQTL workflow.
Different approaches tested to perform eQTL mapping using scRNA-seq profiles. Starting from one gene × cell count matrix, counts were aggregated per sample (i.e. donor, or donor-run combination), either by summing the data first at sample-level and then normalising using methods designed for bulk RNA-seq (i.e. TMM implemented in edgeR [453, 455]) or by first normalising the single cell counts (using scran/scater [409, 438] or baynorm [454]) and then calculating the mean or the median at the sample-level.

3.8.5 Comparing sc-eQTL results across aggregation approaches

Next, the (normalised and 'gaussianised') aggregated expression values resulting from each of the methods described were used to map eQTL. For this comparison, the first 20 principal components were calculated from each of the aggregated matrices and included in the model as covariates (as before, from eq. (3.1), P = 20).

Table 3.2 summarises the results across the aggregation methods tested, when restricting the results to the set of 12,720 genes that were tested in all six eQTL maps. The comparison is first of all in terms of yield, i.e. number of eGenes identified using each of the methods (at FDR<5%; discovery). Next, we assess the degree of replication of the discoveries in the set of results obtained using bulk RNA-seq, both using matched samples only (m-bulk), and using all samples (a-bulk). Replication was assessed by considering significance (FDR<10%) and consistent direction of the effect of each sc-eQTL in the bulk results (**Table 3.2**).

	discovery		m-bulk replication		a-bulk replication	
	# eGenes	% tested	# replicated	% replicated	# replicated	% replicated
dr-mean	1,835	14.43%	889	48.45%	1,367	74.50%
dr-median	1,337	10.51%	650	48.62%	952	71.20%
dr-sum	1,463	11.50%	819	55.98%	1,153	78.81%
d-mean	1,305	10.26%	768	55.85%	1,046	80.15%
d-median	776	6.10%	470	60.57%	625	80.54%
d-sum	1,174	9.23%	709	60.39%	951	81.01%
m-bulk	2,590	20.36%	-	-	2,448	94.52%
a-bulk	9,729	76.49%	-	-	-	-

Table 3.2: Aggregation method comparison.
Number of eGenes and replication of eQTL for the different aggregation & normalisation strategies in Smart-Seq2 iPSC cells. The same set of 12,720 genes were considered in all of the strategies. FDR was controlled at 5% for the discovery; replication was defined as FDR<10% and consistent direction of effect in the two bulk studies, i.e. matched donor bulk (N=88, m-bulk) and all bulk set (N=527, a-bulk). Discoveries using bulk were also added, for reference (last two rows).

We identified between 776 and 1,835 genes with at least one eQTL (i.e. eGenes, at FDR<5%) using the different aggregation methods (out of 12,720 genes tested). To put these numbers in context, the equivalent eQTL map using matched samples with bulk RNA-seq (m-bulk) identified 2,590 eGenes (**Table 3.2**). These results show two main trends. First, aggregation at the donor-run level outperforms aggregation at the donor level only (e.g. dr-mean vs d-mean). Next, our results indicate that mean aggregation (after single cell-specific normalisation; 1,835 eGenes) outperforms sum aggregation (followed by bulk-like normalisation; 1,463 eGenes), and median aggregation performs worst in all cases (1,337 eGenes, **Table 3.2**).

Next, we used two selected sets of bulk iPSC RNA-seq data as described above, i.e. m-bulk (n=87) and a-bulk (n=526), to assess the replication of the iPSC sc-eQTL mapping results in bulk data (assumed to be the gold standard). We assess replication of the top eQTL effects in a single-cell method in bulk (i.e. direct eQTL replication), and define replication as FDR<10% (in the replication set) and a consistent effect direction. Replication rates from the two sets of samples show a very similar picture: on average we find slightly lower replication rates for the single cell normalisation methods, but a substantially higher total number of replicated discoveries at the eQTL level. In particular, the highest number of replicated eQTL is found for dr-mean (1,450 considering a-bulk) and highest fraction of replication is found for d-sum (82%, a-bulk). Moreover, we observe broadly consistent effect sizes between single cell and bulk, across the different methods, with the median once again performing worse than mean and sum aggregations (**Fig. B.6**). Finally, we see higher replication rates considering

a-bulk as compared to m-bulk, indicating that some of the effects found in the single-cell data can only be picked up from bulk datasets with more samples. When specifically looking at the effects that get replicated, we observe that they are highly overlapping (86%) across aggregation strategies. This result indicates that the same effects that get replicated in d-sum, are also replicated in dr-mean, but since there are more effects found in dr-mean, the overall fraction is lower.

We speculate that whilst the sum would perhaps be the most obvious approach to reproduce bulk-like measurements, the mean might perform better because of the normalisation used. Indeed, the normalisation at the cell-level may better balance the differences in read counts across cells prior to the aggregation at individual-level. Next, we also tested an alternative single cell normalisation approach, bayNorm [454], and mapped eQTL using the dr-mean aggregation method, finding that the results were virtually indistinguishable (1,835 eGenes vs 1,860, and Pearson's correlation between both p values and effect sizes: R=0.99, p value < 2.2×10^{-16}, total difference in effect sizes = 10^{-6}; **Fig. B.7**).

Overall, reassuringly, we broadly recapitulate the key results found in the first part of this chapter, confirming that eQTL maps using bulk RNA-seq are better powered that those using single cells, at identical sample size. Additionally, 'dr' approaches (i.e. aggregating at donor and batch) outperformed d approaches (i.e. aggregating only at donor level **Table 3.2**), re-iterating the importance of considering replicated experimental designs for eQTL studies.

3.8.6 Comparing results using different expression covariates

As a second comparison, we varied the type and number of expression covariates included in the model to account for global expression variation (see **section 2.2.4**). Since dr-mean outperformed other aggregation methods, we only perform this comparison using the dr-mean-aggregated values.

In particular, we compared multiple different methods to capture global expression covariates: probabilistic estimation of expression residuals (PEER [325, 326]), principal component analysis (PCA), linearly decoded variational autoencoder (LDVAE or linear scVI [330]), and multi-omic factor analysis (MOFA [331]), for which we considered two different flavours: with and without sparsity constraints). For each approach we tested the effect of including 5-25 factors as covariates in the model (eq. (3.1)), in steps of five. In a first instance, we compared results when running these maps for chromosome 2 genes only (1,421 genes).

To evaluate performance, as before, we first considered the number of eGenes (**Table 3.3**):

	0	5	10	15	20	25
PCA	83	160	165	184	175	167
PEER	83	148	139	146	126	183
MOFA	83	129	168	164	165	154
MOFA-ns	83	155	152	149	154	112
LDVAE	83	113	118	135	158	144

Table 3.3: Number of eGenes for various types and numbers of covariates.
Rows are methods to calculate expression covariates, columns number of covariates considered for the eQTL test. The numbers of eGenes are to be considered out of 1,421 (chromosome 2) genes tested. The dr-mean is used as aggregation method, normalised using scran. MOFA-ns is MOFA non-sparse, i.e. run without the default sparsity constraints.

Next, we considered the number of eQTL that were replicated using a-bulk (using all samples, FDR<10% and same direction of effect, **Table 3.4**; for the equivalent results using 'm-bulk' see **Table A.1**):

	0	5	10	15	20	25
PCA	64	116	124	132	133	123
PEER	64	106	102	109	95	136
MOFA	64	91	119	123	122	114
MOFA-ns	64	118	114	109	114	94
LDVAE	64	82	86	99	113	101

Table 3.4: Number of bulk-replicated eQTL for various types and numbers of covariates.
Similar to **Table 3.3** (dr-mean as aggregation method, 1,421 chromosome 2 genes tested only), but considering replication in the set of results using a-bulk (all samples, FDR<10% and same direction of effect). MOFA-ns is MOFA non-sparse, i.e. run without the default sparsity constraints.

As previously described [326], we observed a big increase in the number of eGenes discovered when considering covariates as compared to not considering covariates: the minimum increase is 75%. However, when comparing the method-specific optimal number of covariates (e.g. 15 for PCA, 25 for PEER), which is commonly done when optimising a eQTL mapping protocol, we observe broadly similar results across methods (**Table 3.3**). LDVAE, the only method included that works directly on the single-cell data, produces the smallest increase in eGene discovery. Furthermore, the replication of the effects in a-bulk, fixed at 20 PCs as previously used, is similar between the different methods (**Table 3.4**). Our results also show that more computationally expensive methods such as LDVAE, PEER or MOFA do not perform measurably better than the historical default in bulk eQTL studies of correcting for unwanted variation using principal components.

In summary, our recommendation for optimising yield of eQTL mapping using single cell expression profiles is to normalise counts at the cell-level, using single cell specific methods such as scran [436] (or baynorm [454], which performed similarly well), then aggregate such counts by considering the mean expression across cells. Depending on the experimental design, if data from a given donor is present across multiple technical batches, we recommend considering those batches as separate replicates for the donor, and calculating the mean for the two separately. Next, principal components can be calculated from such aggregated expression matrix, and the first 15-20 PCs should be included as covariates in the model. We also acknowledge that the number of PCs to include will depend on the sample size of the study. For example, in the case of several hundreds of individuals, 25 or 50 PCs may be more appropriate.

Future work towards establishing a best-practice pipeline for single cell eQTL studies includes confirming such recommendations using droplet-based data, for example using data I will introduce in **Chapter 5**, which have been sequenced using the 10X Genomics Chromium technology [389]. Moreover, future avenues involve the use of experimental-data informed simulation experiments to assess the impact of additional variables including varying numbers of cells per donor, varying sample size, batch effects and genetic effects of varying strength, on our ability (and statistical power) to detect eQTL using single cell expression data.

3.9 Discussion

Since being highlighted as 'Method of the Year' in 2013 [457], sequencing of the genetic material of single cells has become common-practice to investigate cell-to-cell heterogeneity in biological systems [458]. In particular, scRNA-seq enables the quantification of gene expression transcriptome-wide, and at single-cell resolution, allowing for cell type sub-populations to be distinguished [459–464], and the identification of cells transitioning between states [395, 465, 429, 430, 466].

Now an established method, scRNA-seq can be extended to profile single cell expression across several individuals, enabling the study of the effect of different genetic backgrounds on gene expression, at single cell resolution. Single cell eQTL (or as they called them scQTLs) were first introduced in a paper by Wills *et al.*, in 2013 [159]. In their paper, the authors study lymphoblastoid cell lines from only 15 donors, but could already observe some cell type-specific effect, which could not have been identified using bulk profiling. In 2018, Van

der Wijst *et al.* [160] published the second sc-eQTL map, this time across blood cell types from 45 individuals. They showed consistent direction of effects compared to bulk eQTL on similar cell types, but could only replicate a very small percentage of the bulk-identified eQTL ($\sim$10%).

The work I present in this chapter (and in the next, and published in [447]) was the third effort to map eQTL using scRNA-seq profiles, and the best powered at the time, with data from over one hundred individuals. Additionally, the use of bulk and scRNA-seq data from matched samples makes this the first step towards a systematic assessment of the differences between bulk and single-cell transcriptomics, as applied to eQTL mapping. As I have shown, compared to the results from [160], we could replicate approximately 50% of the eQTL found using bulk, and almost all of the strongest signals (**Fig. 3.7**).

Our preliminary results on a best-practice workflow for sc-eQTL studies suggest that mean is the preferable aggregation method, probably due to data normalisation considerations. Additionally, PCA slightly outperforms other linear matrix decomposition approaches for correcting for global expression covariates. This study begins to demonstrate that optimising the sc-eQTL mapping workflow can increase the number of eGenes discovered substantially. However, our conclusions come with some caveats; 1) simply discovering more eGenes does not necessarily mean that an approach is better, as false positives could arise due to data processing decisions; 2) bulk eQTL are a powerful, but not perfect, gold standard for assessing truth, as biases in bulk-eQTL mapping may be replicated in sc-eQTL mapping analyses. Future work includes validating these results when mapping eQTL using 10X data, where we expect normalisation to have an even larger impact, due to the increased sparsity of the data. Moreover, the use of real data-informed simulations will allow a more extensive power analysis, as their use will enable us to scale up the numbers of donors and cells and introduce group structure to comprehensively investigate the role of those parameters as well[13].

Yet, altogether, the results from this chapter illustrate a difference between bulk and single-cell eQTL mapping: there is a trade-off between statistical power and cellular resolution. Indeed, in this analysis of iPS cells, bulk RNA-seq data provided higher statistical power for discovery of eQTL (about 30% more discoveries using bulk). However, iPSCs are a rather homogeneous cell type, displaying relatively consistent expression profiles across

[13]Some of these analyses have now been done and strengthen our conclusions here, showing that our guidelines can be generalised to a 10X data set and to simulated data [467].

cells[14]. In more heterogeneous populations of cells, such as cells in the brain, the single cell transcriptomes may become critical for defining pure populations, thus increasing discovery power to detect eQTL.

A further advantage of the application of single-cell RNA-seq data in this study, was to enable the pooled experimental design. Indeed, this setup allows us to assay cells from many individuals in a single, neatly contained, experiment. As single-cell approaches are extended to more disease-relevant tissues and cell types, this may provide important clues on the causal role of genetic variants in disease.

These future studies are likely to be using droplet-based technologies, which allow the assessment of a much larger number of cells. Although our main results are on (plate-based) SmartSeq2 data, we could validate our approach with a subset of samples assayed with the droplet-based 10X Genomics technology (**Fig. 3.8**), which is a strong indication that single cell eQTL mapping can be performed using droplet-based scRNA-seq data. We observe that 10X data recapitulates bulk eQTL slightly less well than SmartSeq2 (**Fig. 3.8**). This can largely be explained by the lower number of reads obtained per individual using this technology, despite the higher number of cells (**Fig. B.2**). However, the differences are rather small (**Fig. 3.8**), which is reassuring, since droplet-based technologies are likely the only feasible option as we move to larger data sets in terms of both budget and throughput considerations.

Finally, in this chapter we have focused on reproducing standard 'mean' expression level eQTL mapping using scRNA-seq, where the phenotype of interest is expression abundance within a homogeneous population of cells. We can call such efforts 'pseudo-bulk' approaches, where we are essentially replicating bulk-like expression values and performing the eQTL test adapting approaches used for traditional eQTL mapping using bulk RNA-seq. In the applications we and others have described [160, 447], the value of using scRNA-seq lies in the fact that we are able, within a single experiment, to unbiasedly define and map eQTL in multiple different cell types, whilst retaining a single cell resolution.

Now that we have established that such 'mean-level' eQTL maps are feasible, new eQTL analyses, that specifically exploit the single cell resolution, can be performed (**Fig. 3.3**). One such analyses is variance eQTL (varQTL, vQTL [468]) mapping, where one can assess the

[14]That is not to say that there is no sub-structure at all, see for example our results from section 5.4. However, iPSCs are generally similar across protocols, and similar to ESCs [129].

effect of common genetic variants on cell-to-cell transcriptional variability, rather than on expression abundance. Unfortunately, these analyses have proven especially challenging, largely because the variance of gene expression is strongly dependent on its mean, making it hard to disentangle the two effects [469]. Moreover, variance QTL effects may be smaller than anticipated. As a result, studies at current sample sizes are underpowered to detect any variance QTL, as shown for instance by [444].

On the other hand, a single-cell approach allows detailed annotation of changing eQTL effects across heterogeneous cell types and cell states, with the ability to better interpret the context-specific role of individual genetic variants. In particular, dynamic eQTL, where the effect of a genetic variant on gene expression is modulated by differentiation time [470, 302] can be extended to single cell-resolved data, and expanded to include not only differentiation trajectories, but any cellular state. In the next chapter (**Chapter 4**), I will present examples of dynamic eQTL and eQTL affected by other cellular contexts.

4

Identifying dynamic eQTL effects during iPSC differentiation using scRNA-seq

As outlined in **section 1.2**, human iPSCs and cells derived from them have proven to be an excellent system to study cell fate decisions in early human development *in vitro*, which cannot be studied *in vivo*. So far, experiments have been limited to a handful of individuals or have focused on one single time point, thus the extent to which development varies from individual to individual, and the role played by common genetic variants during the process, remain largely unexplored. Here, we combine human iPS cell lines from over one hundred donors, a pooled experimental design, and single-cell RNA-sequencing to study population variation during differentiation to a definitive endoderm fate. We identify molecular markers that can predict the differentiation efficiency of iPSC lines, and exploit natural variation in the genetic background across individuals to map hundreds of eQTL that influence expression dynamically during differentiation and across cellular contexts.

Contributions

This work was a joint effort of the Stegle, Marioni and Vallier labs. In particular, the data was generated by Ludovic Vallier's lab at the Wellcome Trust Sanger Institute, and the experiments were led by Mariya Chhatriwala, who also contributed to the interpretation of the results. Additionally, from the Vallier lab, Iker Martinez performed some of the more recent experiments and Pedro Madrigal performed the ChIP-seq data analysis. All cell lines used are from the HipSci project. The statistical methods and analyses described in this chapter were co-supervised by Oliver Stegle and John Marioni. Davis McCarthy and I processed the scRNA-seq data and performed quality control steps. Davis McCarthy ran Cardelino to demultiplex donors from pooled experiments, and I analysed the results (for example I identified plate swaps based on these results). I performed exploratory data analysis of the single cell data, including the pseudotime analysis and the definition of developmental stages, i.e. the results presented in section 4.3. I developed and implemented the statistical methods for eQTL mapping at the individual developmental stages (section 4.4) and in bins along pseudotime (Fig. 4.13, 4.14). Daniel Seaton quantified allele-specific expression (ASE) and performed the ASE analyses, the results of which I then summarised and present in sections 4.5, 4.6. The analysis presented in section 4.7 was also largely driven by Daniel Seaton (in parallel with similar analyses presented in Chapter 5), with my contribution residing mostly in the definition of differentiation efficiency based on pseudotime, and in summarising the results and generating the figures for this section. The code for processing, analysing and plotting the data is open source and freely accessible here: https://github.com/single-cell-genetics/singlecell_endodiff_paper. Daniel Seaton, John Marioni, Oliver Stegle and I wrote the manuscript. The paper [447] is available at https://www.nature.com/articles/s41467-020-14457-z and has been published as:

Anna S.E. Cuomo*, Daniel D. Seaton*, Davis J. McCarthy*, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, Andrew Knights, Kedar Nath Natarajan, the Hipsci Consortium, Ludovic Vallier, John C. Marioni, Mariya Chhatriwala, Oliver Stegle. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, 2020, (* equal contribution)

I generated all figures presented in this chapter, except where indicated otherwise in figure legends.

4.1 Introduction

As highlighted in **section 1.2.2**, the early stages of embryogenesis entail dramatic and dynamic changes in cellular states. As cells transition from a pluripotent state, where they still have the potential to differentiate to all cell types, to committing to a specific cell fate, many molecular programs and mechanisms are activated and tightly regulated. Our understanding of such mechanisms in humans is still only partial, yet a lot has been learnt in model organisms including fruit flies [471–473], zebrafish [474–476], and mice [477, 478, 463]. For obvious (ethical) reasons, such mechanisms cannot be studied *in vivo* for humans. There exist a few studies that use human embryonic stem cells (hESCs) but the data is hard to access, often limited to a narrow time frame within development and cannot be easily derived from a variety of genetic backgrounds (**section 1.2**).

Additionally, the extent to which an embryo's genetic background influences early development has only been explored in a small number of special cases linked to rare large-effect variants that cause developmental disorders. This missing information is critical and it can provide a better understanding of how genetic heterogeneity is tolerated in normal development, when controlling the expression of key genes is vital.

Human iPSCs and iPSC-derived cells offer great potential to interrogate cell types and states that are challenging if not impossible to access in human, *in vivo* [294]. The generation of population-scale collections of human iPSCs [294, 296] has already allowed for assessing regulatory genetic variants in pluripotent [294, 296] as well as in differentiated cells [301, 300, 297]. Combined with a time-course experiment, iPSC-derived differentiation protocols can be used to mimic human early development *in vitro* in a highly controlled setup. This, in turn, provides a unique opportunity to study the dynamic effect of common genetic variants on gene expression regulation during early development across several genetically distinct individuals. However, iPSC differentiation protocols are challenging to apply in practice: most protocols generate much more cell type diversity than intended [479]. Additionally, extensive batch-to-batch as well as line-to-line heterogeneity has been observed [301, 296]. Finally, the protocols are lengthy and hard to scale, leading to limited throughput. In this study, we employed different strategies to combat some of these issues.

4.2 Single-cell RNA-seq profiling of differentiating hiPSCs

To assess the effect of common genetic variants on early human cell differentiation, human iPSC lines from 125 unrelated donors were differentiated towards a definitive endoderm fate [480]. In order to assess differences in expression and genetic regulation along development, data were collected at four time points: at iPSC stage (day0) and then after 24, 48 and 72 hours post initiation (day1, day2, day3). To get a comprehensive picture of the process, single cells' transcriptomes were assayed using full-length RNA sequencing (using the SmartSeq2 technology [383]). Additionally, the expression of two selected cell surface markers was recorded using fluorescence-activated cell sorting (FACS), a pluripotency marker (Tra-1-60) and a definitive endoderm marker (CXCR4). In order to increase throughput and control for technical batch effects, a pooled experimental design was adopted, where each differentiation experiment (hereon simply 'experiment') consisted of iPS cells from 4 to 6 distinct cell lines from the HipSci collection, which were grown and differentiated in the same plate. In total, data were retained from 28 experiments. Imputed genotypes (see **section 1.1.7**) were also available for all lines from the HipSci resource (**Fig. 4.1**).

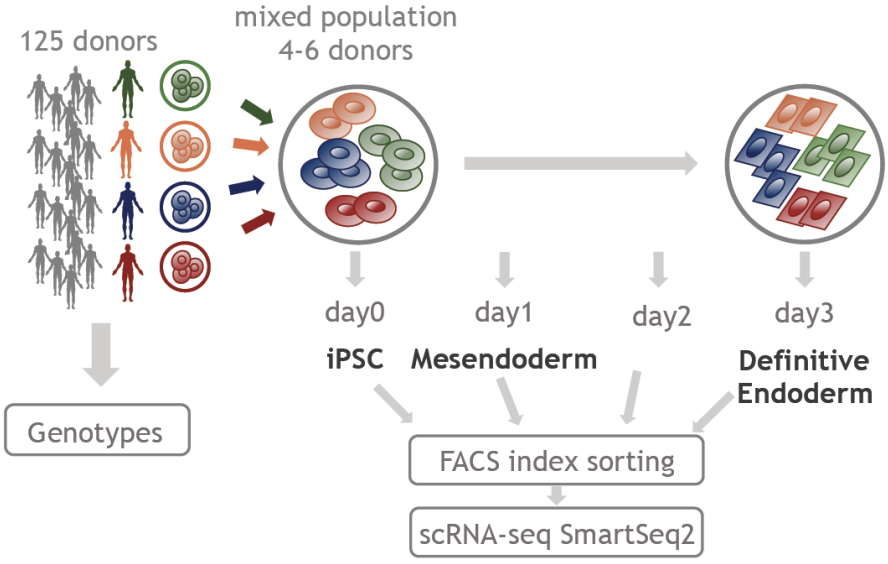


Fig. 4.1: Experimental Design.
Human iPSC lines from 125 unrelated donors were differentiated to definitive endoderm (**Fig. B.8**) using a pooled design, where cells from 4 to 6 lines were grown and differentiated together. Cells were collected prior to differentiation (at day0) and every 24 hours along differentiation to definitive endoderm (at day1, day2 and day3). Cells at day0 are expected to be pluripotent; cells at day1 are considered to be bipotent for either mesoderm or endoderm (mesendoderm); by day3, cells should have reached a definitive endoderm state. At each time point, cells were FAC-sorted and sequenced using the SmartSeq2 technology. Imputed genotypes were also available for all lines.

4.2.1 Data processing and QC

Demultiplexing donors from pooled experiments

In the considered pooled experimental design, cells from multiple donors are differentiated together in the same experiment. To be able to link the genetic background of an individual with their transcriptional profile we need to map the cells back to their donor of origin, without the use of any barcode. Indeed, we find that for the large majority of cells the RNA-seq reads map to a sufficient number of common genetic variants for us to reliably assign each cell to its original donor. In particular, assignment of cells to donors was performed using Cardelino [403]. In short, Cardelino estimates the posterior probability of a cell originating from a specific donor using common genetic variants in scRNA-seq reads, while employing a Bayesian beta binomial-based approach to account for technical factors such as differences in read depth, allelic drop-out, and sequencing accuracy. To perform donor assignment, we considered a larger set of HipSci lines with genotype information (n=490), including the 126 lines used in this study. A cell's assignment to a donor was considered successful if the model identified the match i) with posterior probability > 0.9, and ii) using a minimum of 10 informative variants. Cells for which the donor identification was not successful were discarded and not considered for further analyses. Across the entire dataset, 99% of cells that passed RNA QC steps (see below) were successfully assigned to a donor. In some cases, unexpected donor assignment (where several cells from one experiment were found to be assigned to none of the 4-6 donors used in that experiment) allowed me to identify (and correct) plate swaps that happened in the lab, without losing any data (**Fig. 4.2**).

Flow cytometry

The success of the differentiation protocol was validated using expression of two protein surface markers, a pluripotency marker, Tra-1-60, and a marker of definitive endoderm, CXCR4. We note that while cells were gated using the two markers, we did not discard any cells based on their expression. In contrast, the first cell QC step performed using FACS consisted in identifying dead cells based on 7AAD[2] using FACS staining. These were discarded and were not plated. FACS data were analysed using the openCyto package, implemented in R [481]. The FACS gating strategy we used is illustrated in **Fig. 4.3**.

[2]Staining with 7AAD is used a cell viability assay. 7AAD cannot readily pass through intact cell membranes, thus only cells with compromised membranes will stain.

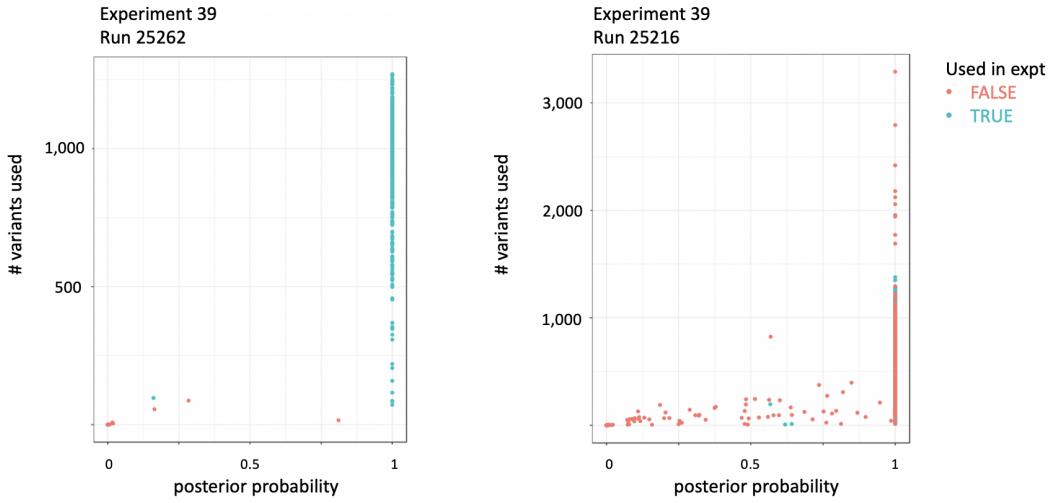


Fig. 4.2: Demultiplexing donors.
Example of how donor assignment of cells helped identifying a plate swap. To explore the results of the donor assignment algorithm [403], I plotted cells along two axes: on the x axis, the posterior probability of being assigned to a certain donor, on the y the number of common variants found on scRNA-seq reads used to perform the assignment. Because we know for each experiment which lines are supposed to have been differentiated, we can colour cells based on whether the donor they have been assigned to was used in the specific experiment or not. On the left, an example of a correct donor assignment: most cells are assigned to one of the correct donors[1] and the few that are not had very few usable genetic variants. On the right, the donor assignment is apparently incorrect. Most cells were assigned to donors that were not differentiated in the experiment, in many cases with a high level of confidence and using many variants, which would generally indicate high quality cells. Indeed, investigating further we realised that all cells were assigned to donors that all belonged to the same experiment, but that it was a different experiment. The wrong label was assigned in the lab: run 225216 actually contained cells from experiment 43 and not 39. By resolving this computationally, we avoided mistakes and retained all of the cells from this sequencing run, which would have otherwise been discarded.

scRNA-seq feature quantification and quality control

Single cell profiles were obtained using the SmartSeq2 technology [383]. This is a plate-based technology that involves single cells being sorted into 384 independent wells on a plate. Adaptors of raw scRNA-seq reads were trimmed using Trim Galore! [482–484], using default settings. Trimmed reads were mapped to the human genome (build 37) using STAR [485]. Gene-level expression quantification was performed using Salmon [451]. Briefly, Salmon quantifies transcript- (rather than gene-) level expression levels, similar to Kallisto [486]. Then, such values are summarised at a gene level (counts per million (reads) (CPM)).

[2] A cell technically could still have been assigned to a wrong donor within the correct experiment, but given a threshold both on the variants used (> 10) and on the posterior probability (> 0.9) I deemed this unlikely.

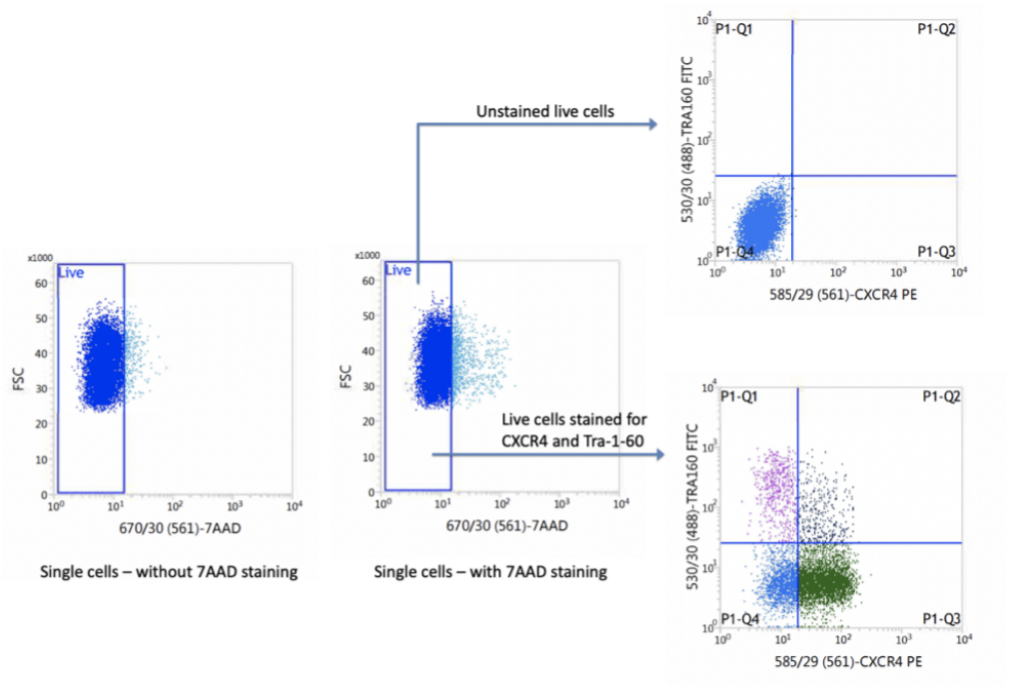


Fig. 4.3: FACS gating strategy.
Figure by Mariya Chhatriwala. FACS gating strategy: first, single cells were stained with 7AAD to exclude dead cells. Unstained live cells were then used to gate for expression of Tra-1-60 and CXCR4.

We performed quality control (QC) of scRNA-seq profiles following a widely used pipeline (see **section 3.1.2**) using Bioconductor packages *scater* and *scran*, implemented in R [436, 438, 439]. In particular, cells were retained for downstream analysis if they had at least 50,000 counts from endogenous genes, at least 5,000 genes with non-zero expression, if less than 90% of counts came from the top 100 most highly-expressed genes, less than 15% of reads mapped to mitochondrial (MT) genes, they had a Salmon mapping rate of at least 60%, based on distribution observation and thresholding (**Fig. 4.4**) [399]. Additionally, cells were only retained if they could be successfully assigned to a donor (QC1, **Fig. 4.5**).

I then performed an additional QC step, where I excluded all cells from plates and experiments that had overall low quality. In the case of plates sequenced twice, I retained the one with most cells. Finally, I retained plates that had enough cells for the majority of the donors considered (QC2, **Fig. 4.5**).

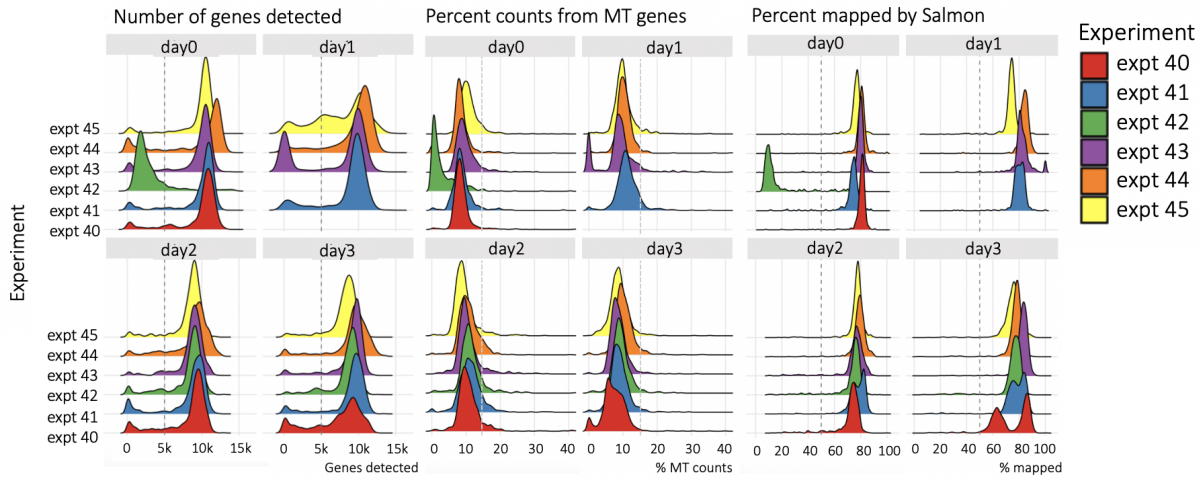


Fig. 4.4: Distributions of QC metrics.
Distributions of three exemplar QC metrics for six differentiation experiments (40-45). Shown are the cell distributions along the metrics (number of genes detected, percentage of counts from mitochondrial genes, Salmon [451] mapping rate), as well as the thresholds we used as dotted lines, stratified by day and experiment. One can immediately spot how poor quality plates[3]perform similarly badly across all metrics (i.e. < 5,000 genes detected, <60% reads mapped by Salmon).

scRNA-seq processing

SmartSeq2 data do not include UMIs, which can be used to accurately detect PCR duplicates and quantify transcript abundance [487, 488, 392]. In the absence of UMIs, we can borrow information from cells with similar total number of reads and correct for overall library size. Such size factor normalisation of counts was performed using *scater* [438]. Expressed genes with an HGNC symbol were retained for analysis, where expressed genes in each batch of samples were defined based on (i) raw count >100 in at least one cell prior to cell QC (i.e. **Fig. 4.5**) and (ii) average log2(CPM+1) >1 after cell QC. Normalised CPM data were log transformed (log2(CPM+1)) for all downstream analyses. As a last QC step, we considered possible differences between cell lines derived from healthy and diseased donors. Specifically, a subset of 11 cell lines in our dataset were derived from monogenic neonatal diabetes patients, and differentiated together with cell lines from healthy donors across 7 differentiation experiments (out of 28). There was no significant difference in differentiation efficiency (see **section 4.7**) between healthy and neonatal diabetes lines in these experiments (p value > 0.05), and cells from both sets of donors overlapped in principal component space (**Fig. B.9**). Thus, we included cells from all donors in our analyses, irrespective of disease state.

[3]e.g. cells from day0, experiment 42.

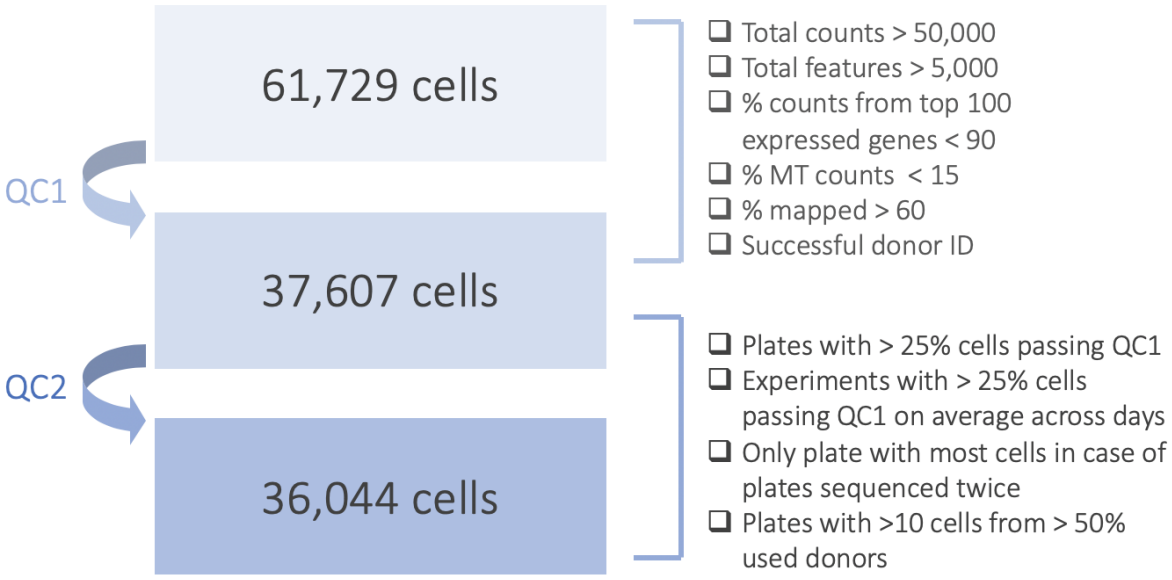


Fig. 4.5: QC workflow.
Two stages of cell QC. First, at the level of single cells, all QC metrics and thresholds are indicated. 61% of cells passed QC1. Second, at the experiment/plate level. If plates had many cells not passing QC1 they were considered poor quality batches and removed altogether. This stage removed far fewer cells, with 96% of cells considered passing QC2.

4.3 Data overview

Following quality control (QC), 36,044 cells were retained for downstream analysis, across which 11,231 genes were expressed. At each time point, cells from between 104 and 112 donors were captured, with each donor being represented by an average of 286 cells (after QC, **Fig. 4.6**). The success of the differentiation protocol was validated using canonical cell-surface marker expression: consistent with previous studies [489], an average of 72% of cells were TRA-1-60(+) in the undifferentiated state (day0) and an average of 49% of cells were CXCR4(+) three days post differentiation (day3, **Fig. 4.6**).

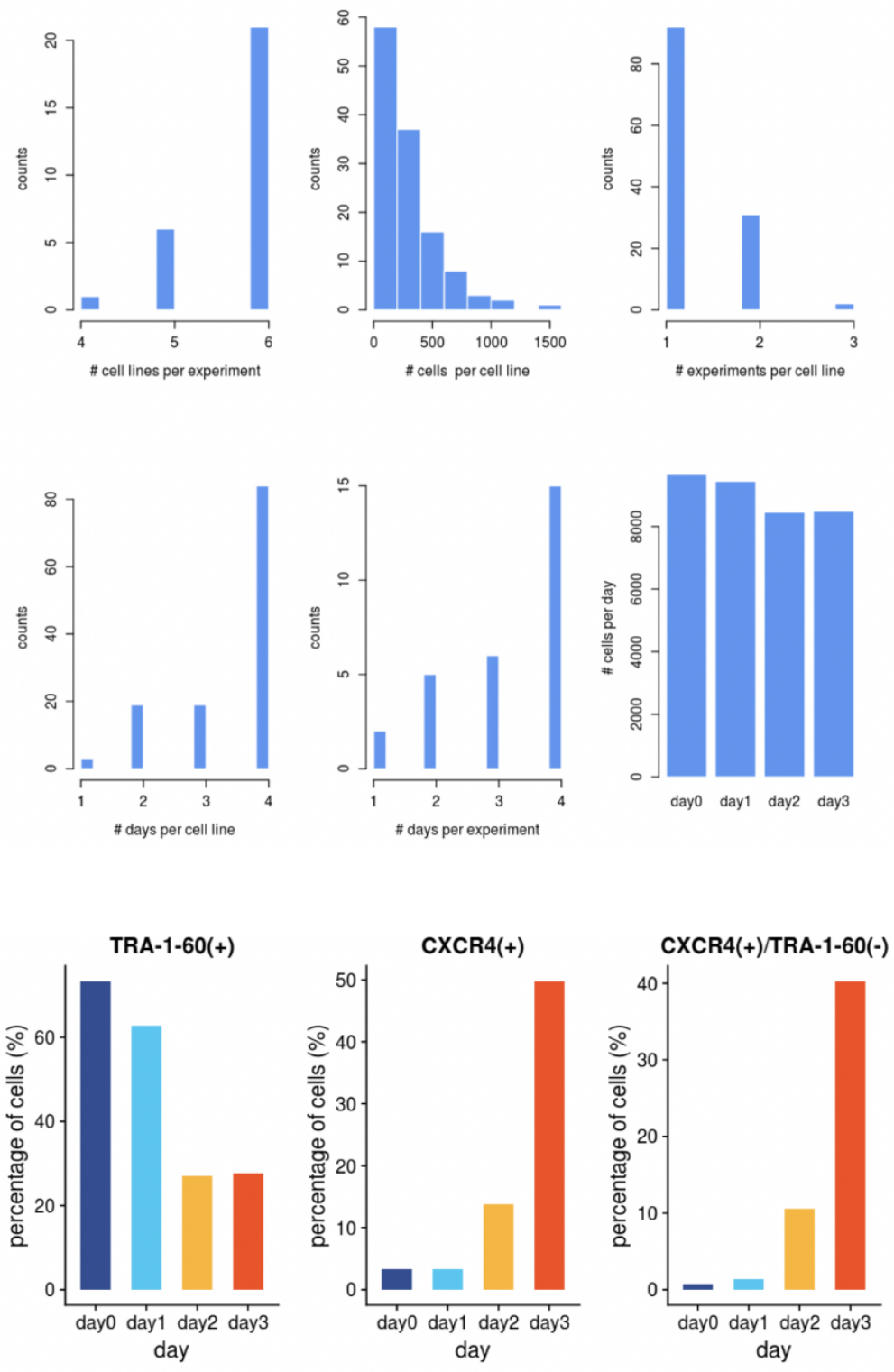


Fig. 4.6: Overview of experimental metrics.
Statistics for number of cells, donors, experiments, days, and combinations. Cell counts are shown after quality control. Additionally, shown are the percentages of cells that are positive for TRA-1-60, a pluripotency marker, positive for CXCR4, a definitive endoderm marker, and positive for CXCR4 and negative for TRA-1-60, across all cell lines and all experiments.

4.3.1 Sources of variation

To identify the main sources of variation in our dataset we performed variance component analysis for each of the genes, using a linear mixed model. Variance component analysis revealed the time point of collection as the main source of variation, followed by the cell line of origin and the experimental batch (**Fig. 4.7**).

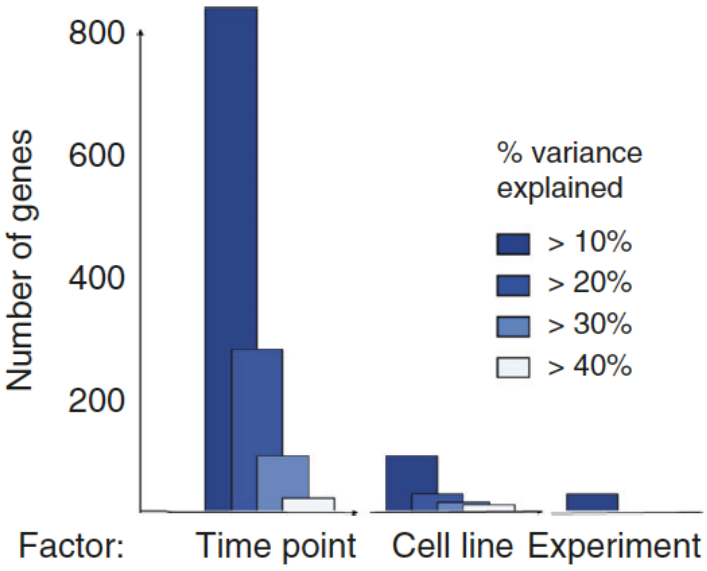


Fig. 4.7: Variance Component Analysis.
Summary of variance component analysis results for each of 4,546 highly variable genes, using a linear mixed model fit to individual genes to decompose expression variation into time point of collection, cell line and experimental batch. The number of genes for which each factor explains 10%, 20%, 30% and 40% of the variance respectively is indicated.

Next, we performed PCA on our dataset. To do so, we first identified the top 500 highly variable genes (HVGs) defined as the most variable genes given a mean-variance trend calculated across all genes, using the function *trendVar* as implemented in the R package scran. Consistent with the results from the variance component analysis (**Fig. 4.7**), the first principal component (PC1) was aligned with differentiation time, motivating its use to order cells by their differentiation status (hereafter 'pseudotime', **Fig. 4.8**).

Pseudotime inference is a common step in the analysis of scRNA-seq data along differentiation and development: while single cells are single snapshots along time, with enough points and considering that cells differentiate at different rates, they can be used to reconstruct a trajectory. In this case, the nature of the short and linear differentiation process of our data (i.e.

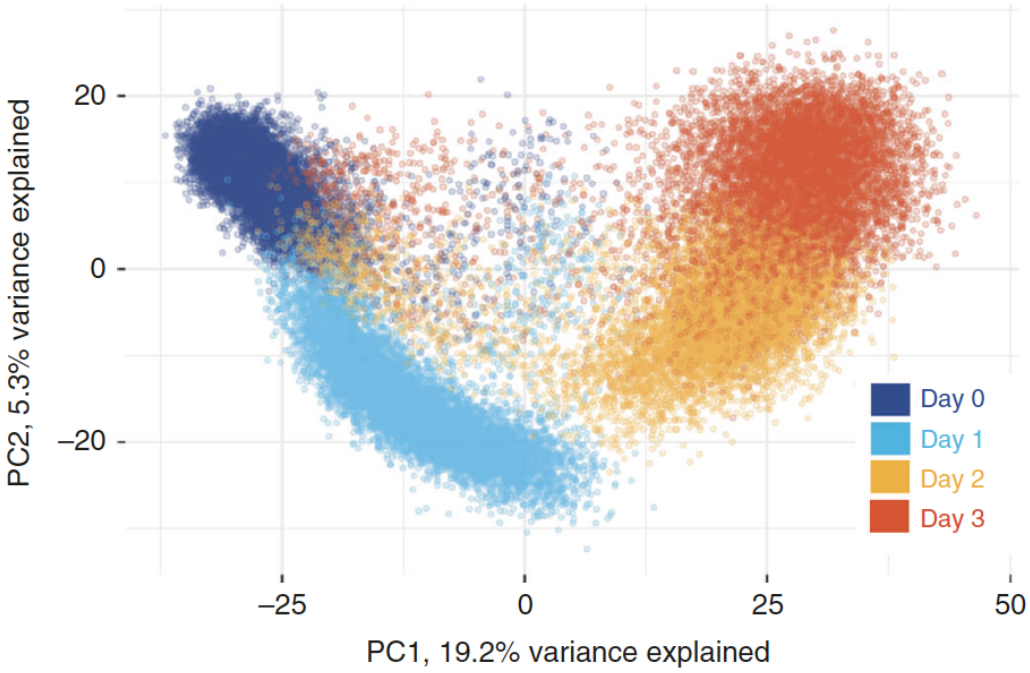


Fig. 4.8: Overview of dataset.
Principal component analysis of gene expression profiles for 36,044 QC-passing cells, coloured by the time point of collection. PC1 effectively captures differentiation time and is defined as pseudotime.

iPSC → mesendoderm → definitive endoderm) meant that PC1 captured the differentiation trajectory. For comparison, we did apply alternative pseudotime inference methods, which yielded similar orderings (**Fig. 4.9**). Further validation of our inferred pseudotime was provided by the temporal expression dynamics of known marker genes that characterise endoderm differentiation, which was captured by our ordering of cells as expected (**Fig. 4.10**).

4.3.2 Defining discrete developmental stages

While the continuous measure of pseudotime nicely highlights the dynamics of gene expression over time, in order to map eQTL, and to be able to exploit methods similar to those described in the previous chapter (**Chapter 3**), it was also important to define homogeneous populations of cells that represent specific developmental stages. To do so, we assign our cells to one of three non-overlapping stages, corresponding to the three canonical stages of endoderm differentiation: iPSC, mesendoderm (mesendo) and definitive endoderm (defendo). In particular, we utilise i) the ordering of cells along our inferred pseudotime ii) the expression of the previously described markers of differentiation progress and iii) the cell's day of collection to determine the cell assignment to each stage (**Fig. 4.10**). Specifically, we

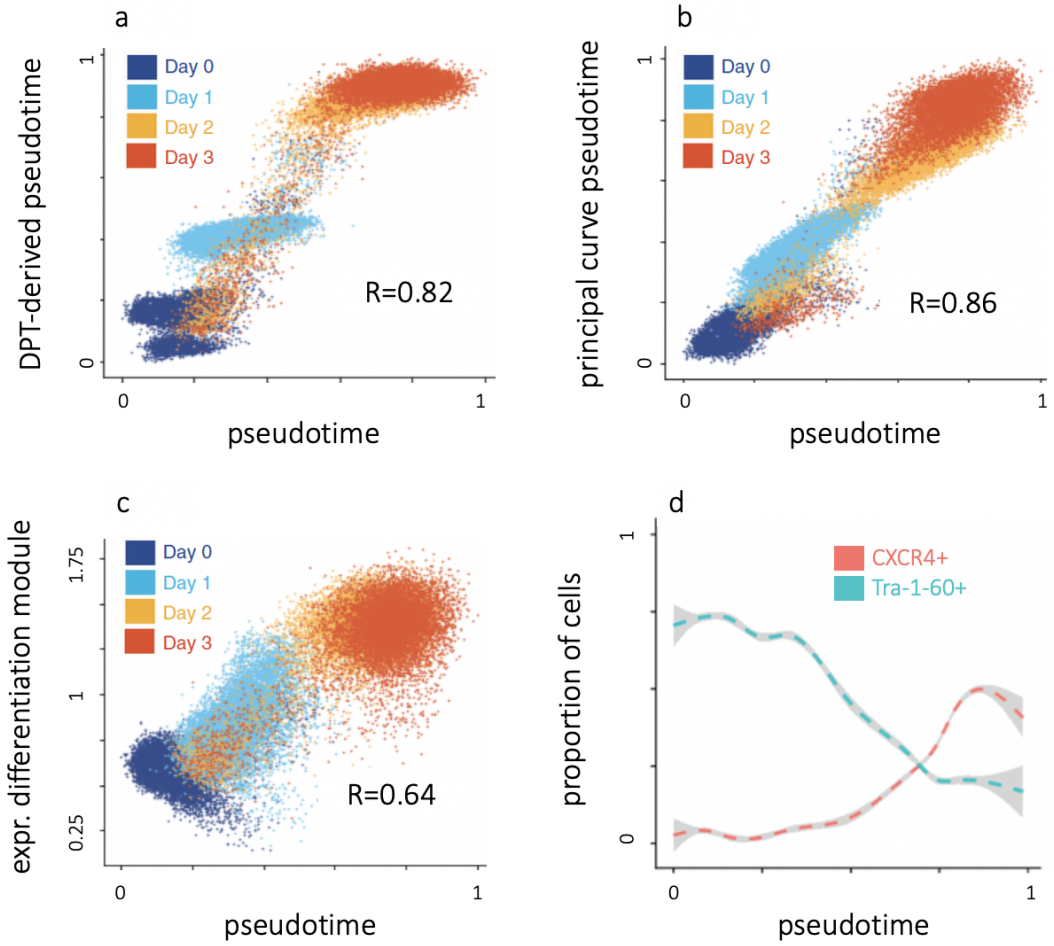


Fig. 4.9: Evaluation of pseudotime definition by comparison with alternative approaches. (a) Comparison of the pseudotime defined based on principal component analysis with diffusion pseudotime (DPT) [428]. The diffusion map was generated using 15 nearest neighbours and the first 20 PCs across the top 500 most highly variable genes. (b) Comparison of our defined pseudotime with an alternative measure of pseudotime based on projection of each cell on to a principal curve (using princurve as implemented in R [490]) calculated using the first two principal components from the top 500 most highly variable genes. (c) Comparison of our pseudotime to the average expression of 124 co-expressed genes associated with cell differentiation. (d) Scatter plot-derived loess curves of FACS markers as a function of the our PCA-based pseudotime, showing expected trends.

assign all day0 cells to the iPSC cluster given their very high homogeneity. Next, cells were assigned to the mesendoderm stage if they were collected at either day1 or day2, and had pseudotime values corresponding to the peak expression of *Brachyury* (*T*) along pseudotime (pseudotime between 0.15 and 0.5, **Fig. 4.10**). Similarly, cells were assigned to definitive endoderm if they were collected at day2 or day3 and had pseudotime values higher than 0.7, corresponding to a pseudotime window with maximal expression of *GATA6* (**Fig. 4.10**). In total, we assigned 28,971 cells (about ∼80% of all cells) to one of the three stages. A smaller fraction of cells with intermediate pseudotime (between 0.5 and 0.7, n=7,073) could not be confidently assigned to a canonical stage of differentiation; these cells were largely collected at day2, at which stage rapid changes in expression profiles are expected, reflecting a transitional population of cells. I note that these cells were excluded for the purposes of the initial stage eQTL mapping (results in **section 4.4**), but are included in all other analyses.

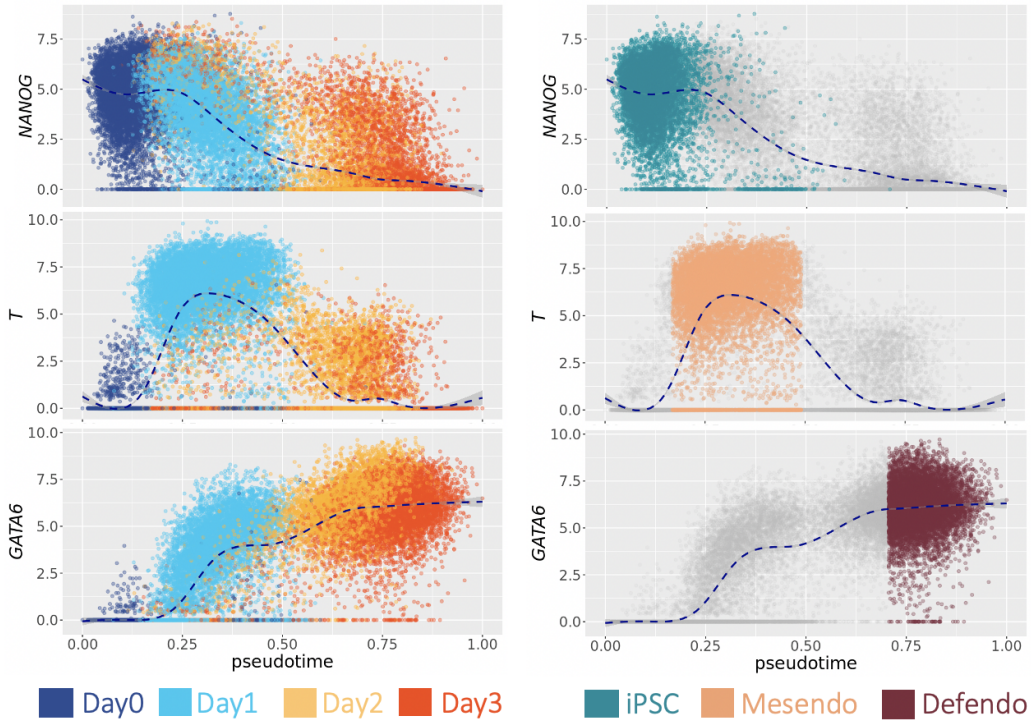


Fig. 4.10: Marker gene expression in pseudotime-based developmental states.
Expression of exemplar canonical markers for iPSC (*NANOG*), mesendoderm (*T*) and definitive endoderm (*GATA6*) along pseudotime. Developmental stages were defined taking into consideration i) the day of collection, ii) the expression of canonical markers, and iii) the position along pseudotime.

4.4 Mapping eQTL in iPSCs, mesendo and defendo

By combining single cell expression profiling and common genetic variation of over one hundred individuals we can begin to assess the impact of genetic variability on expression in a continuous manner across early human development. We have imputed genotypes for all of our 125 samples [294], so this study allows discovery of single cell eQTL along differentiation. Using the developmental stages just described and methods similar to those described in the previous chapter, we mapped eQTL in each of the iPSC, mesendo and defendo populations, yielding 1,833, 1,702 and 1,342 eGenes, respectively. Briefly, we quantified each gene's average expression level for each donor, experiment, and differentiation stage[4], before using a linear mixed model to test for *cis* eQTL, adapting approaches used for bulk RNA-seq profiles (+ and - 250 kb, MAF >5% [294]).

For comparison, we also performed eQTL mapping in cells collected on day1 and day3, i.e. the experimental time points commonly used to identify cells at mesendo and defendo stages [480]. Interestingly, this approach identified markedly fewer eGenes: 1,181 eGenes at day1, and 631 eGenes at day3. These results demonstrate the power of using the single-cell RNA-seq profiles to define relatively homogeneous differentiation stages in a data-driven manner (**Fig. 4.11**). Notably, this observation was not merely a consequence of differences in the number of cells or donors considered in each cell population (**Fig. 4.11**).

Profiling multiple stages of endoderm differentiation allowed us to assess at which stage along this process individual eQTL can be detected as well as the level of sharing of genetic signal across time. We observed substantial regulatory and transcriptional remodelling upon endoderm differentiation of iPSCs, with over 30% of eQTL being specific to a single stage. To define pairwise replication (and conversely specificity) between two sets of test results we considered nominal significance (p value < 0.05) and consistent direction of the effect size. Importantly, we observed that stage-specificity of eQTL was not significantly explained by stage-specific gene expression (**Fig. 4.12**). Our differentiation time course covers developmental stages that have never before been accessible to genetic analyses of molecular traits and thus this study provides the first eQTL maps at mesendoderm and definitive endoderm. We next explored whether any of the eQTL identified in these two studies were novel, and found that 349 of them have not been reported in either a recent iPSC eQTL study based on bulk RNA-seq [449], or in a compendium of eQTL identified from 49 tissues as part of the

[4]This approach is the same as what is described in the first part of **Chapter 3**, and similar to the 'dr-mean' described in **section 3.8**, except that the aggregation is done at the experiment level rather than the sequencing run level.

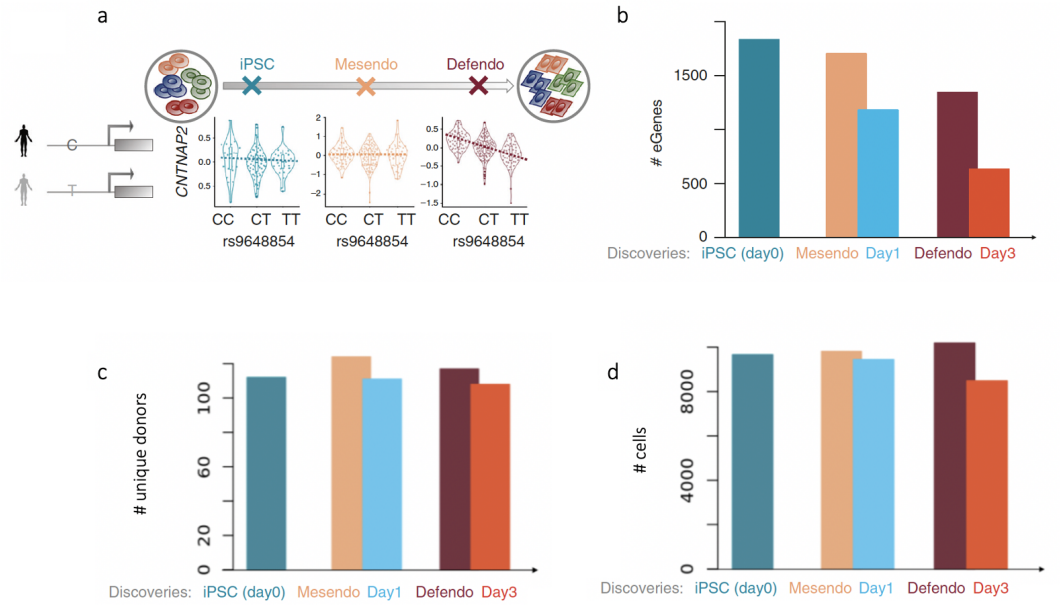


Fig. 4.11: Mapping single cell eQTL at different developmental stages.
(a) Illustration of the single cell eQTL mapping strategy at various stages of differentiation. Shown is an example of a defendo-specific eQTL. Box plots of gene expression stratified by the allelic state of rs9648854 at each stage, showing an association between rs9648854 and *CNTNAP2* expression at the defendo stage, but not at earlier stages. (b) Comparison of eQTL mapping using different strata of all cells. The use of pseudotime-based stages increases the number of detectable eQTL, compared to using the corresponding time point of collection. Bar plots represent number of eGenes (genes with at least one eQTL, at FDR < 10%). (c) Similar to b, the number of donors for which gene expression data were assayed at day0, day1, and day3, compared to the number of donors in the pseudotime-inferred mesendo and defendo stages. (d) As for (c), with the number of cells. https://github.com/ebiwd/EBI-Icon-fonts by EBI Web Development is licensed under CC BY 4.0.

GTEx project [150]. An eQTL was defined as novel when it was not reported as lead variant (FDR < 10%) in any of the tissues considered nor was it in LD (see **section 1.1.7**) with any reported lead variant, LD assessed using $r^2 < 0.2$.

Finally, we investigated the presence of lead switching events. These correspond to two distinct genetic variants that are identified as lead eQTL for the same gene at different stages of differentiation (at LD: $r^2 < 0.2$), We found lead switching events for 155 eGenes (an example in iPSC and defendo is illustrated in **Fig. 4.12**). To explore the potential regulatory role of these variants, we investigated whether the corresponding genetic loci also featured changes in histone modifications during differentiation. To do so, we used ChIP-Sequencing to profile five histone modifications that are associated with promoter and enhancer usage (H3K27ac, H3K4me1, H3K4me3, H3K27me3, and H3K36me3) in human embryonic stem cells (hESCs) that were differentiated towards endoderm (using the same protocol employed

above) and measured at equivalent time points (i.e. day0, day1, day2, day3, see **section C.1.3** for detailed experimental methods). Interestingly, we observed corresponding changes in the epigenetic landscape for 20 of the lead switching events (i.e. stage-specific lead variants overlapped with stage-specific changes in histone modification status), suggesting a direct mechanism (**Fig. 4.12**).

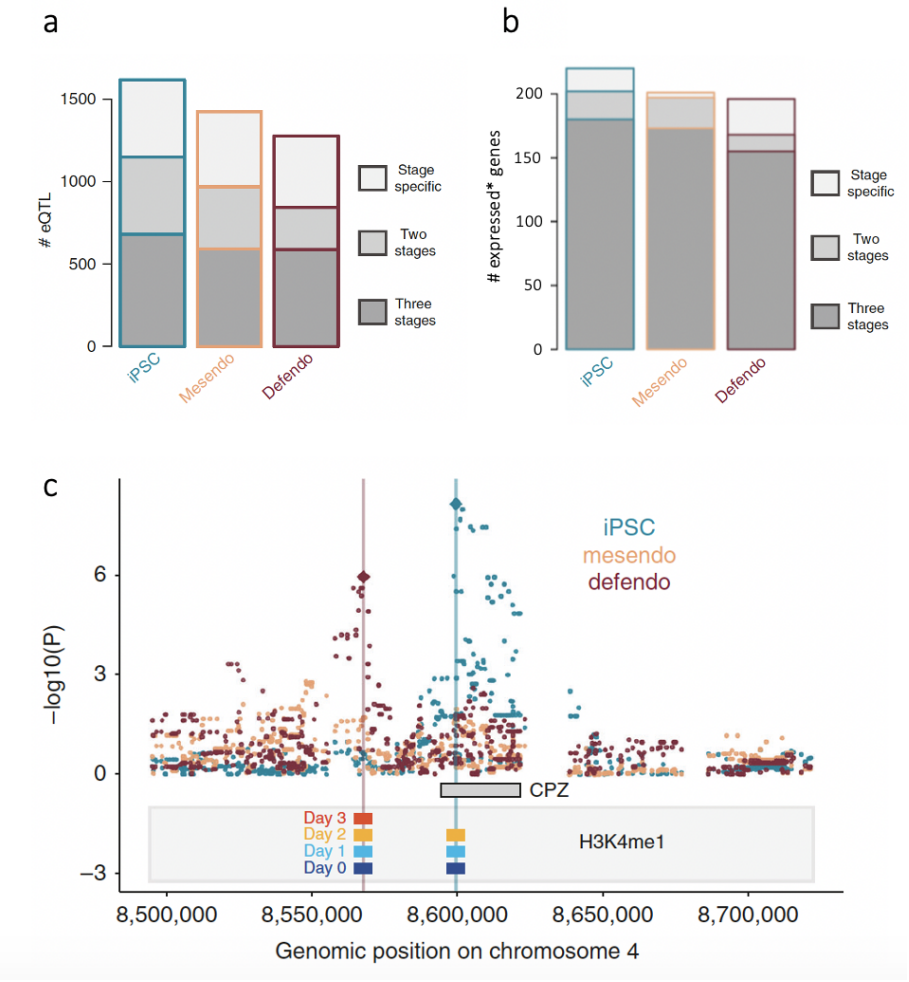


Fig. 4.12: Stage-specific eQTL.
(a) Proportion of eQTL that are specific to a single stage, shared across two stages, or observed across all stages (sharing defined as a lead eQTL variant at one stage with nominal p value < 0.05 and consistent direction at another stage). (b) Proportion of stage-specific eGenes (genes with a stage-specific eQTL) that are expressed only at a single stage, expressed at two stages, or expressed at all stages. Expressed is defined as normalised log2(CPM+1) > 2. CPM: counts per million. (c) A lead switching event consistent with epigenetic remodelling. The overlap of H3K4me1 with the eQTL SNPs across differentiation time points is shown by the coloured bars.

4.5 Dynamic eQTL across iPSC differentiation

The availability of large numbers of cells per donor across a continuous differentiation trajectory from pluripotent stage to definitive endoderm enabled the analysis of dynamic changes of eQTL strength at fine-grained resolution. To formally test for eQTL effects that change dynamically across differentiation (dynamic eQTL), we tested for associations between pseudotime (both linear and quadratic) and the genetic effect size using allele-specific expression (ASE) and a linear model (see **Box 3**):

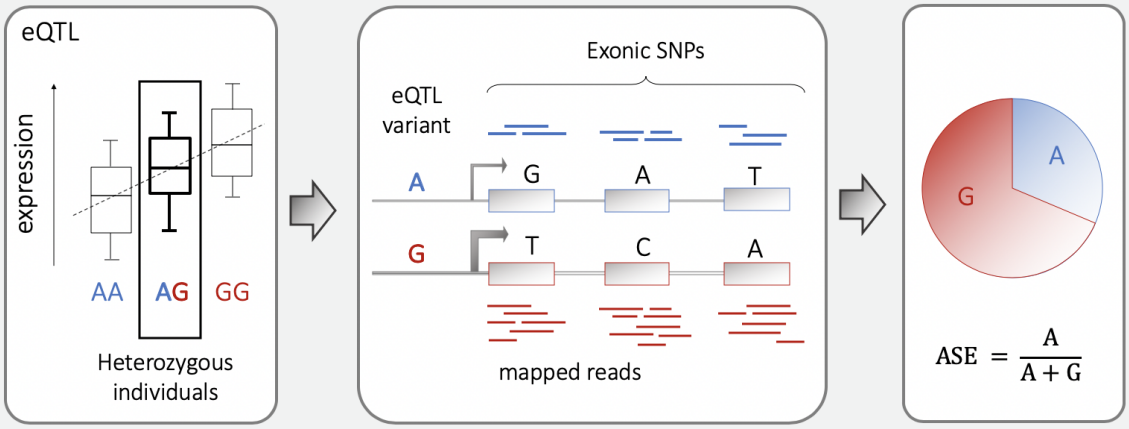
$$\text{ASE} = \alpha_1 \text{pseudotime} + \alpha_2 \text{pseudotime}^2 + \boldsymbol{\psi}, \tag{4.1}$$

where the genetic effect is defined based on ASE at the level of single cells, i.e. quantified as fractional read counts overlapping each allele for a given gene-SNP pair (see details in **Box 3**). We assessed significance using a likelihood ratio test with two degrees of freedoom (i.e. $H_0 : \alpha_1 = \alpha_2 = 0$). For this analysis, we focused on the joint set of 4,422 eQTL lead variants (4,470 SNP-gene pairs) discovered at the iPSC, mesendo, and defendo stages and explored how they were modulated by developmental time (using our inferred pseudotime). In this way, we uncovered a total of 899 time dynamic eQTL at FDR < 10%, including a substantial fraction of eQTL that were not identified as stage-specific by the discrete approach previously used (**page 113**). This analysis is somewhat complementary to the eQTL map perfomed on discrete differentiation stages (**Fig. 4.11**), which identified substantial stage-specific effects (**Fig. 4.12**). Namely, we observe that in general stage-specific effects are weaker and unique to certain cell types. In contrast, the dynamic eQTL identified are detected to a certain extent across all cell types, but the strength of the effect is modulated by differentiation time.

One obvious explanation for these subtle dynamic changes could be that they are simply reflecting changes in overall expression. To visualise this, we used a sliding-window approach. Because we need a rather large amount of cells to reliably estimate expression abundance for each individual, we slide a window containing 25% of the cells along pseudotime by a step of 2.5% cells. In each window, we considered average expression quantifications and estimate genetic effects using eQTL mapping, essentially performing the same analysis we performed in developmental stages in **section 4.4**, now in each window. In parallel, we reassessed each eQTL in each window taking advantage of the full length transcript sequencing to measure ASE. Here, in each window, we quantified the deviation from 0.5 of the expression of the minor allele at the eQTL (ratio of reads phased to eQTL variants, **Fig. 4.13**). Notably, ASE can be quantified in each cell and is independent of expression level, thus mitigating technical correlations between differentiation stage and genetic effect estimates (**Box 3**).

Box3: Quantifying genetic effects using ASE

When full-transcript (phased) data is available, ASE can be used to quantify the genetic effect of a variant on expression as a single ratio, by quantifying the relative expression of one allele over the other. For a given eQTL - and the corresponding eQTL variant and eGene - we i) select individuals that are heterozygous at the eQTL SNP of interest, ii) consider all exonic heterozygous variants on the corresponding eGene, iii) map reads to these SNPs, iv) aggregate all reads coming from the same chromosome and v) compute the ratio. Conventionally, we look at ratios < 0.5 i.e. in the numerator goes the allele with fewer reads mapped to it:



If one of these alleles is more responsive to a particular environmental factor (e.g. because of preferential transcription factor binding), then ASE is expected to vary consistently with that factor. This observation has previously been used to identify GxE interactions in gene expression across individuals [491]. Critically, these ASE tests are internally matched, because potentially confounding batch effects and technical variation should affect both alleles in each cell similarly. Additionally, this test increases power by reducing the number of parameters to estimate, i.e. instead of the standard test for interactions:

$$\text{expression} = \sum_{k=1}^{K} \mathbf{e}_k \alpha_k + \mathbf{g}\beta + \sum_{k=1}^{K} \mathbf{g} \odot \mathbf{e}_k \gamma_k + \boldsymbol{\psi},$$

where for each of K environments $\mathbf{e}_k$ two terms must be added (to account for E and GxE), resulting in (2*K + 2) parameters needing to be estimated (including one effect size β and the variance explained by the noise term σ_n^2, from $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I_n})$), we can run:

$$\text{ASE} = \sum_{k=1}^{K} \mathbf{e}_k \alpha_k + \boldsymbol{\psi},$$

where we can test directly the effect of the K environments on ASE and therefore only K+1 parameters need estimation (the K α_k's and σ_n^2).

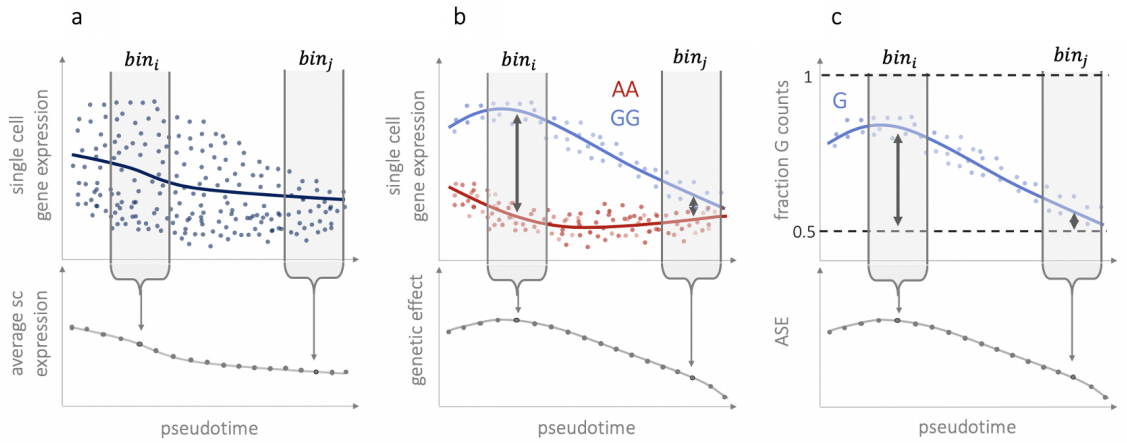


Fig. 4.13: Schematic of the sliding window approach.
Cells are binned based on pseudotime ordering, to (a) quantify average expression, (b) perform eQTL mapping, and (c) quantify average ASE. Each bin includes 25% of cells, binned at incremental steps of 2.5%.

Both methods result in a measure of genetic effect dynamics, i.e. changing strength of genetic effects along differentiation. Reassuringly, the two approaches were highly consistent across pseudotime (**Fig. 4.14**). To explore this hypothesis, we clustered the top dynamic eQTL (FDR <1%) jointly based on both the relative gene expression dynamics (global expression changes along pseudotime, quantified in sliding windows as above), and on the genetic effect dynamics (using ASE). This identified four basic dynamic patterns (**Fig. 4.14**): decreasing early (cluster A), decreasing late (cluster B), transiently increasing (cluster C), and increasing (cluster D). As expected, stage-specific eQTL were grouped together in particular clusters (e.g. defendo specific eQTL in cluster D, **Fig. 4.15**). Notably, the dynamic profiles of gene expression and those of eQTL effects tended to be distinct, demonstrating that expression level is not the primary mechanism controlling variation in genetic effects. In particular, genetic effects were not most pronounced when gene expression was high (**Fig. 4.14**). Distinct combinations of expression and eQTL dynamics result in different patterns of allelic expression over time. This is illustrated by the mesendoderm-specific eQTL for *VAT1L*. Overall expression of *VAT1L* decreases during differentiation, but expression of the alternative allele is repressed more quickly than that of the reference allele (**Fig. 4.14**). This illustrates how *cis* regulatory sequence variation can modulates the timing of expression changes in response to differentiation, similar to observations previously made in *C. elegans* using recombinant inbred lines [470]. In other cases, the genetic effect coincides with high or low expression, for example in the cases of *THUMPD1* and *PHC2* (**Fig. 4.14**). These examples illustrate how common genetic variation is closely linked to the dynamics of gene regulation.

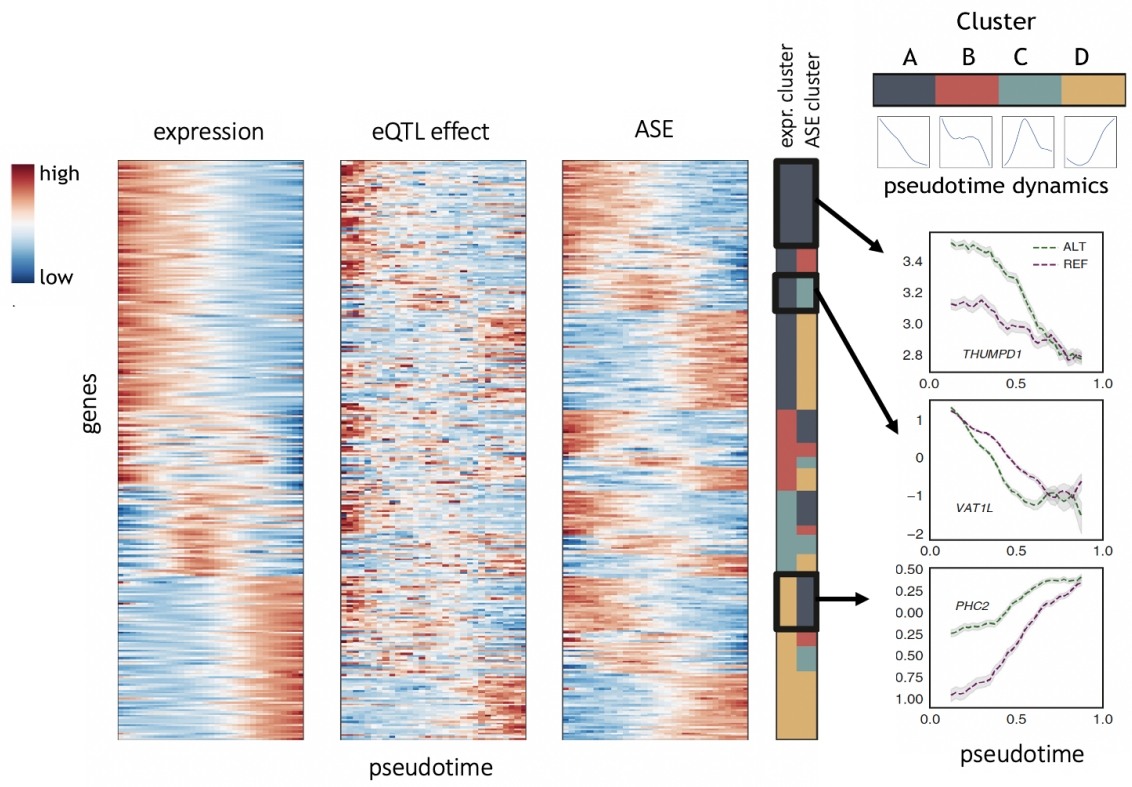


Fig. 4.14: Dynamic eQTL.
Clustered heatmap of global expression levels, eQTL effect sizes, and ASE across pseudotime for the top 311 genes with the strongest dynamic eQTL effects (FDR < 1%; out of 785 at FDR < 10%). For each gene, the dynamic profiles of gene expression and ASE were jointly grouped using clustering analysis, with 4 clusters. The membership of a genes's expression and ASE dynamics to one of these clusters is shown by colours in the right-hand panel. All values in the heatmaps are z-scores, normalised by gene (row). In particular, for ASE, average ASE values are plotted such that red indicates highest deviation from 0.5. The diagram in the top right summarises the four identified cluster dynamics, displaying the average dynamic profile of each cluster, computed as the average across z-score normalised gene expression/ASE profiles. Selected examples of the dynamics of allele expression for different cluster-combinations are shown in the bottom right panel. Shaded regions indicate standard error (+/- 1 SEM). This figure is based on one by Daniel Seaton.

We next asked whether dynamic eQTL were located in specific regulatory regions. To do this, we evaluated the overlap of our identified dynamic eQTL with epigenetic marks defined using the hESC differentiation time series[5] (**Fig. 4.15**). This revealed an enrichment of dynamic eQTL in enhancer (i.e. H3K27ac and H3K4me1), and promoter (H3K4me3) marks as compared to non-dynamic eQTL (i.e. eQTL we identified which did not display dynamic changes along pseudotime, **Fig. 4.15**), consistent with these SNPs being located in active regulatory elements.

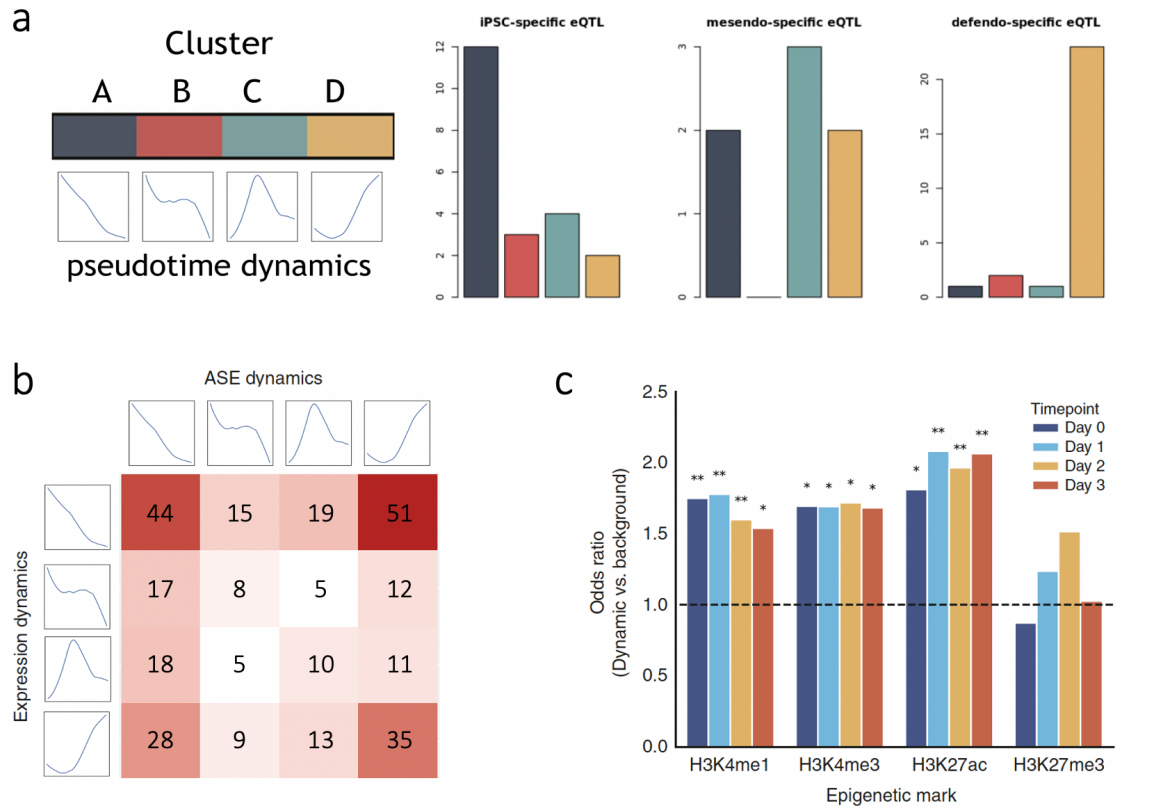


Fig. 4.15: Characterisation of dynamic eQTL.
(a) Summary of the identified cluster dynamics[6]and assignment of stage-specific eQTLs to dynamic eQTL clusters. The numbers of each of the 3 classes of stage-specific eQTL (i.e. iPSC-, mesendo- , and defendo-specific eQTLs) that are assigned to each of the 4 dynamic eQTL clusters. (b) Number of genes categorised by the combination of expression and ASE cluster from (a). Average dynamics of expression clusters (rows) and ASE clusters (columns) are shown. (c) Overlap of dynamic eQTL variants from a with histone marks. The odds ratio compared to the background of all other eQTL variants is shown (*p value < 0.01; **p value < $1x10^{-4}$; Fisher's exact test).

[5]using bulk, at equivalent time points along endoderm differentiation, see **section C.1.3** for methods.
[6]displaying the average dynamic profile of each cluster, computed as the average values across z-score normalised gene expression/ASE profiles.

4.6 Cellular environment modulates eQTL effects

Whilst differentiation was the main source of variation in the dataset, single cell RNA-seq profiles can be used to characterise cell-to-cell variation across a much wider range of cell state dimensions [465, 327, 492]. Next, we identified sets of genes that varied in a co-regulated manner (co-expressed genes) using clustering. In particular, grouping of genes by single-cell co-expression was performed using affinity propagation[7] [493], as implemented by the Python scikit-learn library [494], using the top 8,000 highest expressed genes. This resulted in a set of 60 clusters of co-expressed genes. Exemplar co-expression clusters were selected to represent 4 dimensions of cellular state: cell cycle G1/S transition (cluster 10), cell cycle G2/M transition (cluster 30), cellular respiration (cluster 0), and sterol biosynthesis (cluster 28). This selection was done according to two criteria: (i) strongest enrichment of relevant GO terms[8], and (ii) a priori expectation of sources of cell-to-cell variation. Indeed, variation in cell cycle stage is a common feature of single-cell datasets [465], while variation in metabolic state during iPSC differentiation is well known [496]. These functional annotations were further supported by enrichment of relevant transcription factor binding (e.g. enrichment of *SMAD3* and *E2F7* targets in the differentiation and cell cycle modules, respectively). Additionally, expression of the cell differentiation module (cluster 6) was correlated with pseudotime, as expected (Pearson's R=0.62; **Fig. 4.9**, panel d).

Using the same ASE-based interaction test as applied to identify dynamic eQTL, reflecting ASE variation across pseudotime, we assessed how the genetic regulation of gene expression responded to these cellular contexts. Briefly, we tested for genotype by environment (GxE) interactions using a subset of four co-expression modules as markers of cellular state, while accounting for effects that can be explained by interactions with pseudotime:

$$\text{ASE} = \alpha_1 \text{pseudotime} + \alpha_2 \text{pseudotime}^2 + \beta \text{factor} + \boldsymbol{\psi}, \tag{4.2}$$

where the test performed is: $H_1 : \beta \neq 0$ (using a likelihood ratio test), and factor represented one of the four co-expression modules, and was quantified as the normalised mean expression levels in each cell across all genes in the cluster. This approach extends previous work using ASE to discover GxE interactions [491, 497], taking advantage of the resolution provided by single-cell data.

[7] The Pearson correlation across all cells was used as the similarity/'affinity' metric.
[8] GO enrichment of each cluster was performed by Fisher's exact test in Python using GOATOOLS [495].

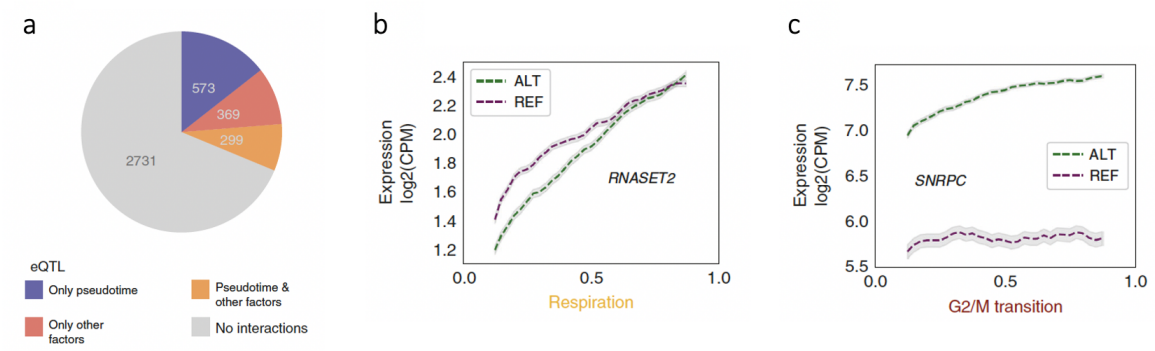


Fig. 4.16: Allele-specific expression reveals interactions with fundamental cellular processes. (a) Results summary: numbers of eQTL identified as displaying GxE interactions with pseudotime (purple), displaying GxE interactions with other cellular contexts but not with pseudotime, (after appropriately accounting for pseudotime, red), displaying GxE interactions with both pseudotime and at least one other cellular context (yellow), and displaying no GxE interactions at all (grey). Significance is assessed at FDR < 10%. (b) ASE variation for two example eQTL SNPs that show GxE interactions (FDR < 10%) and tag cancer-associated GWAS variants. The eQTL for *RNASET2* (rs2247315) tags a risk variant for basal cell carcinoma, and its effect size is modulated by cellular respiration, while the eQTL for *SNRPC* (rs9380455) tags a risk variant for prostate cancer and is responsive to the G2/M transition of the cell cycle. For each cell, cellular contexts were inferred using GO annotations of co-expression modules. Shaded regions indicate standard error (+/- 1 SEM).

We identified 668 eQTL that had an interaction effect with at least one factor (at FDR < 10%), with many of these eQTL having no evidence for an interaction with differentiation. Indeed, 369 genes had no association with pseudotime, but responded to at least one other factor. Conversely, of the 872 dynamic eQTL, 299 were also associated with GxE effects with other factors, whereas 573 were exclusively associated with pseudotime.

These interactions encompass regulatory effects on genes and SNPs with important functional roles. Specifically, 95 interaction eQTL variants overlap with variants previously identified in genome-wide association studies (GWAS, LD $r^2 > 0.8$). For example, the effect size of a *RNASET2* eQTL is sensitive to cellular respiratory metabolic state. This eQTL SNP is in LD ($r^2 = 0.86$) with a GWAS risk variant for basal cell carcinoma [498]. Furthermore, an eQTL for *SNRPC* showed sensitivity to the G2/M state, and is in LD ($r^2 = 0.92$) with a GWAS risk variant for prostate cancer [499] (**Fig. 4.16**). These cellular factors vary not only across cells in the experiments considered here, but also across cells *in vivo*, across individuals, and across environments. Thus, these examples illustrate the versatility of our single cell dataset and how it can provide regulatory information about variants in contexts beyond early human development.

Finally, we explored whether we could detect higher order interaction effects, where the genetic effect varies with a cellular state in different ways along differentiation, effectively testing for genotype x environment x environment (GxExE) interactions. To this end, we fitted a linear model with fixed effects for differentiation and each of the factors, plus a combined term (factor x pseudotime):

$$\text{ASE} = \alpha_1 \text{pseudotime} + \alpha_2 \text{pseudotime}^2 + \alpha_3 \text{factor} + \beta(\text{pseudotime} * \text{factor}) + \boldsymbol{\psi}, \quad (4.3)$$

where we test: $H_1 : \beta \neq 0$.

This approach identified 176 genes with significant higher order (GxExE) interactions between a genetic variant, pseudotime, and at least one other factor. For example, an eQTL for *EIF5A*, was responsive to G2/M state, especially early in differentiation, as measured using ASE (**Fig. 4.17**).

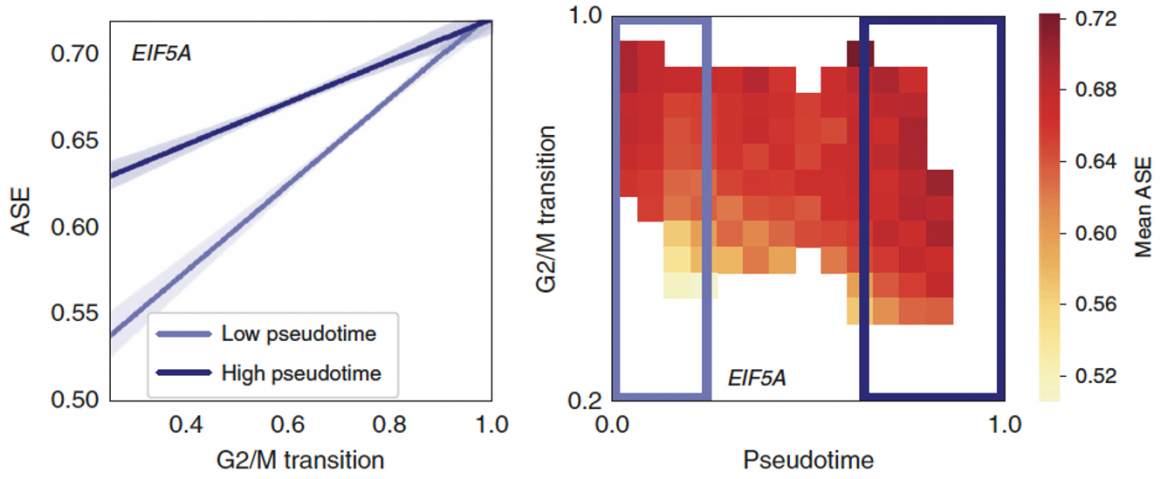


Fig. 4.17: Second order GxE interactions with fundamental cellular processes.
Figure by Daniel Seaton. Higher order interaction example: an eQTL variant for *EIF5A* (rs7503161) is affected by a GxExE higher order interaction with both pseudotime and the G2/M transition. Left panel: effects of G2/M transition on ASE for cells with low and high pseudotime. Regression lines are indicated with 95% confidence intervals for the 30% of cells with lowest and highest pseudotime values. Right panel: heatmap of averaged ASE for cells falling within the specified windows of pseudotime and G2/M transition. Only values for windows containing more than 30 cells are shown (n=6,423 cells in total).

4.7 Early markers are predictive of differentiation efficiency

The final piece of analysis we performed on this dataset was based on the observation that iPSC lines have been shown to vary in their capacity to differentiate, as demonstrated by previous studies [479]. This motivated us to look at whether we could detect clear differences in the ability to differentiate among our 126 lines. We used the average pseudotime value at the latest time point considered (day3) as a measure of differentiation efficiency in our analysis. This was motivated by the observation that there was significant variation across cell lines, which remained consistent across replicate differentiation experiments of the same cell line (**Fig. 4.18**, n=33). Exploiting the scale of our study and the pooled experimental design, we set out to look for potential markers of differentiation efficiency that are accessible prior to differentiation.

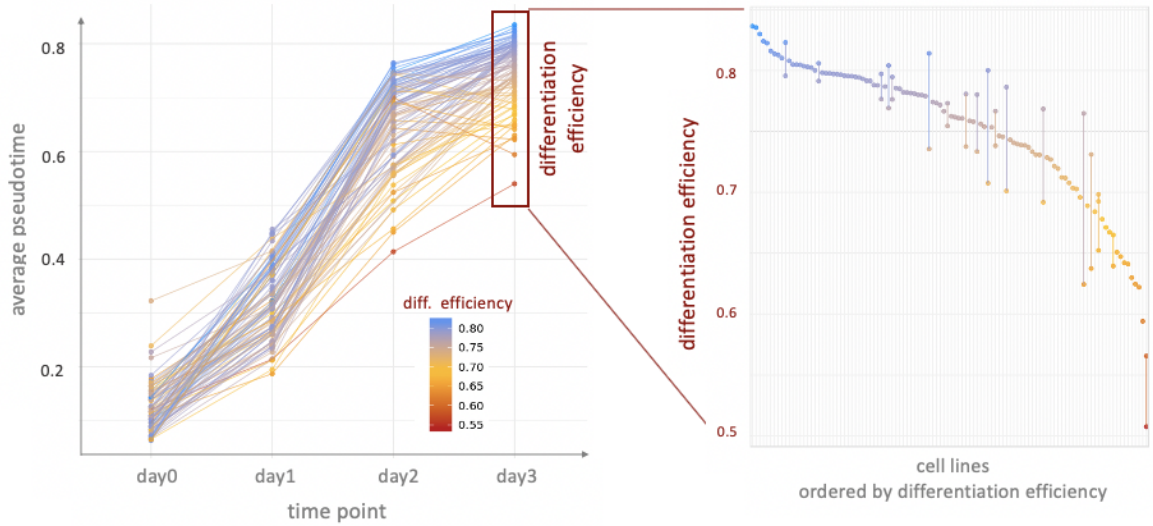


Fig. 4.18: Line-to-line variation in differentiation efficiency.
Variation in differentiation efficiency across cell lines. Left: average pseudotime values for each line, showing trajectories over time for 98 cell lines, coloured by differentiation efficiency. Shown are 98 cell lines (out of 126) for which we had sufficient data at all time points (> 10 cells at each time point). Differentiation efficiency of a cell line was defined as its average pseudotime value across all cells on day 3. Right: differentiation efficiency across cell lines (points), and consistency of individual cell lines differentiated in multiple experiments (vertical bars, n=33).

First, we looked for genetic markers. Given our small sample size and consequent low statistical power, we limited our pool to the set of 4,422 eQTL lead variants at any of the

three developmental stages[9]. We tested each variant for association with differentiation efficiency using a linear mixed model, similar to eq. (2.31):

$$\text{Differentiation efficiency} = \beta * \text{Marker} + \text{Experiment} + \text{Donor} + \boldsymbol{\psi}, \tag{4.4}$$

where Experiment is a random effect grouping sets of samples from the same experiment, and Donor is a random effect grouping samples from the same donor (and cell line). Here, a Marker is a genetic variant (i.e., eQTL SNP), and it is modelled as a fixed effect, with corresponding weight β. The models were fitted using the lme4 package in R [500], and significance was determined using a likelihood ratio test ($H_1 : \beta \neq 0$). This identified only one significant association, with the eQTL variant for *DPH3*, at FDR < 10%. In an attempt to validate this finding, we performed an additional set of differentiations in HipSci iPSC lines derived from individuals that were not part of the variant discovery, selected based on genotype at this variant (n=20). In these experiments, differentiation efficiency was measured by the percentage of CXCR4+ cells on day3:

$$\%\text{CXCR4+} = \beta * \text{Marker} + \boldsymbol{\psi}. \tag{4.5}$$

While the direction of effect was consistent, the association was not statistically significant (p value = 0.24, Student's t-test), probably reflecting low power at this sample size. We conclude that larger sample sizes will be required to conclusively identify genetic predictors of *in vitro* differentiation efficiency.

Secondly, we asked whether levels of gene expression at the iPSC stage (prior to differentiation) could represent molecular markers of differentiation efficiency. We used the same model as in eq. (4.4), with Marker in this case indicating the expression of genes at iPSC stage. This revealed 38 associations at FDR < 10%, out of 11,231 genes tested (**Fig. 4.19**). A subset of those genes (9/38) were also observed when using independent bulk RNA-seq data from the same cell lines to look for an association between gene expression and differentiation efficiency (nominal p value < 0.05). We note that expression of these marker genes is largely orthogonal to pseudotime itself, i.e. these are not genes that vary across pseudotime, and only as a consequence of that are associated with differentiation efficiency (which is defined as average pseudotime at day3, **Fig. 4.19**). We noted that 17 of the 38 differentiation-associated genes were located on the X chromosome, reflecting a significant enrichment of X chromosome genes (24.5-fold enrichment, p value = $8x10^{-16}$, Fisher's exact test). For all of the X chromosome genes, higher expression was associated with reduced

[9] rather than testing all variants genome-wide. Note that this would be essentially a GWAS, and normally a GWAS is considered to be well powered at around 1,000 individuals.

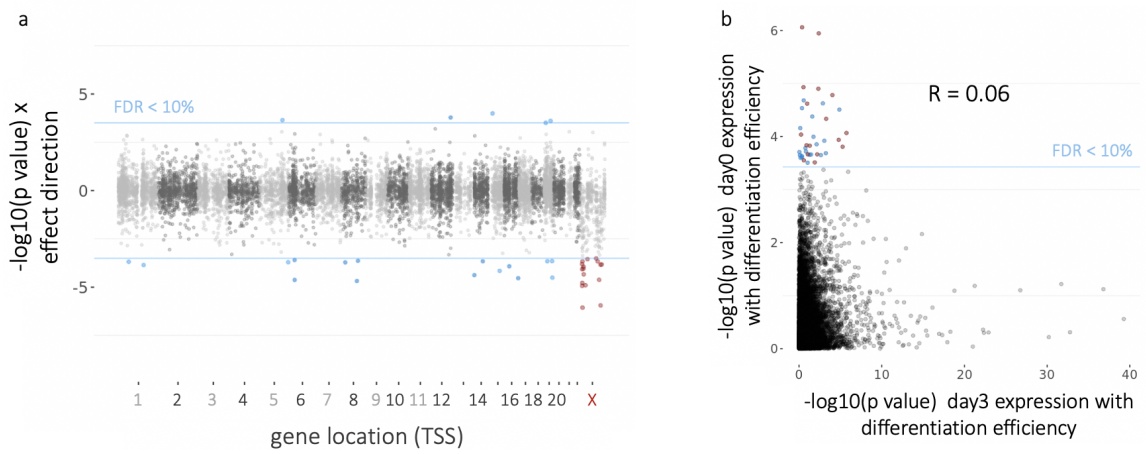


Fig. 4.19: Associations between iPSC gene expression levels and differentiation efficiency. (a) Genome-wide analysis to identify markers of differentiation efficiency (defined as average pseudotime at day3), considering iPSC gene expression levels. Displayed are negative log (nominal) p values signed by the direction of the effect. The horizontal blue lines denote the significance threshold (FDR = 10%, Benjamini-Hochberg adjusted). Significant associations are shown in blue for autosomal genes and in red for X chromosome genes. (b) Scatter plot comparing, for all genes, the statistical significance of associations with differentiation efficiency for their expression on day0 and day3. The y axis shows the association between the expression of the gene on day0 (i.e. in iPSC) and differentiation efficiency of cell lines. The x axis shows the association between the expression of the gene on day3 and differentiation efficiency. The correlation between these associations is R = 0.06, demonstrating that the genes identified as predictive markers of differentiation efficiency (i.e. those with high values on the y axis) are not the genes that define differentiation capacity (i.e. those with high values on the x axis). Significance measured as -log10(p value). Significant genes are coloured as in (a): autosomal genes in blue, X chromosome genes in red.

differentiation efficiency (**Fig. 4.19**). The majority of these associations (14/17) persisted when performing the same analysis for only female lines (p value < 0.05), indicating variation beyond differences between sexes.

These results are consistent with observations made by other groups, showing that X chromosome reactivation is a marker of poor differentiation capacity for iPSCs in general (see **page 38**) [501, 502]. We did not identify any other striking patterns in the genes identified other than the reported over-representation of chromosome X genes, partly due to a small sample size.

4.8 Discussion

Here, we generated a map of early endoderm differentiation across human iPSC lines from 125 unrelated individuals (**Fig. 4.1, 4.8**). This offers a unique and powerful tool which allows

the interrogation of the role of genetic heterogeneity in early human development.

First, we characterise the effects of common genetic variants on gene expression at three distinct developmental stages, adapting methods traditionally used for bulk RNA-sequencing to single cell expression data (**Fig 4.11** and see **Chapter 3** for ample discussion on single cell eQTL mapping). This was one of the first single cell eQTL mapping studies (after [159] and [160]), and the first with over 100 individuals (previous largest sample size was 45). Additionally, we mapped eQTL in mesendoderm and definitive endoderm cells, providing the first eQTL maps (single cell or otherwise) at these key developmental stages. In this application, the benefit of using single cell expression profiles resides mainly in the ability to define homogeneous cell populations in an unbiased manner, resulting in a higher number of detected eQTL (**Fig. 4.11**). We can also use this tool to start assessing the amount of sharing of eQTL signal between closely related developmental stages. In particular, we found that about one third of the identified eQTL at any given stage was specific to that stage (**Fig. 4.12**).

While relevant from a developmental biology perspective, the discretisation into three developmental stages (iPSC, mesendoderm, definitive endoderm) is somewhat arbitrary, as the differentiation trajectory clearly appears as a continuum (**Fig. 4.8, 4.10**). To reflect this, we exploited this resource to identify hundreds of dynamic eQTL, i.e. eQTL whose strength is modulated by differentiation time in a continuous manner (**Fig. 4.14**). Reassuringly, we found that eQTL dynamics were largely independent of total gene expression dynamics (**Fig. 4.15**). These findings nicely complement results from a similar study [302], where they identify eQTL in cells from iPSC lines that are differentiated toward cardiomyocytes across more time points (16) but for fewer individuals (19). I note that we cannot completely rule out that differentiation itself may be genetically regulated, although the results presented in **section 4.7**, where we found no common genetic variant associated with differentiation efficiency indicates that those effects, if present, are probably negligible for our dynamic eQTL analysis.

Additionally, we extend the concept of context-specific eQTL [143, 161, 491] to single cell resolution, identifying eQTL which are modulated by specific cellular states including cell cycle phases and preferential metabolic pathways, thus fully utilising the power of single-cell transcriptomics. These results highlight the power of using a single cell approach for eQTL mapping, which allows detailed annotation of changing eQTL effects across heterogeneous cell types and cell states, with the ability to better interpret the context-specific role of individual genetic variants (**Figures 4.16, 4.17**).

A further advantage of the application of single-cell transcriptomics in this study was to enable the pooling strategy. While the feasibility of pooling samples has already been demonstrated for PBMCs [402], in this study we have extended this to cell lines differentiated together in culture. This strategy provided higher throughput, and enabled the characterisation of line-to-line variability in terms of differentiation efficiency in a controlled setting. While the differentiation protocol considered here (to definitive endoderm) is short and efficient, other protocols (e.g. to generate neurons [503]) are much more challenging, making a pooling strategy useful for scaling up these protocols to population-scale (this aspect is further discussed in the next chapter, **Chapter 5**). There are some possible drawbacks of a pooled design. For example, paracrine signalling [504] could affect differentiation dynamics of cell lines grown together and obscure genetic effects. Additionally, although we have considered replicates of the same line across different experimental pools, our study is based on a single iPSC line per donor. In the future, experimental designs may consider multiple lines per donor, which, however, would require a different barcoding scheme as these cannot be discriminated using genetic barcodes. As a result of this experimental design, we cannot definitively distinguish between donor and cell line effects.

In summary, our results demonstrate the power of combining iPSC line pooling and scRNA-seq to investigate development and genetics *in vitro*. Sorting of cells along different cellular states allows eQTL context-specificity to be probed in detail across many axes of cellular variation. The scRNA-seq readouts also provide a rich description of the progress of differentiation over time across different cell lines.

This work acts as a proof of principle study, where we establish the feasibility of combining a pooled experimental design, differentiation of human iPSCs and scRNA-seq readouts across several individuals. In the next chapter (**Chapter 5**), we apply a similar set up to a larger scale and much more complex differentiation protocol, with cells differentiating to dopaminergic neurons, which are preferentially lost in Parkinson's disease. This other protocol generates more mature cell types, which are directly disease-relevant, and thus allows us to characterise the genetic component of differentiation across a larger spectrum of human development and disease. Additionally, we will have more power to explore differences in differentiation capacity across more lines and after a much longer differentiation period.

5

Population-scale differentiation of iPSCs to a neuronal fate

The work described in **Chapter 4** acted as a proof of principle study, where we demonstrated the feasibility of pooling cells from several lines prior to differentiating them towards an endodermal fate. This means that in a single experiment we can obtain data from many independent donors, which in turn allows us to increase throughput of these studies thus enabling population-scale genetics to be performed. Additionally, the single cell readouts make it possible to trace back the donor of origin of each cell, without the need for any barcoding. We and others have shown that single cell RNA-seq can be used to map eQTL and, despite the pooling, we retain enough cells per individual to do so successfully. Finally, by profiling differentiations of several lines we can start to disentangle differences in differentiation efficiency across lines and experimental batches.

In this second study, we considered a larger-scale experiment in terms of both the number of donors (from 125 to 215) and cells (from around 40,000 to over 1 million) and apply similar principles to a more challenging differentiation protocol, considering iPSCs differentiating towards a midbrain neuronal fate. First, the use of a droplet-based scRNA-seq technology allows us to assay a much larger number of cells, providing an overview of the cell types generated by this protocol. Second, the larger number of cell lines included, and the longer protocol, allows us to dive deeper into the differences across lines in terms of their efficiency to differentiate, and allows us to start exploring possible causes. Lastly, the closer resemblance of the differentiated cells to primary tissues enables the exploration of the effects of disease-associated variants on relevant cell types both at a specific stage and across development.

Contributions This work is the result of a collaboration between the Stegle, Merkle, Marioni and Gaffney labs, and was funded by Open Targets (https://www.opentargets.org). The data was generated by Dan Gaffney's lab at the Wellcome Trust Sanger Institute, and the experiments were led by Julie Jerber, who also contributed to the interpretation of the results. Madeline Lancaster generated and helped annotating the organoid data (section 5.3.1). The statistical methods and analyses described in this chapter were co-supervised by Dan Gaffney and Oliver Stegle with some input from Florian Merkle, John Marioni and Natsuhiko Kumasaka. Daniel Seaton was originally the lead computational member of the team, and I replaced him more recently upon his departure from the lab. As a result, Daniel Seaton processed the data and performed QC (sections 5.2.2., 5.2.3). I then refined the cell type annotation and mapped our data to existing scRNA-seq datasets (section 5.2.4). The work presented in sections 5.3 and 5.4 is intrinsically joint work between Daniel Seaton and myself, as these are originally analyses and observations made by Daniel which I later followed up and added to upon his departure from the lab, as well as in response to reviewers' comments. Daniel Seaton and I developed and implemented the statistical methods to perform eQTL mapping (largely building on methods I had developed previously, described in Chapter 4) and I led the eQTL analysis, including comparison of different methods (section 5.5.1) and between results across cell types and other tissues (sections 5.5.2, 5.5.3). Finally, Natsuhiko Kumasaka ran the colocalisation analysis (section 5.6), the results of which Julie Jerber and I then explored and summarised.

The code for processing, analysing and plotting the data is open source and freely accessible here: https://github.com/single-cell-genetics/singlecell_neuroseq_paper. Julie Jerber, Daniel Seaton, Dan Gaffney, Oliver Stegle and I wrote the manuscript, with input from Florian Merkle and John Marioni. The paper [445] is available at : https://www.nature.com/articles/s41588-021-00801-6 as:

Julie Jerber*, Daniel D. Seaton*, Anna S.E. Cuomo*, Natsuhiko Kumasaka, James Haldane, Juliette Steer, M Patel, D Pearce, M Andersson, Marc Jan Bonder, Ed Mountjoy, Maya Ghoussaini, Madeline A. Lancaster, the HipSci Consortium, John C. Marioni, Florian T. Merkle, Oliver Stegle, Daniel J. Gaffney. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics*, 2021 (* equal contributions).

I generated all figures presented in this chapter, except where indicated otherwise in figure legends.

5.1 Introduction

As discussed in previous chapters, genetic variation can significantly alter cell function, for example by leading to changes in gene expression. Human iPSCs are a promising cellular model for assessing the cellular consequences of human genetic variation across different lineages, developmental states and cell types. In particular, human iPSCs enable the study of developmental stages and stimulation conditions that would be challenging to access *in vivo*. The creation of cell banks containing hundreds of iPSC lines [294] provides an exciting opportunity to carry out population-scale studies *in vitro* [447, 302, 301, 300]. However, differentiating iPSCs is costly and labour-intensive, and differentiation experiments are difficult to compare due to substantial batch-to-batch variation (**section 1.2.5**). Thus, experiments with more than a handful of cell lines remain a significant challenge. Moreover, most iPSC differentiation protocols generate a heterogenous cell population, of which the target cell type represents only a subset [505, 298, 506, 507]. This variability in differentiation outcomes hampers efforts to assess the genetic contributions to cellular phenotypes.

Single cell profiling has enabled 'multiplexed' experimental designs, where cells from multiple individuals are pooled together [447, 507]. Pooling improves throughput and allows experimental variability between differentiation batches to be rigorously controlled, by enabling cell type heterogeneity to be accounted for in downstream analysis. As we have seen in the previous chapter (**Chapter 4**), multiplexed experimental designs have only been applied to one short differentiation protocol [447], which generated cells corresponding to very early stages of development, and thus have not captured differentiation progression toward a mature cell fate. Population-scale pooling during long-term differentiation offers the opportunity to examine the effect of common genetic variants on gene expression in each cell population produced over neural development, providing a foundation for future mechanistic studies.

Here, we develop and apply a multiplexing strategy to profile the differentiation and maturation of more than two hundred iPSC lines derived from the HipSci towards a midbrain neural fate, including dopaminergic neurons (DA). DA are involved in motor function and other cognitive processes and play key roles in neurological disorders, including Parkinson's Disease (PD)[1] [509, 510]. Additionally, we expose some cells to an oxidative stress, which is thought to play a role in PD [511].

[1] Parkinson's disease (PD) is a progressive neurodegenerative disorder, characterised by the loss of midbrain DA neurons. These neurons control motor behavior, and, as they degenerate, they result in several motor features of the disease, such as bradykinesia, rigidity, resting tremor, gait disturbances and postural instability [508].

5.2 Single cell map of iPSCs neuronal differentiation

5.2.1 Experimental strategy and data generation

215 feeder-free iPSC lines were selected from 215 unique, healthy, unrelated donors from the HipSci consortium [294]. Cells from multiple iPSC lines were pooled together in 17 pools, each containing cells from 7 to 24 lines. 24h after plating, neuronal differentiation of the pooled lines to a midbrain lineage was performed, as described by Kriks *et al.* [284]. To capture transcriptional changes during neurogenesis and neuronal maturation, scRNA-seq was performed from cells captured at day 11 (midbrain floorplate progenitors), day 30 (young post-mitotic midbrain neurons) and day 52 (more mature midbrain neurons). The three time points were selected based on the data available in the original paper, where molecular profiling, biochemical and electro-physiological data defined developmental progression of midbrain DA neurons [284]. The timeline was aligned to theirs: in their paper they described days 11, 25, 50 as, respectively, midbrain DA progenitors, time of cell cycle exit, and long term neurons. Day 30 was selected instead of day 25 to enrich for young post-mitotic neurons. Additionally, half of the cells on day 51 were exposed to a sub-lethal dose of rotenone, a chemical stressor that preferentially leads to DA death in models of PD [511]. Droplet-based scRNA-seq was performed using the 10X Genomics™ technology [389]. After QC, a total of 1,027,401 cells was retained across 17 cell pools and four conditions - day 11, day 30, day 52 untreated and day 52 rotenone-treated (**Fig. 5.1**).

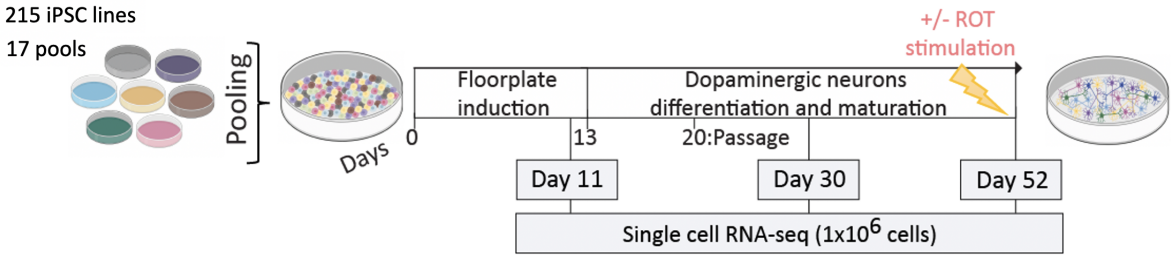


Fig. 5.1: Experimental Design.
Figure created by Julie Jerber. Experimental design for pooled differentiations of iPSCs to midbrain dopaminergic neurons. The three time points (day 11, day 30, day 52) at which cells were collected for scRNA-seq profiling are shown. On day 51, 50% of the cells were stimulated with rotenone (ROT) for 24h, to induce an oxidative stress. Single cell RNA-seq data from 215 iPSC lines (for 215 donors) across 17 pools were collected for a total of over a million cells.

5.2.2 Demultiplexing donors from pooled experiments

For each of the 17 pooled experiments, donors (i.e. cell lines) were demultiplexed using demuxlet [402], considering genotypes of common exonic variants (MAF > 1%) available from the HipSci bank, and a doublet prior of 0.05. Only single cells for which donor identification was successful were considered further. This QC step filtered out two kinds of droplet: droplets that contained two or more cells from different donors, and droplets containing no cells, but enough free-floating RNA to pass the CellRanger UMI filter.

5.2.3 Normalisation, dimensionality reduction, and clustering

Independent analysis of each time point allowed efficient batch effect correction, as all samples were from the same time point, containing similar mixtures of cell types. Moreover, by reducing the number of cells analysed together, computational tasks were made more tractable. In particular, the following steps were performed (at each time point): counts were normalised to the total number of counts per cell. Next, only genes with non-zero counts in at least 0.5% of cells were retained and the top 3,000 highly variable genes were selected, after correcting for the mean-variance relationship in expression data. The first 50 principal components (PCs) were then calculated, and batch correction was applied on the level of PCs using Harmony [512], with each 10X sample treated as a distinct batch. UMAP and clustering was performed using the resulting transformed PCs. In particular, clustering was performed using Louvain clustering [513] with 10 nearest neighbours. Data processing steps besides batch correction were performed using the Scanpy package [440]. This identified a total of 26 clusters (6, 7 and 13 clusters at day 11, day 30, day 52, respectively **Fig. 5.2**).

Next, clusters were assigned to cell types using a set of literature-curated marker genes for major brain cell types (n=48 marker genes, see **Fig. B.11**). When two clusters showed the same gene set enrichment, they were assigned the same cell type identity (see next section).

5.2.4 Cell type annotation

Cell type annotation was carried out independently at each time point (day 11, day 30 and day 52). For midbrain dopaminergic neurons, which is the target cell type of this protocol, I also performed additional analyses to verify the cell type identity. In the next section, I describe the mapping from clusters (identified unbiasedly using the entire transcriptome) to cell types (using literature-curated gene markers).

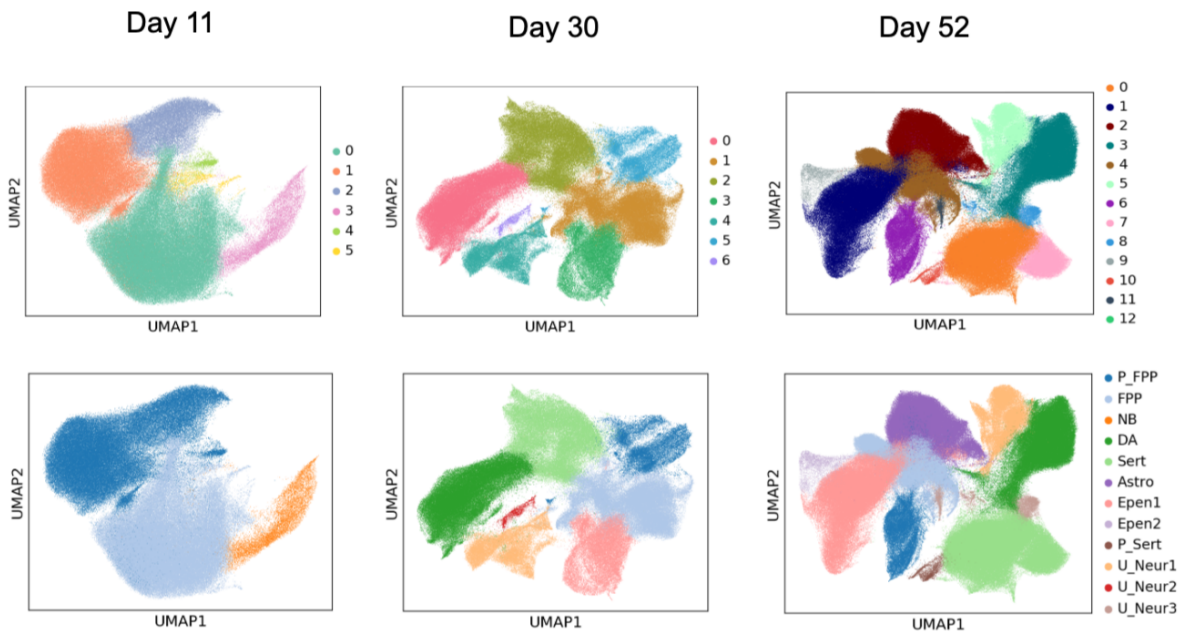


Fig. 5.2: Clustering and cell type assignment.
At each individual time point (day 11, day 30, day 52), cells were clustered using Louvain clustering [513], after normalisation and batch correction using Harmony [512]. Subsequently, clusters were annotated as cell types using known marker genes. When two clusters showed the same gene set enrichment they were computationally assigned to the same cell type identity. (a) UMAPs of cells sampled at each time point and coloured by cell clusters. (b) Same UMAPs as in (a), this time coloured by assigned cell types. Astro: Astrocyte-like, DA: Midbrain dopaminergic neurons, Epen1,2: Ependymal-like, FPP: Floor plate progenitors, NB: Neuroblasts, P_FPP: Proliferating floor plate progenitors, P_Sert: Proliferating serotonergic-like neurons, Sert: Serotonergic-like neurons, U_Neur1,2,3: Unknown neurons.

Day 11

Three cell type populations were identified at day 11. The two most prevalent ones, which constituted circa 96% of all cells at this time point, were classified as proliferating and non-proliferating midbrain floorplate progenitors (both expressing *LMX1A* and *FOXA2* and expressing *MIK67, TOP2A* when proliferating [514], **Fig. B.11**). The third cell population (making up the remaining 4% of day 11 cells) was labelled a neuroblast (NB) population, based on the expression of pro-neuronal genes *NEUROD1*, *NEUROG2* and *NHLH1* [515, 516], **Fig. 5.2, B.11**).

Day 30

At day 30, cells with floorplate progenitor (23%) and proliferating progenitor (7%) identity could still be detected, whereas the neuroblast population was not seen any longer.

Additionally, five new cell types were identified. Four of these additional cell types appeared neuronal and one was non-neuronal, as characterised by the expression (or lack thereof) of the pan-neuronal markers *SNAP25* and *SYT1* [517]. Of the four neuronal populations, two could be assigned to a midbrain neuronal identity. The first expressed canonical DA markers *NR4A2*, *PBX1*, and *TMCC3* [514, 518, 519] and was labelled as a population of midbrain dopaminergic neurons (DA, 27%). The second cell population expressed some serotonergic neuronal markers (*TPH2, GATA2* [520]) and was categorised as serotonergic-like (Sert, 21%) neurons. One additional large neuronal population (expressing *SNAP25* and *SYT1*), expressed both midbrain DA markers and cortical markers, and thus could not be assigned to a specific neuronal identity (Unknown neurons 1, around 8%). Finally, one smaller neuronal population (less than 2%) could also not be assigned to a specific identity (Unknown neurons 2). The only non-neuronal cell type identified at day 30 expressed all the classical markers of ependymal cells (Ependymal 1 [521], 11%, **Fig. 5.2, B.11**).

Day 52

At day 52, the cell types identified at day 30 were largely recapitulated (**Fig. 5.2**). Floor-plate progenitors were present in smaller proportions (13 and 5%). In addition to DA, Sert, the mixed neuronal population 1 and the ependymal-like cell population 1, a population of astrocyte-like cells could be identified, which were unique to day 52 (Astrocyte-like [522, 523]). Finally, three additional rare cell types (present in less than 2% of cells sampled at any time point) were detected, namely a second ependymal-like population (Ependymal 2), a population of proliferating neuronal serotonergic-like cells (Prolif. serotonergic-like neurons), and one additional neuronal population which could not be annotated unambiguously (Unknown neurons 3, **Fig. 5.2, B.11**).

We note that in general, we are careful to clarify that these are *in vitro*-generated cell types, which will not be exactly the same as their *in vivo* counterpart, especially for cell types that were not the target of the protocol used - thus the nomenclature xx-like, e.g. astrocyte-like, and serotonergic-like. In the next section I discuss this further for the two cell populations which we assigned to a midbrain neuronal identity.

Serotonergic-like neuronal population

First, the population we call serotonergic-like was an especially hard one to define. Serotonergic neurons are located in the same brain region as dopaminergic neurons (the midbrain), and share some common functions and gene markers. However, whilst dopaminergic neurons have been very well characterised, partly because of their involvement in PD, serotonergic

neurons have not been studied as much, and there are no well defined markers (at least not in human, whereas there are a few mouse studies [520]). Additionally, there is no *in vivo* single cell reference dataset. The only study containing a human serotonergic neuronal cell population to the best of my knowledge is La Manno *et al.* [514], which contained only 14 cells. In the same study they also derive midbrain neurons from human iPSCs, but do not obtain any serotonergic neurons. Since our population expressed some, but not all, canonical serotonergic markers, we could not unambiguously say that these were serotonergic neurons, hence the name serotonergic-like.

Dopaminergic neuronal population

In contrast, human midbrain dopaminergic neurons are much better annotated, and in particular there exist published *in vivo* datasets we can compare to. Specifically, to confirm the dopaminergic identity of our DA cell population, I compared our cells to three datasets: human iPSC-derived dopaminergic neurons and human fetal cells from La Manno *et al.* [514] and substantia nigra samples from post-mortem donors from Welch *et al.* [416].

To perform the mapping, I performed joint PCA (using the multiBatchPCA from the batchelor package, implemented in R) and batch correction (using MNN [412]) of log-normalised counts (using scater [438]) from our data and each of the three reference datasets. The set of genes used was the union of 2,000 highly variable genes (HVGs, using the trendVar function from scran) from our data and 2,000 HVGs from the reference dataset. Next, I asked which reference cell each of our cells was most similar to (i.e. 'mapped to', using queryKNN as implemented in BiocNeighbors, with k=1 nearest neighbour).

I mapped our DA cells to the set of all neurons from each of three datasets. First, I compared to the La Manno *et al.* iPSC data, and found that 99% of our DA cells mapped to the 'iDAb' population. Second, to the La Manno *et al.* embryonic data. 85% of our DA cells mapped to one of the dopaminergic populations in the reference, i.e. 39% to 'hDA1', 36% to 'hDA2', and 10% to 'hDA0'. Finally, we mapped our DA cells to the Welch *et al* post-mortem data, and found that for 91% cells mapped to 'NEURODop', with the remaining 8% mapping to a population of inhibitory neurons, 'NEUROinh1'. These combined analyses provide confidence in the identity of DA neurons from our iPSC differentiation model.

5.2.5 Data overview

For visualisation purposes, we also performed a combined analysis of a random subsample of 20% of all cells (after QC) from the three time points. In this case, the Harmony batch correction was performed across pools (rather than across individual 10X samples, as before). A joint UMAP projection of cells collected across all time points, stimuli and lines revealed broad co-clustering of cell types (using the labels described previously, see **Fig 5.2**), but with noticeable differences between time points and stimuli (**Fig. 5.3**).

Substantial variation in the cell type proportions could be observed, across time points and stimuli (**Fig. 5.3**). For example, the proportion of DA cells was significantly reduced upon rotenone stimulation (30% reduction, Fisher's exact test, p value = 2.2×10^{-16}), which was consistent with previous observations that dopaminergic neurons are most affected by apoptosis due to oxidative stress [524–526]. Collectively, our population-scale scRNA-seq analysis revealed a diverse repertoire of cell types, enabling both the study of cell line differentiation propensity (**sections 5.3, 5.4**) and the identification of genetic variants that affect expression in a cell type-specific manner (**sections 5.5, 5.6**).

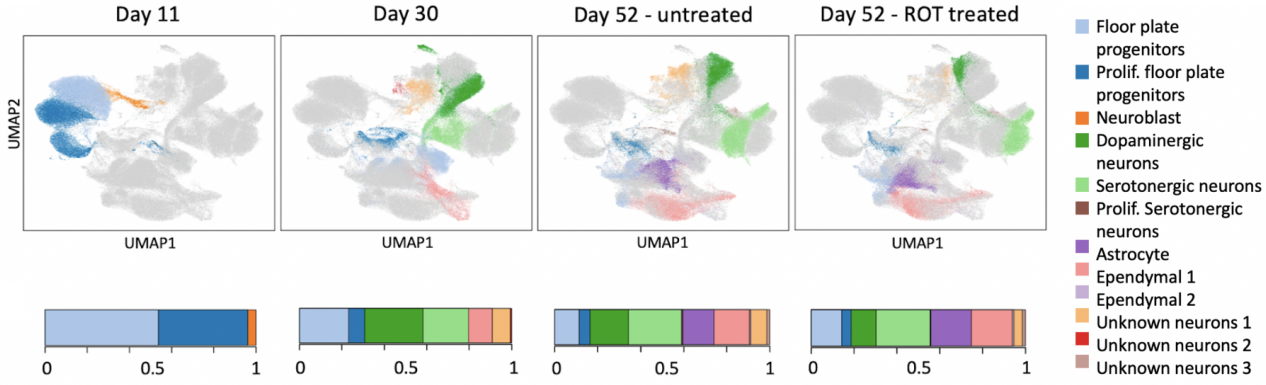


Fig. 5.3: Overview of study.
Top: UMAP plots of a subset of 205,416 cells assayed (20% of the total), coloured by cell type identity. Cells that were not collected at a given (time point, stimulus) condition are shown in light grey. ROT: rotenone; Prolif: Proliferating. Bottom: Bar plots showing, for each condition, the fraction of cells assigned to each cell type.

5.3 Line-to-line variation in neural differentiation efficiency

The great diversity of cell types generated by this protocol raised the question of whether it may be driven by variation in differentiation outcome between different iPSC lines, which, as we have seen, is prevalent in iPSC differentiation studies (see **section 4.7** as well as other studies, e.g. [527, 506]). Yet, as we discussed, the biological basis for this high variability in differentiation outcomes between lines remains largely obscure, which complicates efforts to rationally select cell lines for different applications. Here, we found substantial variation in the proportions of different cell types produced by different iPSC cell lines at each time point. For example, the proportion of day 52 untreated cells assigned to DA neurons ranged from 1% to 100% from line to line. (**Fig. 5.4**).

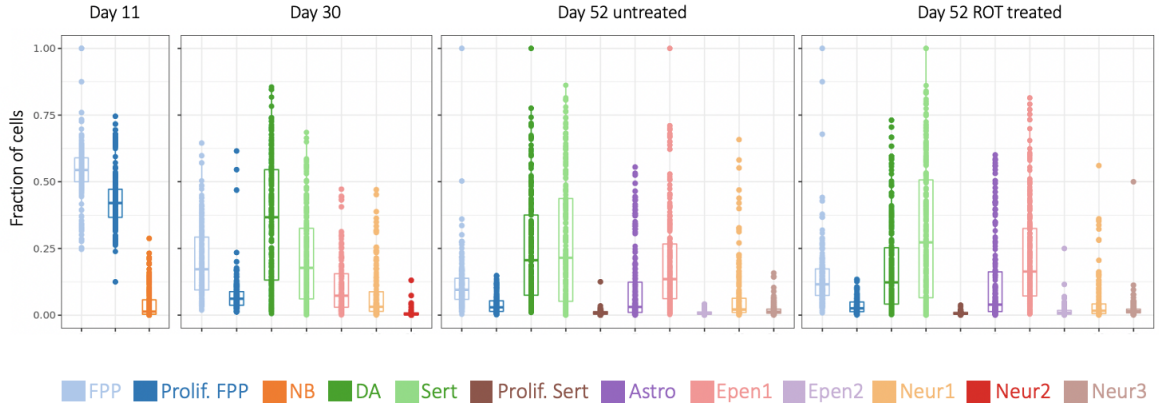


Fig. 5.4: Cell type fractions across lines. Box plots showing, for each cell type, the proportions of that cell type across cell lines at day 11, day 30, untreated day 52, rotenone (ROT) treated day 52. Each point indicates a different cell line. Astro: Astrocyte-like, DA: Dopaminergic neurons, Epen1,Epen2: Ependymal-like, FPP: Floor Plate Progenitors, NB: Neuroblasts, P_FPP: Proliferating FPP, Sert: Serotonergic-like neurons, P_Sert: Proliferating Sert, U_Neur: Unknown Neurons.

Looking at the cell type fractions per cell line and pool across time points, we observed a bimodality in the data, with roughly 2/3 of the iPSC lines mostly making DA and Sert at day 30 and day 52, and the other 1/3 making very few midbrain neurons but many glial cells (Ependymal-like and Astrocyte-like) instead (**Fig. 5.5**). When we performed PCA of such cell type fractions matrix, we identified the proportion of midbrain neurons (DA and Sert) on day 52 as the largest axis of variation (PC1, 47% variance, **Fig. 5.5**). Since DA and Sert cells are derived from similar progenitor populations *in vivo*, it is not surprising that both populations are observed in our differentiation experiment [528, 529]. This motivated us to estimate a 'neuronal differentiation efficiency' for each iPSC line, defined as the sum of the fractions of DA and Sert cells produced on day 52 (**Fig. 5.5**).

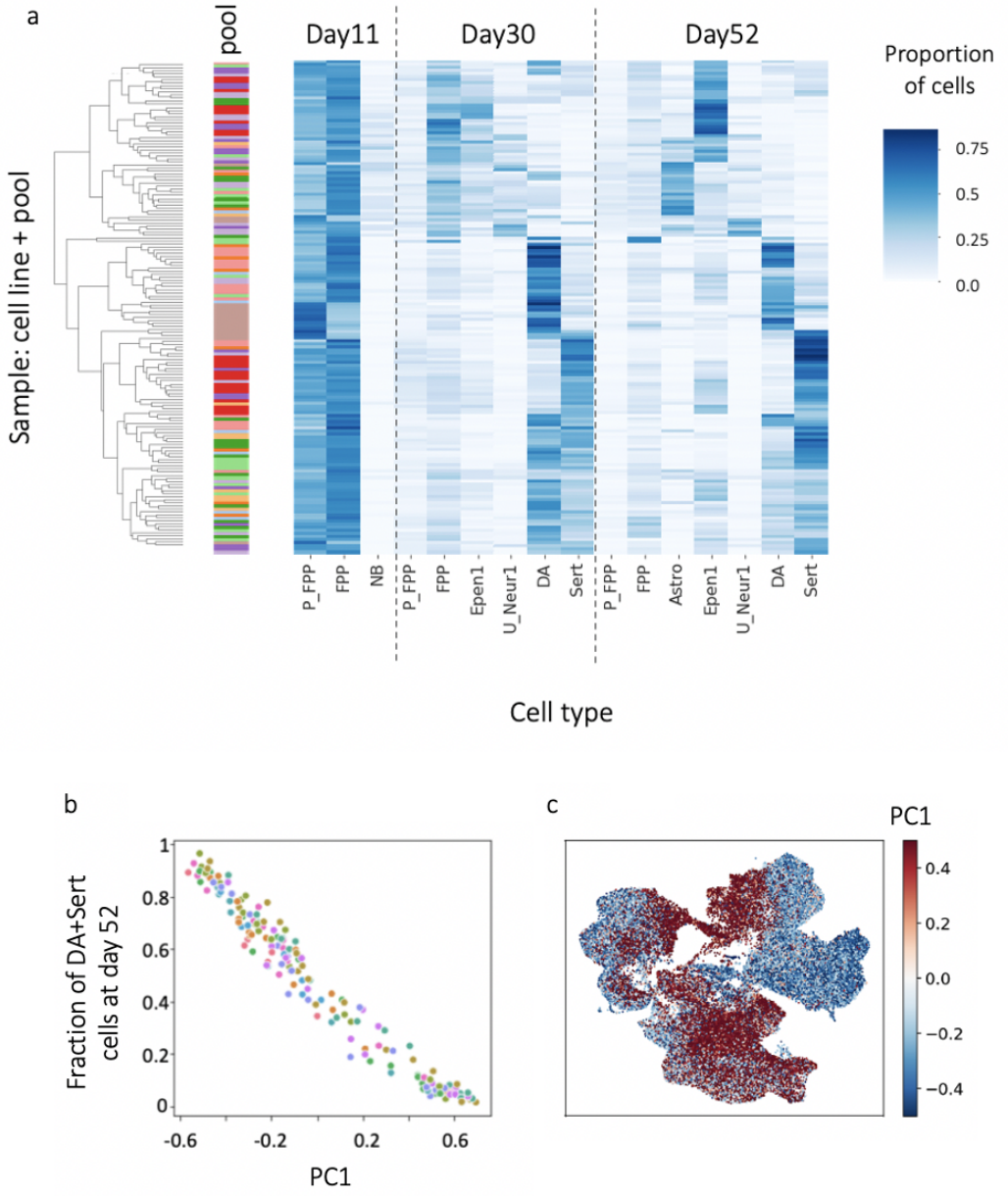


Fig. 5.5: Distribution of cell proportions at day 52 and definition of neuronal differentiation efficiency. Cell proportions were generated for each cell type and time point for all combinations of cell lines and pools with at least 10 cells at all time points (10 pools). (a) Heatmap of the resulting cell proportion matrix. Pools are shown in the first bar and the colours indicate in which of the 10 pools each line was differentiated. Rows (i.e. cell line, pool combinations) were hierarchically clustered according to Euclidean distance. (b) Comparison of the first principal component (PC1) to the sum of fractions of dopaminergic and serotonergic-like neurons present on day 52. (c) UMAP of cells included in (a), coloured by PC1. Astro: Astrocyte-like, DA: Dopaminergic neurons, Epen1: Ependymal-like, FPP: Floor Plate Progenitors, NB: Neuroblasts, P_FPP: Proliferating FPP, Sert: Serotonergic-like neurons, U_Neur: Unknown Neurons.

We assessed the reproducibility of this measure of neuronal differentiation efficiency using data from 32 lines that were differentiated twice, in two different pools. Importantly, we found that iPSC line neuronal differentiation efficiency defined in this way was highly reproducible between different pools (Pearson's R = 0.75; p value = 2×10^{-6}, **Fig. 5.6**).

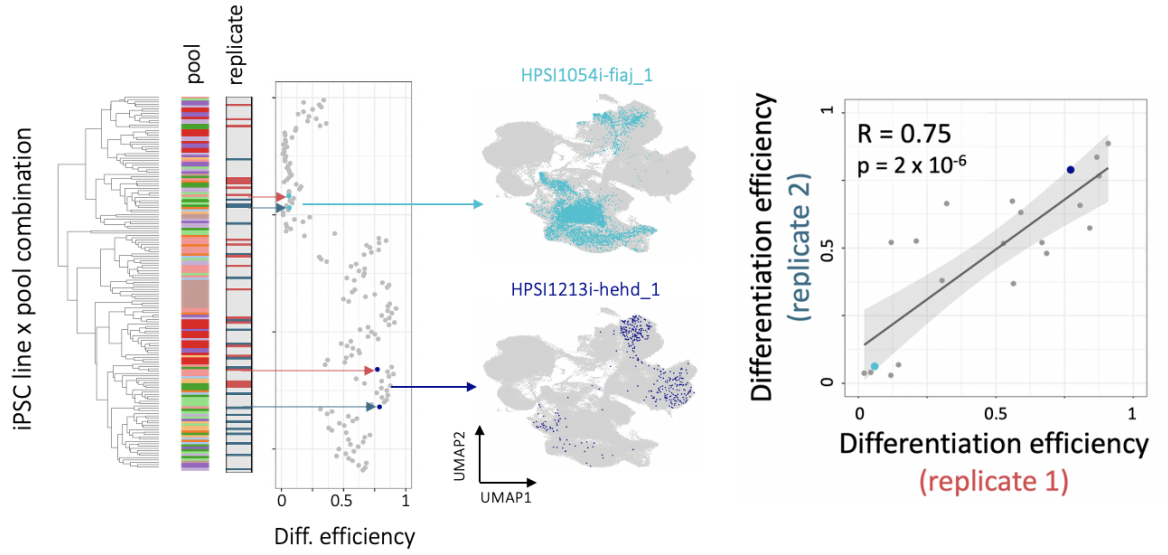


Fig. 5.6: Reproducible variation in differentiation trajectories.
(left) Hierarchical clustering of (cell line, pool) combinations (same as **Fig. 5.5**) by neuronal differentiation efficiency. Colours in the first bar indicate in which of 10 pools (for which we had data at all time points, used to define neuronal differentiation efficiency) each line was differentiated. Differentiation replicates for lines present in two pools, are indicated in the second bar (red for replicate 1 blue for replicate 2). (middle) UMAPs, highlighting the distributions of cells on day 52 for two selected cell lines with low and high differentiation efficiencies respectively (HPSI0514i-fiaj_1, in seagreen and HPSI1213i-hehd_1, in dark blue). (right) Scatter plot showing estimated neuronal differentiation efficiency between differentiation replicates (i.e. cell lines differentiated in two different pools, out of the 10 pools considered here, n=21). Highlighted are the same two cell lines.

5.3.1 Organoids

Given the robustness of our measure of neuronal differentiation efficiency, we next wondered if it was generalisable to other differentiation approaches. We therefore differentiated a pool of 18 lines (pool 4) into cerebral organoids for 113 days (as previously described in Lancaster *et al.* [530]) and profiled the resulting cell populations using scRNA-seq (11,445 cells). The same steps of dimensionality reduction, batch correction and clustering applied to the midbrain dataset were applied to the cerebral organoid data. These steps identified eight clusters that were labelled as different cell types (i.e. neuronal cells, intermediate progenitor cells, radial glial progenitor cells, satellite cells, mesenchymal cells, myotube and Wnt and

PAX7 positive cells) using 24 marker genes (**Fig. 5.7**). We found that the proportion of brain cell types (all neuronal, glial, and neural progenitor cells) produced by each line in the cerebral organoids was strongly correlated with neuronal differentiation efficiency as estimated from the dopaminergic differentiation (R = 0.94; p value = 2×10^{-5}; n=12). Taken together, these results strongly suggest that variation in iPSC neuronal differentiation efficiencies arise primarily due to cell-intrinsic factors. Furthermore, the consistency of neuronal differentiation efficiency suggests that these properties extend to neuronal differentiation more generally.

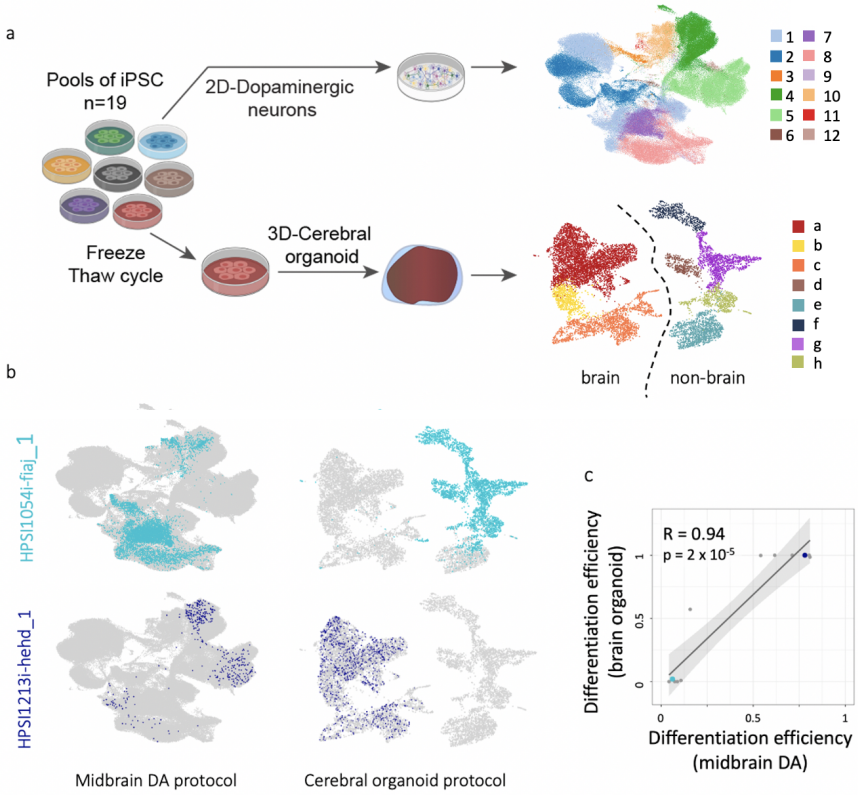


Fig. 5.7: Differentiation efficiency in cerebral organoids.
(a) Experimental workflow for single cell profiling of iPSC-derived cerebral organoids using one pool containing 18 cell lines, profiled using scRNA-seq after 113 days of differentiation. UMAPs for both our dopaminergic neuron study and the cerebral organoid study (1: floor plate progenitors (FPP), 2: proliferating FPP, 3: neuroblasts, 4: dopaminergic neurons (DA), 5: serotonergic-like neurons (Sert), 6: proliferating Sert, 7: astrocyte-like, 8: ependymal-like (Epen) 1, 9: Epen2, 10: unknown neurons (UN) 1, 11: UN2, 12: UN3. a: neurons, b: intermediate progenitors, c: radial glial progenitors, d: satellite cells, e: mesenchymal cells, f: myotube, g: paired box (PAX)7+ cells, h: Wnt+ cells.) (b) UMAPs of two representative cell lines making non-brain and brain cell types in the organoid study. (c) Scatter plot of neuronal differentiation efficiency as measured using midbrain dopaminergic neuronal differentiation (x axis) versus neural differentiation efficiency as measured in organoid differentiation (y axis) for a subset of 12 iPS cell lines in common. Highlighted are the same two cell lines as in (b).

5.4 iPSC expression can predict neuronal differentiation efficiency

Motivated by the reproducibility of differentiation outcomes across multiple independent pools, we set out to explore possible predictors (similar to the analysis described in **section 4.7**). The idea was that, if we could find characteristics that could be measured in iPSCs and that would predict a bad differentiation outcome, they could become a useful tool to select the most suitable lines prior to differentiation.

We began by testing for associations between neuronal differentiation efficiency and other experimental and biological factors. Those included cell line passage number (p value = 0.77), donor sex (p value = 0.008), chromosome X activation status (p value = 0.01), and PluriTest scores [280] (p value = 0.01). Although some of these were nominally significant, they explained little variation as compared to line-specific effects, when we performed variance component analysis, by modelling:

$$\text{neuronal differentiation efficiency} = \text{Donor/Line} + \text{Pool} + \text{Sex} + \text{Age} + \boldsymbol{\psi}, \qquad (5.1)$$

where Line (which cannot be distinguished from Donor), Pool, Sex and Age are all modeled as random effects (n=230 line-pool combinations). To assess specifically the effect of X chromosome inactivation status, we fitted an alternative model which was limited to the female donors (n=115 line-pool combinations):

$$\text{neuronal differentiation efficiency} = \text{Donor/Line} + \text{Pool} + \text{XCI} + \text{Age} + \boldsymbol{\psi}. \qquad (5.2)$$

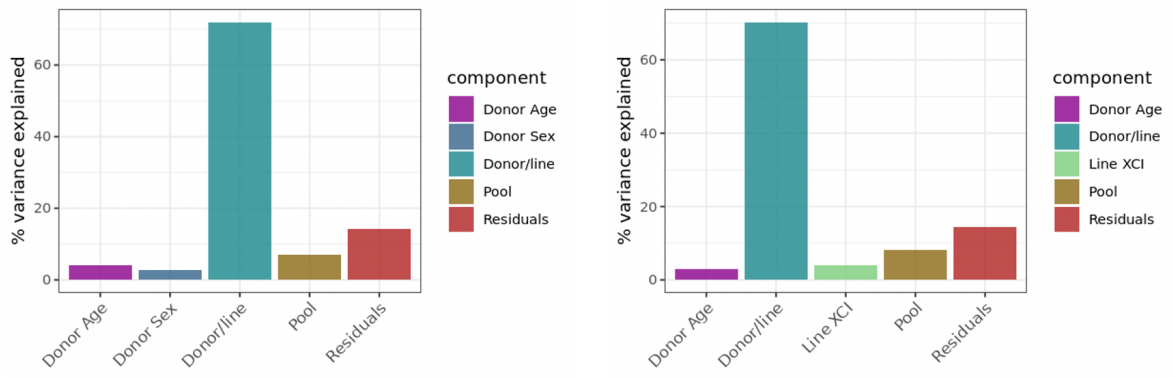


Fig. 5.8: Variance component analysis of neuronal differentiation efficiency.
Results from variance component models in eq. (5.1) and (5.2), respectively. The variance explained by each component was re-scaled to sum up to 100. XCI is categorised into 0.1-wide bins: [0-0.1]..[0.4-0.5].

We note that there is some effect of the technical batch iPSC lines were differentiated in, as it has been observed before [294, 301], yet line-specific effects are prevalent (**Fig. 5.8**). This is confirmed when we consider data from 6 lines (from pools 1, 2 and 3) that were differentiated individually, as well as in pools. When we compared our measure of neuronal differentiation efficiency for each of the lines when differentiated alone or in a pool, we found rather concordant results (R = 0.83, p value = 0.034, n = 6).

Next, we assessed whether neuronal differentiation efficiency was associated with particular patterns of gene expression in undifferentiated iPSCs. Using data from independent bulk RNA-seq data available for a subset of 184 iPSC lines included in this study [294, 129] we identified significant associations with neuronal differentiation efficiency for 2,045 genes (983 positive and 1,062 negative associations; F-test, FDR < 5%, **Fig. 5.9**).

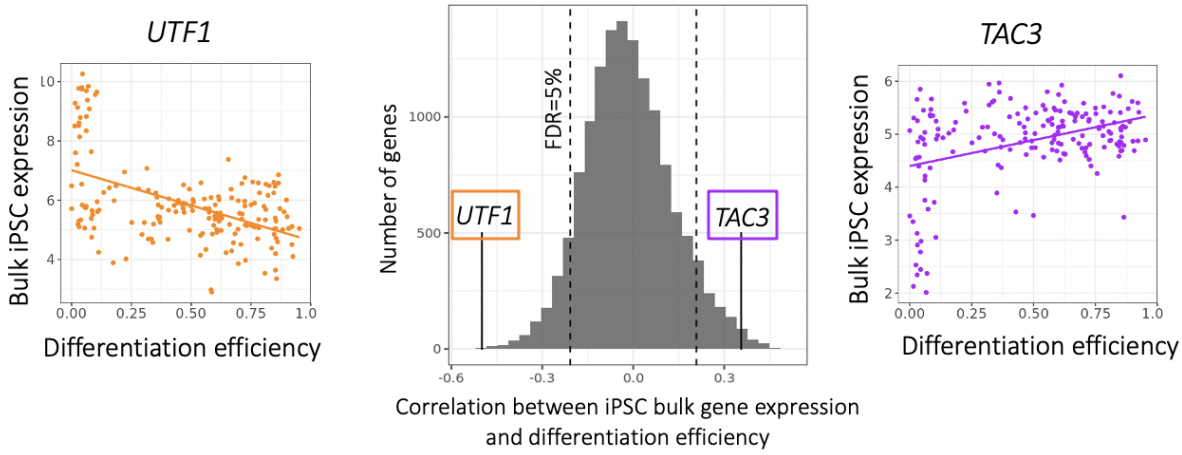


Fig. 5.9: An iPSC expression signature is associated with neuronal differentiation efficiency. (middle) Histogram of Pearson correlation coefficients between iPSC gene expression of individual genes (measured using bulk RNA-seq [129]) and neuronal differentiation efficiency. Two exemplar genes (*UTF1, TAC3*) are highlighted (left and right, respectively). *UTF1* is an example of a gene whose expression level in iPSC (based on bulk RNA-seq) is negatively correlated with neuronal differentiation efficiency (R = -0.5, p value = 3.5×10^{-13}), whereas *TAC3* is positively correlated (R = 0.38, p value = 9.8×10^{-8}).

5.4.1 A predictor of (poor) differentiation using iPSC gene expression

The examples shown in **Fig. 5.9** suggest that differentiation potential and especially poor differentiation (< 0.2) may be associated with clear expression signatures. Motivated by this observation, we used the genome-wide gene expression signature in undifferentiated iPSCs to build a model to predict poor differentiation outcomes, where we defined poor differentiation as a binary outcome (neuronal differentiation efficiency < 0.2). We used a logistic regression and obtained 100% precision at 35% recall as assessed by cross-validation. This result was robust to alternative thresholds for defining poor differentiation outcomes (**Fig. 5.10**).

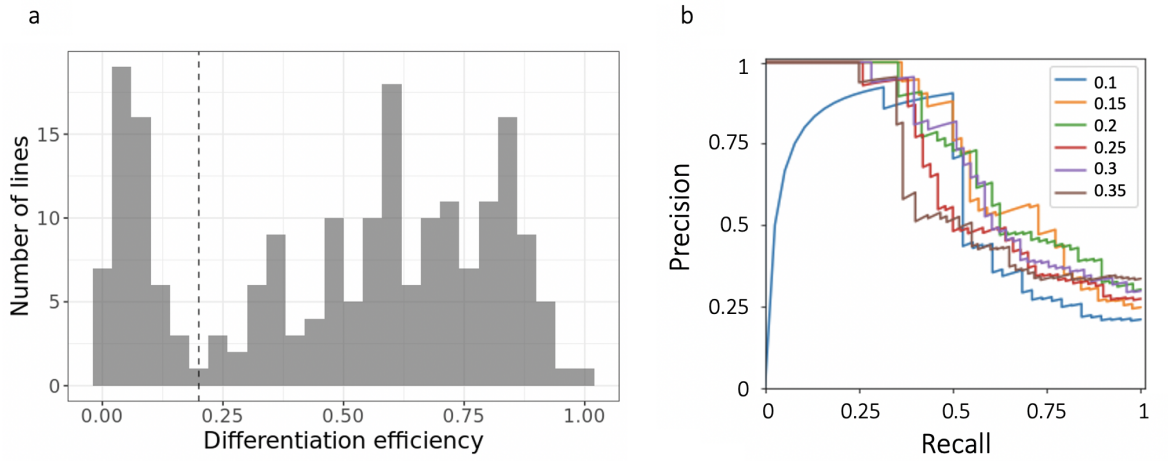


Fig. 5.10: Predicting differentiation failure from iPSC gene expression.
(a) Histogram of neuronal differentiation efficiencies across cell lines. The threshold chosen to define differentiation success or failure (i.e. neuronal differentiation efficiency = 0.2) is shown by the dashed line, separating the two modes of the distribution. (b) Precision-recall curves for a logistic regression model predicting differentiation failure from iPSC gene expression data [129] using a range of thresholds between 0.1 and 0.35 to define differentiation failure. Results are presented from leave-one-out cross validation.

We then used this model to generate predicted scores for all 812 HipSci lines for which bulk RNA-seq data was available. This analysis indicated that a substantial fraction of lines in the HipSci resource (26%) were predicted to produce < 20% neuronal cells under the differentiation conditions we tested. Furthermore, we tested whether the same experimental and biological factors previously associated with neuronal differentiation efficiency replicated in this larger sample and found consistent results. Finally, we did not observe strong concordance between the predicted differentiation outcomes of different cell lines from the same donor, suggesting that donor genetic background is unlikely to play an important role in driving differentiation biases (**Fig. B.12**).

5.4.2 A subpopulation of iPSCs is associated with poor differentiation

Next, we hypothesised that the predictive gene expression signature identified in bulk RNA-seq at iPSC state may reflect variation in the proportion of subpopulations in iPSCs. To test this hypothesis, we re-analysed scRNA-seq data from 112 iPSC lines that were assayed previously under iPSC culture conditions similar to those used here [447], 45 of which were also included in this study[2]. After processing the data using the same pipeline as used above (i.e. Harmony batch correction, Louvain clustering), we identified 5 clusters, which expressed similarly high levels of core pluripotency markers (*NANOG, SOX2, POU5F1*, **Fig. 5.11, B.13**).

We found that genes whose expression predicted poor differentiation (e.g. *UTF1*) were highly enriched in one of those clusters (cluster 2), while genes whose expression were predictive of successful differentiation (e.g. *TAC3*), were downregulated in cluster 2 relative to the remaining iPSC clusters (**Fig. 5.11**). As a validation of this hypothesis, we also tested for and confirmed a significant association between the fraction of cells in cluster 2 and neuronal differentiation efficiency for each cell line (Pearson R = -0.76, p value = 2.05×10^{-9}, **Fig. 5.11**). We used additional data from [447] to assess the consistency of the portion of cluster 2 cells across replication experiments, finding good concordance (Pearson R = 0.9; n=23, **Fig. 5.11**). Using the known relationship between iPSC bulk RNA-seq and the proportion of cluster 2 cells, we predicted this proportion for 182 cell lines included in our differentiation experiments, and confirmed the negative correlation with neuronal differentiation efficiency (Pearson R = -0.49; p value = 3×10^{-12}, **Fig. B.13**).

Finally, we also analysed an additional scRNA-seq dataset from iPSCs derived from Lymphoblastoid Cell Lines (LCLs) [444]. Using our single cell analysis pipeline, we identified a cluster of cells with a concordant (R^2=0.4) expression profile to cluster 2 (**Fig. B.14**). Combined, these results provide further evidence that an iPSC sub-population with poor differentiation capability can be consistently detected across iPSCs from different banks, and that this bias can be predicted robustly using gene expression at iPSC stage.

Importantly, despite the variability in neuronal differentiation efficiencies, we still retained significant numbers of cells across many cell lines and disease-relevant cell types and stimuli, which enabled us to explore the impact of common genetic variants on gene expression, across such cell populations (next section, **section 5.5**).

[2]This is the day 0 population from the data presented in **Chapter 4**, and the same iPSC single cell population used in **Chapter 3**.

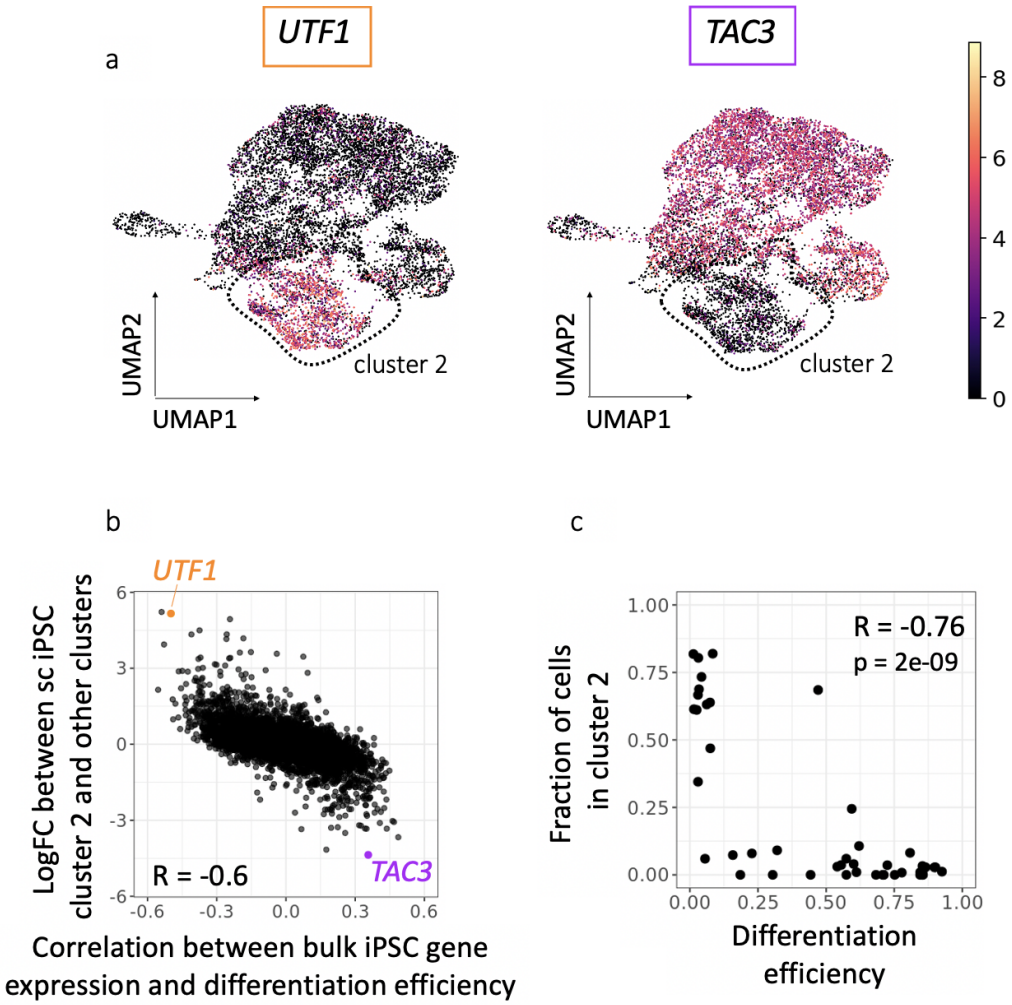


Fig. 5.11: A subpopulation of iPSCs is associated with poor differentiation.
(a) UMAPs of single-cell RNA-seq profiles in iPSCs from 112 donors from [447]. Colours denote the expression level of the two example genes from **Fig. 5.9**: *UTF1* and *TAC3*. Cluster 2 is shown by the dashed lines. (b) Comparison of marker gene association results with expression markers of the cluster 2. For each gene, the Pearson correlation coefficient of association between the gene's iPSC expression and neuronal differentiation efficiency (x axis; iPSC gene expression assessed using bulk RNA-seq, as in **Fig. 5.9**) is compared to its log fold change between cluster 2 and all other clusters (y axis, scRNA-seq). *UTF1*, and *TAC3* are highlighted. (c) Scatter plot between the proportion of cells assigned to cluster 2 (y axis) and neuronal differentiation efficiency (x axis) across 45 cell lines which were included in both sets of experiments. Where measurements across multiple pools were available for a cell line, these were averaged.

5.5 Mapping eQTL in neuronal cell types

Next, in order to understand how individual-to-individual genetic variation influenced gene expression in this system, we mapped eQTL across our identified cell types during differentiation, and in response to stimulation. Specifically, we mapped *cis* eQTL in each of the well represented[3] 'cell type'-'condition' contexts defined above i.e. the 14 distinct cell populations shown in **Table 5.1**.

	FPP	P_FPP	DA	Sert	Epen1	Astro
Day 11	✓	✓				
Day 30	✓		✓	✓	✓	
Day 52 - untreated			✓	✓	✓	✓
Day 52 - ROT treated			✓	✓	✓	✓

Table 5.1: Overview of the 14 cell populations we mapped eQTL for (rows: conditions, columns: cell types). FPP: floor plate progenitors; P_FPP: proliferating FPP; DA: dopaminergic neurons; Sert: serotonergic-like neurons; Epen1: ependymal-like cell population 1; Astro: astrocyte-like.

Cis eQTL were mapped by calculating average expression levels for each donor[4], considering common gene-proximal variants (MAF > 0.05, +/- 250 kb around genes). For each context (cell type, condition), all genes detected in at least 1% of the cells of that context were tested, and expression quantification was only included for a donor if it represented the mean of at least 10 cells. The observed variability in neuronal differentiation efficiency between lines[5] (**Fig. 5.5**) resulted in substantial differences in the number of cells collected for each donor, affecting accuracy of the estimates of aggregated expression. To account for this source of noise, we adapted commonly used eQTL mapping strategies [447] based on LMMs (**Chapter 2**) by incorporating an additional variance component into the model:

$$\mathbf{y} = \sum_{i}^{15} \alpha_i \mathbf{PC}_i + \mathbf{g}\beta + \tilde{\mathbf{u}} + \boldsymbol{\psi}, \tag{5.3}$$

where $\tilde{\mathbf{u}} \sim \mathcal{N}(\mathbf{0}, diag(\frac{1}{n_i}))$, where n_i is the number of cells for each individual i. Note that since our LMM implementation only allows one random effect component to be considered (see **page 64**), in this model we are not accounting for population structure. Consequently, we have to rely on samples being unrelated (**Fig. B.15**), and cannot consider multiple

[3] top 4 cell types per condition with at least 20% cells.

[4] Similar to the d-mean aggregation method described in **Chapter 3, section 3.8**.

[5] i.e. as we have seen, some lines made mostly neuronal cell types and barely any non-neuronal, thus expression estimates for those lines in non-neuronal cell types will be less accurate because they are estimated using very few cells, and vice versa for lines that mostly made non-neurons, and very few neurons.

observations for the same lines (e.g. across pools). Therefore, expression was aggregated at the cell line/donor level, averaged across pools for the lines assessed in more than one pool. Using this approach, we found a total of 4,828 genes with at least one eQTL in any of the contexts (hereafter 'eGenes', FDR < 5%, Storey procedure, **Table 5.2**).

	FPP	P_FPP	DA	Sert	Epen1	Astro
Day 11	2,560	2,457	-	-	-	-
Day 30	881	-	872	776	1,011	-
Day 52 - untreated	-	-	1,024	1,436	1,391	257
Day 52 - ROT treated	-	-	458	1,043	1,122	205

Table 5.2: Number of eGenes at FDR < 5% for each assessed eQTL map. FPP: floor plate progenitors; P_FPP: proliferating FPP; DA: dopaminergic neurons; Sert: serotonergic-like neurons; Epen1: ependymal-like cell population 1; Astro: astrocyte-like.

This approach greatly increased the number of detected eQTL, as compared to the base-model which does not include the noise term, i.e.:

$$\mathbf{y} = \sum_{i}^{15} \alpha_i \mathbf{PC}_i + \mathbf{g}\beta + \boldsymbol{\psi}, \tag{5.4}$$

confirming the importance of taking into account the large effect that the number of cells for each individual has on the uncertainty of the mean expression estimation (**Fig. 5.12**).

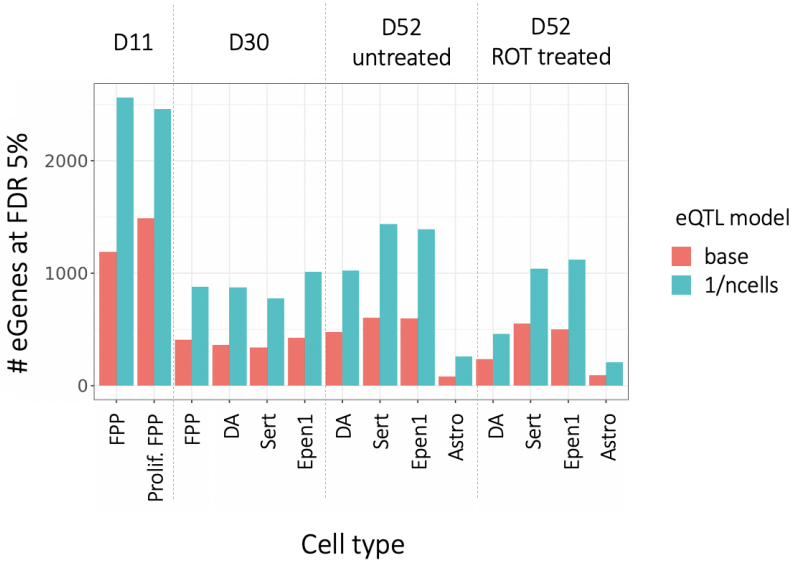


Fig. 5.12: Increase in number of discovered eQTL. Number of eGenes for each cell type, time point and stimulation discovered using either a traditional linear model (coral, from eq. (5.4)) or our enhanced model accounting for noise due to variation in the number of cells collected for each donor (seagreen, from eq. (5.3)). FPP: floor plate progenitors; P_FPP: proliferating FPP; DA: dopaminergic neurons; Sert: serotonergic-like neurons; Epen1: ependymal-like cell population 1; Astro: astrocyte-like.

The main insight from this specific analysis is that variation in cell count across donors for a given cell type/condition (**Fig. 5.4,5.5**) is a substantial source of variation in single-cell based designs. Accounting for this effect in the noise model substantially improves the ability to detect eQTL (**Fig. 5.12**). This is especially pronounced in this dataset, and motivated us to prioritise accounting for this source of variability over considering data across multiple batches (i.e. dr-mean, which had emerged as the best approach in **Chapter 3, section 3.8**).

5.5.1 Comparison with alternative eQTL methods

In order to more generally compare alternative approaches to map eQTL, for one selected cell population (untreated dopaminergic neurons at day 52), we compared our results (which yelded 1,024 eGenes, **Table 5.2**) to those obtained from alternative eQTL methods. Specifically, these methods differ in the approach taken to account for variability in cell count (or not), and in how we deal with replicate lines in the analysis. In particular, by aggregating at the donor level, we may not be accounting properly for batch differences. As an alternative, we could aggregate at the line and experiment level (similar to the dr-mean approach described in **Chapter 3**, which is also adopted in the eQTL analysis from **Chapter 4**), and use a standard kinship matrix-approach to account for the repeatedness (rather than the number of cells noise term), i.e. test the following model:

$$\mathbf{y} = \sum_{i}^{15} \alpha_i \mathbf{PC}_i + \mathbf{g}\beta + \mathbf{u} + \boldsymbol{\psi}, \tag{5.5}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{K})$; In this model, we do not account for the variability in cell count; this resulted in 471 eGenes at FDR < 5%. Another possibility would be to include the batch (and sex) directly as (several) covariates, such that:

$$\mathbf{y} = \sum_{i}^{16} \alpha_i \mathbf{pool}_i + \gamma\, \mathbf{sex} + \mathbf{g}\beta + \mathbf{u} + \boldsymbol{\psi}, \tag{5.6}$$

where $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{K})$. This resulted in markedly fewer eGenes - 320. Finally, in order to account for batch effects whilst still including the number-of-cell noise term, it is possible to only consider one experiment per line, and correct for pool, as well as sex, as covariates:

$$\mathbf{y} = \sum_{i}^{16} \alpha_i \mathbf{pool}_i + \gamma\, \mathbf{sex} + \mathbf{g}\beta + \tilde{\mathbf{u}} + \boldsymbol{\psi}, \tag{5.7}$$

where $\tilde{\mathbf{u}} \sim \mathcal{N}(\mathbf{0}, diag(\frac{1}{n_i}))$, and n_i is the number of cells for each individual i, as above. This approach, too, resulted in fewer eGenes (608) compared to our chosen approach.

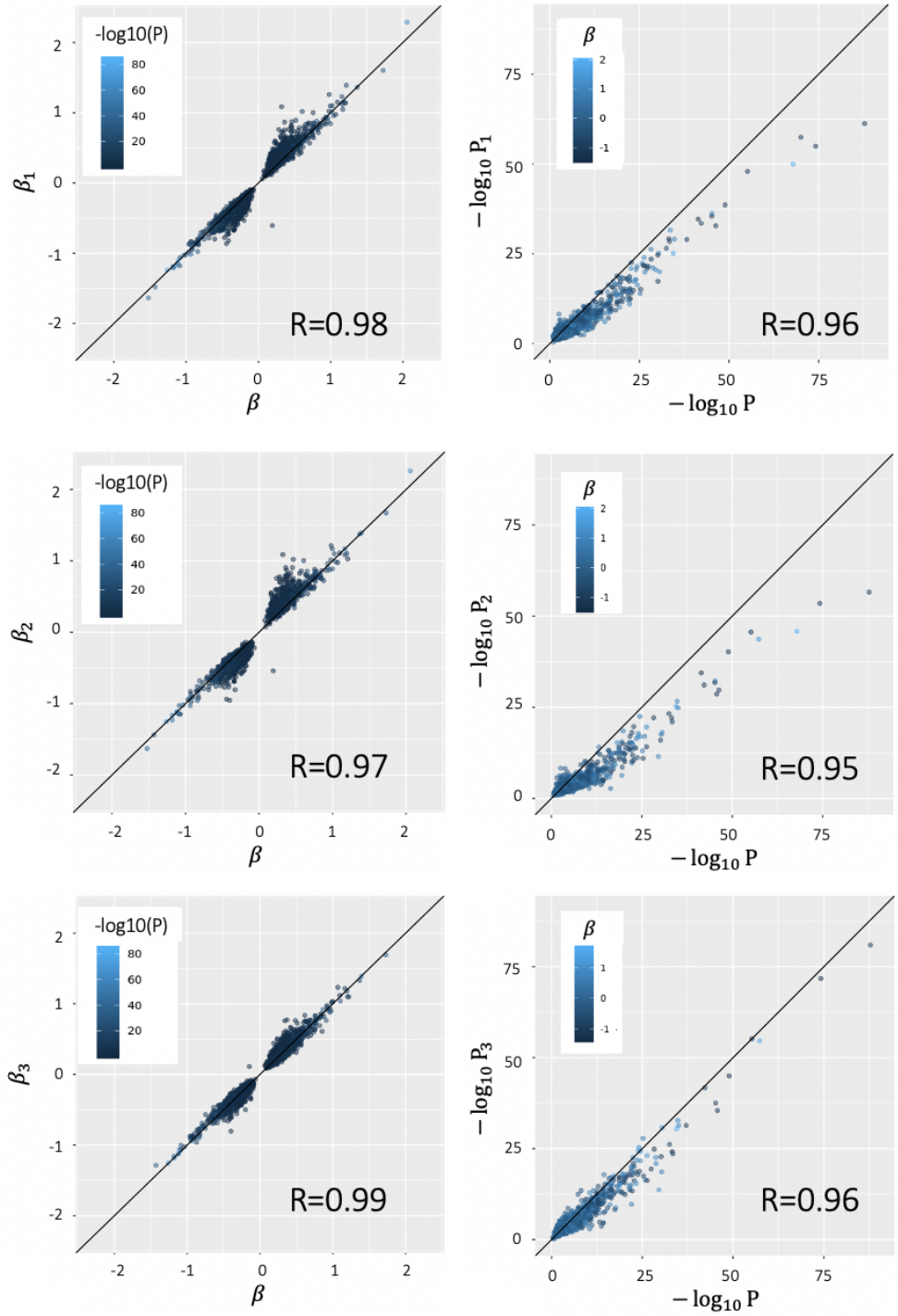


Fig. 5.13: eQTL methods comparison.
Scatter plots of eQTL effect sizes (left) and p values (right) obtained when testing association of untreated day 52 dopaminergic neuron eQTL discovered using our approach (from eq. (5.3), x axis) and each of the three alternative methods described in equations (5.5), (5.6) and (5.7), respectively (y axis). Pearson's correlations (R) are indicated.

Compared to these alternative methods, our approach resulted in the most discoveries (with 1,024 eGenes, see **Table 5.2**), yet the results were highly consistent between methods (**Fig. 5.13**), excluding the possibility that our model may be generating mostly false positives. This demonstrates that our results are robust to these specific choices, with the strategy we chose yielding more total eQTL discoveries. As an additional quality control metric, frequently used by other studies (e.g. [127]), we assess and confirm an enrichment of eQTL variants at gene promoters (**Fig. 5.14**).

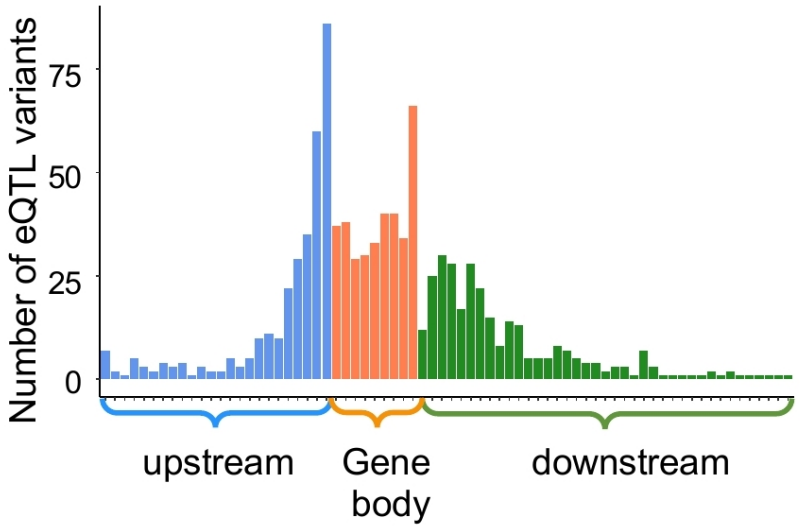


Fig. 5.14: Distribution of eQTL genomic locations.
Genomic location of eQTL lead variants relative to normalised gene coordinates, considering 1,024 eQTL identified in day 52 untreated dopaminergic neurons (using eq. (5.3)).

Moreover, the incorporation of PCs is an efficient approach for capturing global trends and hence can often replace the use of a potentially large number of technical covariates (e.g. we have 16 pools in our data alone). Additionally, accounting for pool is not straightforward in our study, as some lines (n=35) were included in two pools. Finally, we have also tested the extent to which the 15 PCs we have included in our model capture the key known covariates, which we do not model directly. When fitting a linear model to explain different covariates as a function of the 15 PCs:

$$\mathbf{cov} = \mathbf{PC}_1 + .. + \mathbf{PC}_{15} + \boldsymbol{\varepsilon}, \tag{5.8}$$

We observed that the 15 PCs explained 57% of the variance across pools (where only one pool replicate was considered for lines differentiated in multiple pools), 67% of the variance of the donor sex covariate (male=0, female=1), 78% of the X chromosome status, and 9% of average age.

5.5.2 Comparison of eQTL across cell types and conditions

Next, we set out to compare eQTL maps across cell types and conditions. First, we observed that the largest number of eQTL were detected in progenitor cell populations, likely reflecting increased detection power due to the larger number of well-represented donors (> 100 cells per donor). Next, we noted that the cumulative number of eGenes (genes with an eQTL) in each cell type increased considerably when taking into account cells further progressed along the differentiation axis, as well as upon stimulation (**Fig. 5.15**).

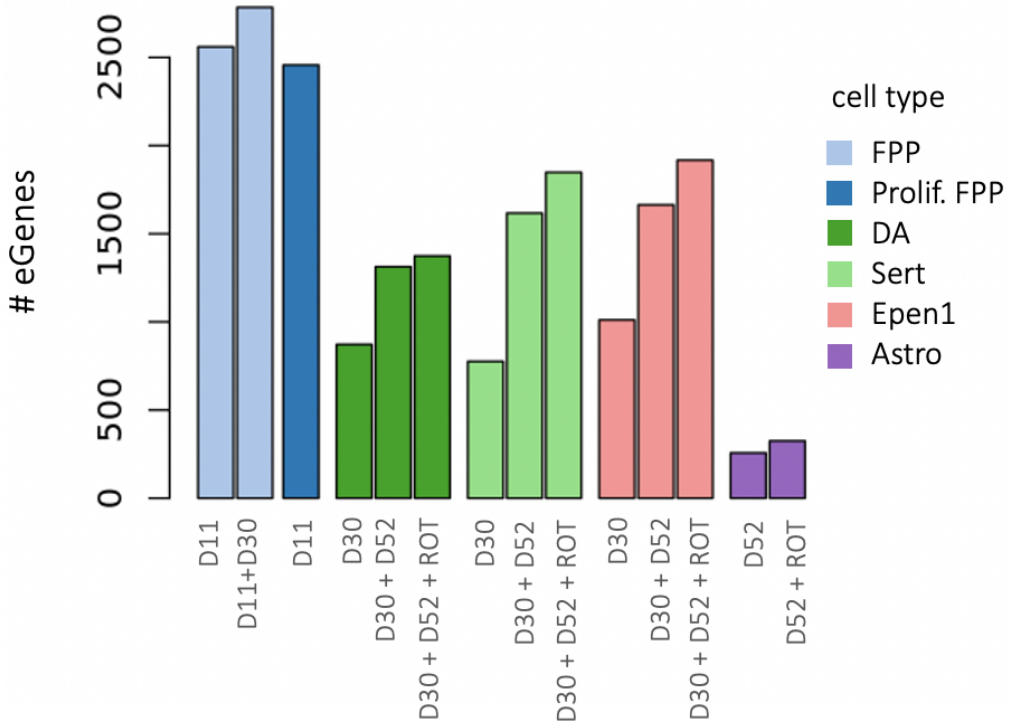


Fig. 5.15: Mapping *cis* eQTL in distinct cell contexts across midbrain differentiation.
Cumulative number of eGenes for each cell type and condition (D11 = day 11; D30 = day 30; D52 = day 52 (untreated); ROT = rotenone-treated day 52).

For example, in DA cells, eQTL mapping in matured (untreated) cells (day 52) identified an additional set of 441 eGenes (at FDR < 5%) compared to day 30 cells. An example of a timepoint-specific eGene is *HSPB1*, for which SNP rs6465098 is an eQTL in day 52 cells, but not day 30 (**Fig. 5.16**). *HSPB1* encodes a heat shock protein that plays a key role in neuronal differentiation [531] and for which changes in gene expression have been observed in neurons after ischemia [532] and associated with toxic protein accumulation in Alzheimer's disease [533, 534].

Similarly, we detected 248 additional eGenes with a rotenone-specific effect in DA and Sert neurons. As an example, the SNP variant rs12597281 is an eQTL for *ACSF3* in rotenone-stimulated serotonergic-like neurons at day 52, but not in unstimulated cells (**Fig. 5.16**). *ACSF3* encodes an acyl-CoA synthetase localised in the mitochondria and for which inherited mutations have been associated with a metabolic disorder, combined malonic and methylmalonic aciduria (CMAMMA), where patients exhibits a wide range of neurological symptoms including memory problems, psychiatric problems and/or cognitive decline [535].

These examples highlight how changes in the expression of genes known to be associated with human disease can be transient and specific to a cell type and state. More importantly, this data shows how our experimental design brings an extra level of resolution to understand disease mechanisms that were previously inaccessible from primary tissues, and opens up new experimental avenues.

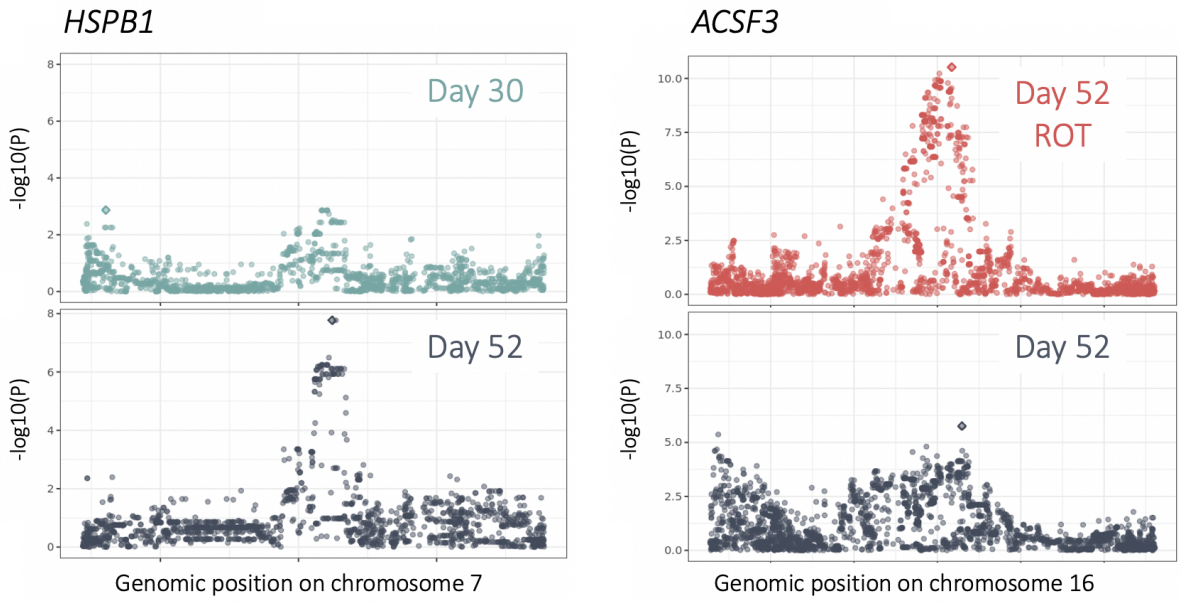


Fig. 5.16: Context-specific eQTL examples.
Left: day 52-specific eQTL for *HSPB1* in DA (rs6465098; FDR < 5%). In figure are Manhattan plots for DA cells at day 30 (top) and day 52 (bottom). Right: a rotenone stimulus-specific eQTL for *ACSF3* in serotonergic-like neuronal cells (rs12597281, right). Manhattan plots are shown for rotenone-stimulated (top) and unstimulated (bottom) Sert day 52 cells.

5.5.3 Comparison of eQTL from our study with *in vivo* maps

In order to put our eGene discovery in relation to previous studies, we compared the number of eGenes identified here with bulk eQTL maps from *in vivo* tissues from the GTEx consortium [150]. For a first coarse-grain comparison between bulk and single-cell eQTL maps, we aggregated[6] eQTL across cell types and found that the number of discovered eGenes was similar to that expected in a primary tissue of the same sample size (**Fig. 5.17**).

However, when focusing on individual cell populations, we observed fewer 'cell type'-'condition' eGenes than detected in GTEx tissues of similar sample size (**Fig. 5.17**), likely due to the uneven representation of donors across cells, which in turn results in noisier expression estimates compared to the GTEx results using bulk measurements. This result is consistent with what we observed in work presented in **Chapter 3**, where we found increased number of discoveries when mapping eQTL using bulk compared to single cell RNA-seq, even when considering the same cell type and matched individuals.

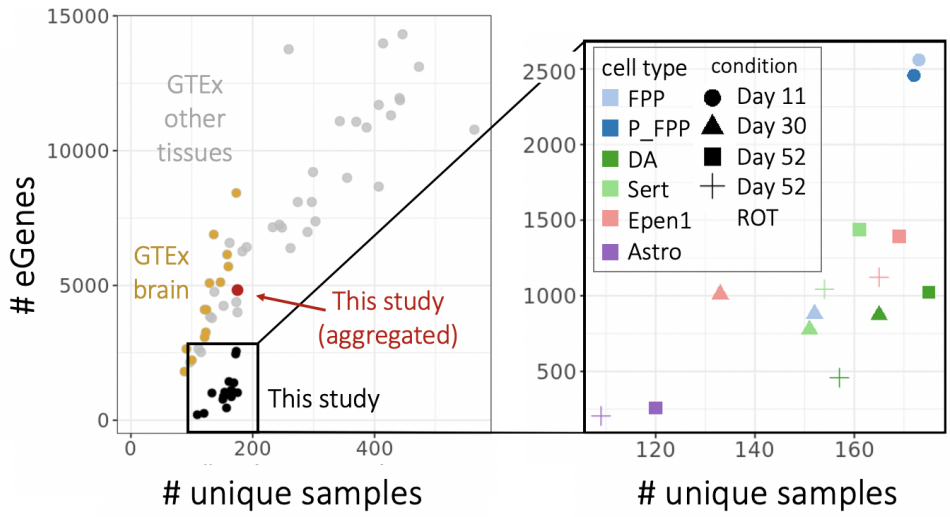


Fig. 5.17: Sample size vs number of discoveries.
Comparison of the number of genes with at least one eQTL (number of eGenes; FDR < 5%; y axis) as a function of effective sample size (number of unique donors; x axis) across studies and cell types. Left: results from overlapping eQTL results in this study with *in vivo* eQTL maps from GTEx, divided into brain tissues and non-brain tissues. The result from our study when aggregating across cell types and conditions is coloured in red. The right panel shows a magnified view of results from our study coloured by cell type and shaped by condition. Astro: Astrocyte-like, DA: Midbrain dopaminergic neurons, Epen1: Ependymal-like, FPP: Floor Plate Progenitors, P_FPP: Proliferating FPP, Sert: Serotonergic-like neurons. ROT: rotenone-treated.

[6] Considered the union of eQTL identified in any of our 14 cell populations.

A key question of eQTL maps from *in vitro* iPSC-based models is how closely these resemble eQTL maps from the equivalent primary tissues, which typically differ in cell type composition. To explore this, we tested the extent to which regulatory variants were shared between our eQTL maps and 48 *in vivo* maps from the GTEx consortium, as measured by genome-wide consistency of eQTL effect sizes (using MASHR [536]). First, reassuringly, we observed that the sharing of genetic signal[7] between our eQTL maps and GTEx tissues is considerably higher when we consider brain tissues compared to all other tissues (using the subset of 6,205 genes that were assessed in each of our cell types and in all GTEx tissues; **Fig. 5.18**, panel a).

Next, we performed a second MASHR analysis, this time including only the 13 GTEx brain tissues (as well as our 14 maps, and one iPSC map from [129]). The main motivation to do so is that in order to quantify the amount of sharing between eQTL results obtained from several tissues or conditions, MASHR only considers gene-SNP pairs that have been assessed in every one of the conditions considered, which naturally will depend on the number of genes expressed in the various conditions and that can be quantified by the different technologies used.

As a consequence, the number of genes considered decreases as the number of conditions included increases, which in turn results in the inflation of the amount of sharing. Here, in particular, excluding non-brain eQTL maps from GTEx rescued several brain-specific genes, and allowed us to assess sharing for 8,706 genes ($\sim$2,500 more). This second analysis enabled us to assess the similarity of our maps to the brain tissues in particular. We found that the extent of eQTL sharing between our eQTL maps and GTEx brain tissues increased as iPSCs were differentiated to increasingly mature neuronal cell types (**Fig. 5.18**, panel b). This result provides confidence that eQTL discovered in iPSC-derived neuronal cell types mimic eQTL maps from *in vivo* tissues. Consistent with the trend of increased sharing of eQTL signal, we also observed that the fraction of eQTL that are not represented in GTEx brain tissues decreases as the cells become increasingly mature. In particular, we identified 2,366 eQTL that could not be detected in GTEx brain tissues (q value $>$ 0.05 in any of the 13 tissues), demonstrating the utility of iPSC and scRNA-seq analysis to assess previously unexplored cell populations and therein discover regulatory changes in disease associated genes.

[7] Following recommendations by the MASHR authors [537], for each pair of conditions, we considered eQTL that were significant (local false sign rate $<$ 0.05) in at least one of the two conditions, and then assessed sharing as the fraction of those for which posterior estimates of effect size were of similar magnitude ($0.5 <$ ratio < 2) and of concordant direction of effect.

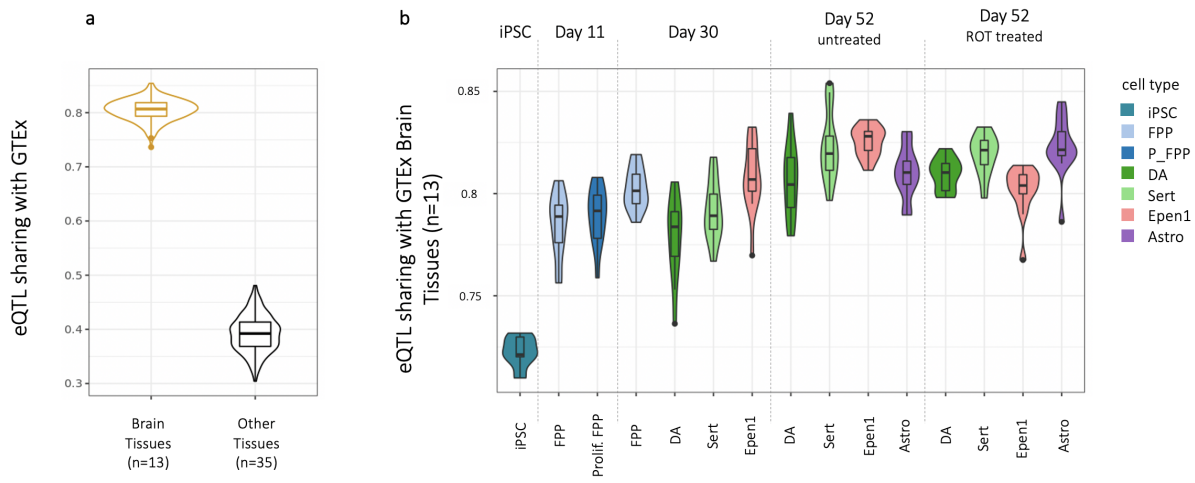


Fig. 5.18: Brain sharing is higher than non-brain sharing, and it increases over time.
(a) Box plots indicating the amount of sharing as quantified by MASHR [536] of our 14 eQTL maps considered together, with each of the GTEx brain tissues (yellow, n=13) and with each of other (non-brain) GTEx tissues (black, n=35). (b) Sharing of eQTL signals discovered in iPSCs and in our study (for each of our 14 cell types and conditions), with *in vivo* brain eQTL maps (from GTEx). Violin plots show the extent of eQTL sharing with each of 13 GTEx brain eQTL maps. Astro: Astrocyte-like, DA: Midbrain dopaminergic neurons, Epen1: Ependymal-like, FPP: Floor Plate Progenitors, iPSC: induced pluripotent stem cells, P_FPP: Proliferating FPP, Sert: Serotonergic-like neurons.

Finally, we note that quantifying the amount of sharing of eQTL signal is a notoriously difficult problem. As mentioned, MASHR considers only genes assessed in all conditions analysed. As a result, the degree of sharing may be inflated, because only relatively few highly expressed genes are included in the analysis, and those are more likely to be common eQTL. In particular using scRNA-seq we can assay fewer genes (as compared to bulk), thus the number of genes that we can assess in our single cell maps becomes the limiting factor in terms of genes included in the analysis (i.e. 8,706/8,738 genes assessed in all of our maps are also assessed in all GTEx brain maps). On the other hand, as a complementary strategy to quantify eQTL sharing, we have also considered conventional definitions of eQTL replication, based on nominal significance of lead eQTL variants discovered in each of the 13 GTEx brain tissue eQTL maps, in each of our 14 eQTL maps (**Fig. 5.19**). Notably, this comparison allows for teasing apart lack of replication versus lack of assessment of an eGene because of difference in expression. We found that of the eGenes identified at FDR<5% in each of the GTEx maps, approximately 50% were tested in our different maps (**Fig. 5.19**, panel a). For the shared fraction of genes assessed, 20-40% eQTL were nominally significant (p value < 0.05) across our 14 maps. Cumulatively, this means that 10-20% of the eQTL from a GTEx brain map could be re-discovered in our single cell maps (**Fig. 5.19**, panel b).

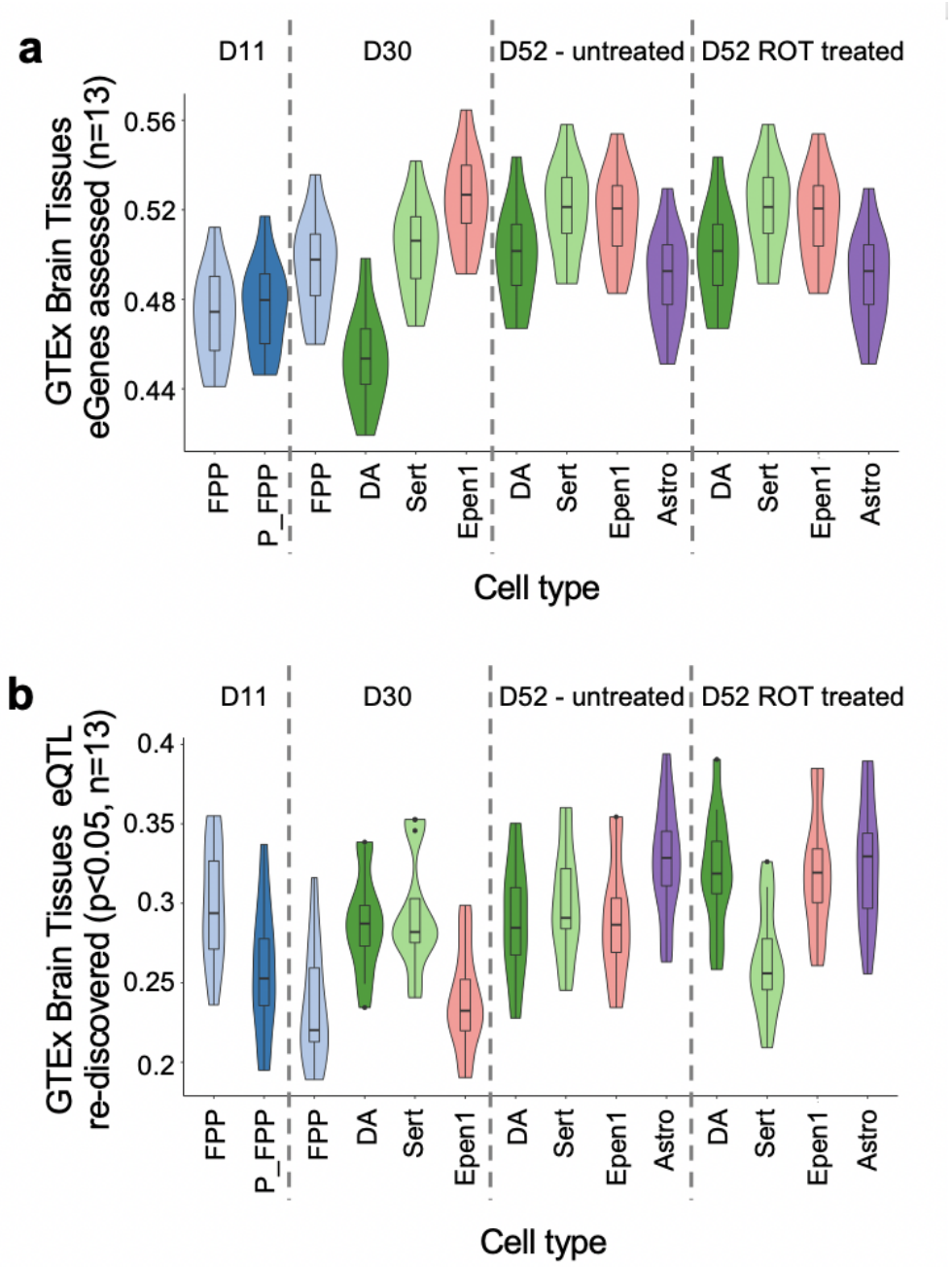


Fig. 5.19: Rediscovery of GTEx brain eQTL maps.
(a)Fraction of GTEx brain eGenes that could be assessed in each of the considered contexts (cell type-conditions). (b) Fraction of GTEx brain eQTL that were replicated in this study (nominal p value < 0.05; fraction relative to the set of assessed genes from a). Astro: Astrocyte-like, DA: Midbrain dopaminergic neurons, Epen1: Ependymal-like, FPP: Floor Plate Progenitors, P_FPP: Proliferating FPP, Sert: Serotonergic-like neurons.

5.6 Colocalisation of eQTL with disease risk variants

The identified cell-type specific eQTL maps across different differentiation contexts provide an exciting opportunity to improve our understanding of human disease traits and their genetic risk factors identified by GWA studies. To systematically test for colocalisation events (**page 21**), we applied COLOC [175] to the summary statistics from 25 neurological traits[8], eQTL discovered in our study, as well as eQTL obtained from GTEx (v7) [150].

In total, we identified 1,284 eQTL in our study with evidence of colocalisation (PP4>0.5) with at least one disease trait, 597 of which were found only in our dataset. This corresponds to an additional >10% of colocalisation events of GWAS variants compared to eQTL across all GTEx tissues (5,028 across 48 tissues, **Fig. 5.20**). Notably, 401 (67%) of the colocalisations in our data were associated with eQTL detected in later differentiation stages (day 52) or upon stimulation (day 52 ROT).

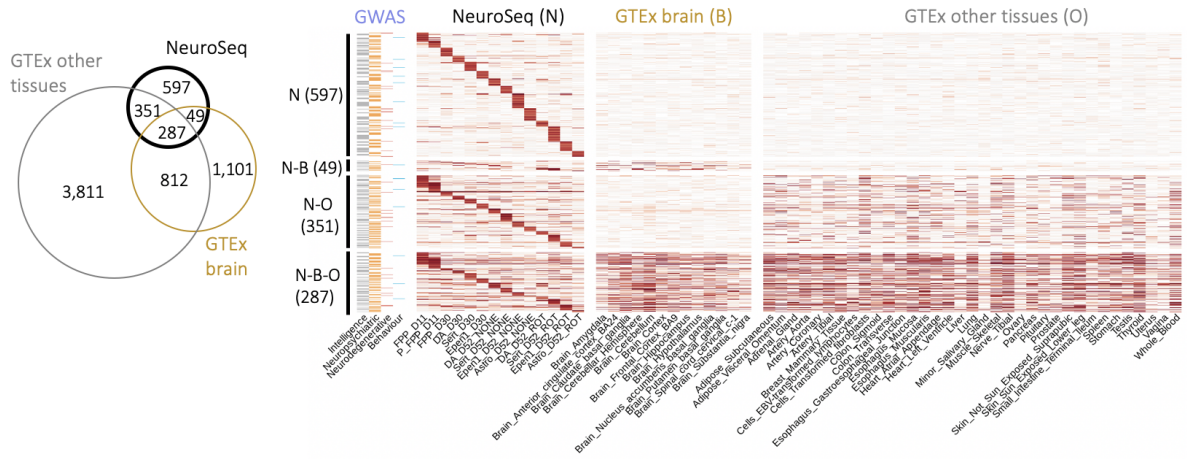


Fig. 5.20: Coloc overview.
Figure by Natsuhiko Kumasaka. Overview of colocalisation analysis between our eQTL maps and 25 neuro-related GWAS traits. (left) Venn diagram showing the numbers of colocalisation events overlapping between our study, GTEx brain and GTEx non brain tissues. (right) Heatmap showing the posterior probability of colocalisation (PP4 from COLOC [175]) for our eQTL that colocalised with one or more GWAS traits. N: Neuronal differentiation (this study), B: GTEx Brain tissues, O: Other GTEx tissues.

[8] including Parkinson's disease, Alzheimer's disease, schizophrenia, bipolar disorder, neuroticism, depression, and other behaviour and intelligence-related traits, see **Table A.4**.

Among the most interesting colocalisation events was an eQTL for *SFXN5*, a mitochondrial amino-acid transporter, which was specific to the rotenone-stimulated serotonergic neurons at day 52, and which colocalised with a Schizophrenia hit (PP4 = 0.78, **Fig. 5.21**). Exposure to rotenone is known to induce oxidative stress by inhibiting the mitochondrial respiratory chain complex I [538, 539]. We therefore speculate that the specific genetic signal observed for the mitochondrial gene *SFXN5* in serotonergic neurons is a possible factor modulating environmental stress response.

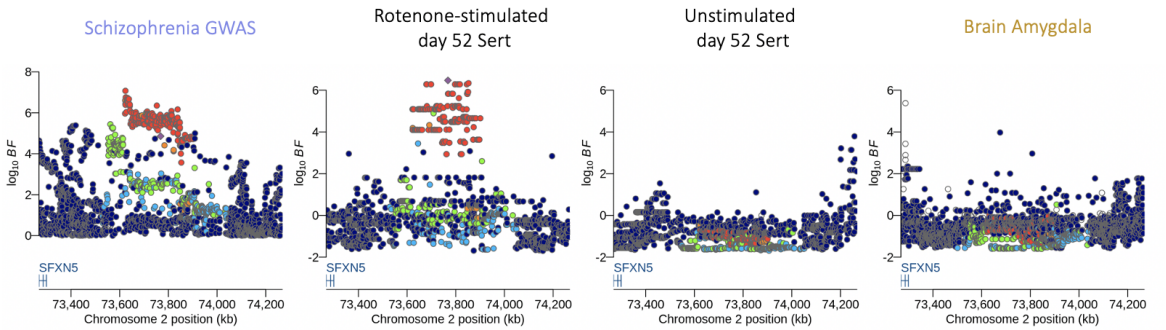


Fig. 5.21: A colocalisation event between a rotenone-specific eQTL and schizophrenia.
Figure by Natsuhiko Kumasaka. Locus zoom plots around the *SFXN5* gene. The Schizophrenia GWAS association (left) is colocalised with the eQTL in rotenone-stimulated serotonergic-like neurons at day 52 (second panel from the left). No colocalisation signal was found in unstimulated serotonergic-like neurons at day 52 (third panel from the left) or any other brain GTEx tissues as illustrated here with GTEx Brain Amygdala (rightmost panel). The lead variant is indicated with a purple diamond and other points were coloured according to the LD index (r^2 value) with the lead variant.

Another example that colocalised with a Schizophrenia GWAS variant was an eQTL for *FGFR1*, detected both in proliferating and non proliferating floor plate progenitors at day 11 (PP4 = 0.93 and 0.88 respectively, **Fig 5.22**) . Previous studies have shown that nuclear *FGFR1* plays a key role in regulating neural stem cell proliferation and central nervous system development, in part, by binding to the promoters of genes that control the transition from proliferation to cell differentiation [540]. Additionally, it was shown that altered *FGFR1* signaling was linked to the progression of the cortical malformation observed in schizophrenia [541].

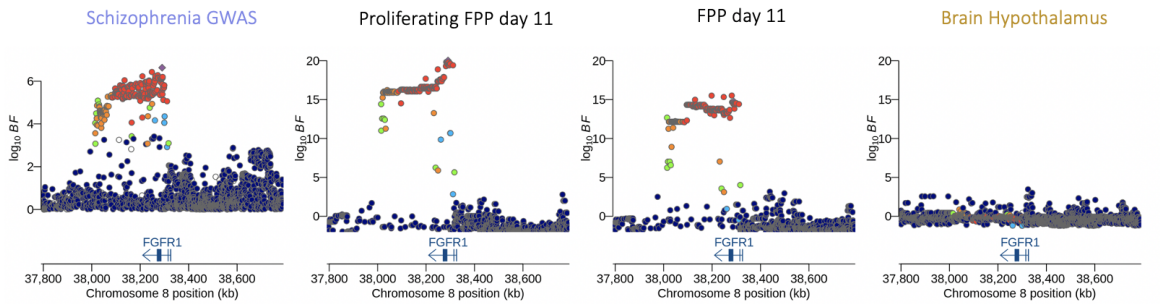


Fig. 5.22: A schizophrenia colocalisation event with a developmental eQTL.
Figure by Natsuhiko Kumasaka. A midbrain progenitor-specific eQTL for *FGFR1* associated with schizophrenia. We identified a colocalisation event with this eQTL in both proliferating (second panel from the left) and non-proliferating floor plate progenitors (third panel from the left) at day 11. No colocalisation was found in any other cell type from our study (not shown) nor in any brain GTEx tissues (shown with GTEx Brain Hypothalamus, rightmost panel).

These examples suggest that a combination of genetic and environmental factors during early development might contribute to schizophrenia pathology and illustrate how these data represent a valuable resource to understand the molecular basis of complex neurological diseases.

5.7 Discussion

The characterisation of the function of human trait-associated genetic variation requires large-scale studies, performed in disease-relevant cell types and states. Here, we demonstrate how human iPSCs can be efficiently profiled at scale throughout a 52 day-long differentiation to a midbrain neuronal cell fate.

First, we demonstrate high heterogeneity of cell types generated by this protocol (**Fig. 5.3**), and uncover a highly reproducible (**Fig. 5.6, 5.7**), cell line-intrinsic neuronal differentiation bias (**Fig. 5.5**). Next, we show how this bias can be robustly predicted using gene expression profiling at iPSC state (**Fig. 5.9**). This is an important step towards the optimised design of future large-scale iPSC experiments, where cell lines can be rationally selected a priori without the need for laborious testing of differentiation capacity.

Indeed, the 'quality' of human iPS cells has been carefully examined by several studies using both genetic and functional genomic data [280, 542, 543, 479], see also **section 1.2.5**. Despite these efforts, variation in differentiation potential between cell lines has been widely

acknowledged, yet poorly understood. To the best of our knowledge, the work we presented here is the first effort to systematically survey differentiation biases at the scale of an entire iPSC bank. To address this question, we leveraged the detailed phenotyping of cell lines in the HipSci bank. We excluded the cell type of origin hypothesis [544] in this instance since all HipSci lines were derived from fibroblasts. Moreover, we observed rather weak associations between neuronal differentiation efficiency and other biological factors, including sex and X chromosome activation status, which has been described as relevant for other differentiation lineages (i.e. endoderm lineage, see **Chapter 4** and [447]). In this work, we focused on the cell-line effects, that were especially prevalent (**Fig. 5.8**), but further investigation into the effects of sex and X chromosome inactivation (as well donor ethnicity, although that could not have been assessed in this study) is left for future work.

Our analysis indicates that the reduced production of neurons was best correlated with increased abundance of a specific subpopulation (cluster 2) of iPS cells that express the transcription factor *UTF1* and other genes at elevated levels (**Fig. 5.11**). Counter-intuitively, the proportion of cells in this subpopulation was positively correlated with the proportion of neuroblast cells on day 11, but lower fractions of dopaminergic and serotonergic-like neurons at later stages of differentiation. One possible explanation is that cell lines that commit earlier to a neuronal fate disproportionately lose neurons upon passaging at day 20 (**Fig. 5.1**, cells are passaged at day 20 as from original protocol [284]; this would explain the lack of clear differences between lines at day 11, with instead divergence appearing at the day 30 time point and then becoming even more evident by day 52).

In alternative, cluster 2 may preferentially differentiate to radial glial cells which are more prone to switch to an astroglial and ependymal differentiation programme [545]. In support of this hypothesis, we identified several genes that were upregulated in cluster 2, including *SIX3*, *MT1F* and *PITX2*, that are thought to play a role in astrocyte and ependymal cell biogenesis [546–548].

A second implication of our study is that, despite growth competition between cell lines, pooled experiments retain sufficient cells per donor to carry out genetic analysis, even following extended periods in culture. Although cells from different lines were pooled in similar numbers, we observed extensive variation throughout our differentiation experiments in the numbers of cells produced by different lines. For example, 50% of the cells we sequenced were produced by only 12% of lines. As we have demonstrated, this was an important effect to take into account in our eQTL analysis (**Fig. 5.12**). Future technical improvements, for

instance more precise matching of growth rates of cell lines within pools, or line selection based on predicted differentiation efficiency using markers in the iPSC state may further increase the utility of multiplexed iPSC differentiation experiments.

Finally, we could map eQTL across several neuronal cell types and in response to oxidative stress (**Fig. 5.15**). As we have seen, in order to understand the functional role of trait-associated variants it is crucial to perform genetic analyses in relevant cell types. Indeed, as a community, we have largely been unable to identify genetic variants that drive expression changes in narrowly-defined cell populations. This is due to our reliance on tissue-level data (e.g. GTEx), our incomplete knowledge of cell populations present in a tissue, or because rare cell populations do not provide enough substrate for common genomics assays.

Our analysis attempts to be a step towards investigating cell type-specific genetic effect, which are often masked in tissue-level assays. Indeed, despite a modest sample size, our study reveals a disproportionately large number of novel colocalisations between neurological traits and diseases and eQTL (**Fig. 5.20**) compared with GTEx tissues of equivalent sample size. For example, the number of novel trait/disease-eQTL colocalisations added by GTEx liver or cerebellar hemisphere (n=208, 215 respectively) are 80 and 107, respectively, compared to 597 in this study. A simple explanation for this result is that our experiment profiled expression states that are hard to capture using post-mortem tissue, including time points during neuronal differentiation and rotenone exposure.

Additionally, the single-cell resolution of our study enabled the detection of many eQTL that were specific to individual cell types, or could only be detected upon stimulation (**Fig. 5.16**). These signals, while present, are challenging to detect in bulk tissue because the relevant cell types are often rare. Combined, these results suggest that many 'missing', disease-relevant eQTL remain to be discovered using single cell profiling of both primary tissues and *in vitro* models.

6

Concluding remarks

The genomes of any two unrelated people are 99.9% identical. Yet, the 0.1% that differs is critical: it explains why individuals look different, and also why some are more predisposed than others to certain diseases. Thus, identifying DNA variants that are associated with complex disorders, and understanding the molecular mechanisms that mediate such associations, can lead in the future to better disease diagnosis, treatment and prevention. While GWAS (**section 1.1.7**) have identified thousands of associations between genetic variants and traits and diseases, the mechanisms involved have proven hard to disentangle. Associations between genetic variants and gene expression levels (i.e. eQTL, **section 1.1.8**) can help uncover such mechanisms, as gene expression often acts as an intermediate between DNA sequence and organismal phenotypes. Importantly, since these regulatory effects often arise in specific tissues or under specific stimuli [300], eQTL mapping studies need to be conducted in disease-relevant cell types. These are often hard to access, historically limiting studies to easily accessible tissues such as skin and blood [161, 114], or to cell lines [549]. More recently, the GTEx consortium released eQTL maps across over 50 human post-mortem tissues [151]. Whilst this represents a great resource, these tissues have been probed using bulk RNA-seq, making it difficult to isolate specific disease-relevant cell types, especially since these are often rare. Moreover, very little is known about the genetic regulation of gene expression at early stages of human development, most of which are impossible to access *in vivo*. Human iPSCs have proven to be a versatile *in vitro* model to study early development in a neatly controlled setup (**section 1.2.5**). Human iPSCs can be derived in a donor-specific manner, and, critically, they can be differentiated towards virtually any cell type of interest. Recently, large cohorts of human iPSCs across hundreds of individuals have enabled eQTL analyses in both iPSCs and a number of iPSC-derived cell types [294, 301].

In this thesis, I have shown that human iPSC technology combined with single cell expression readouts (which allow the isolation of cell types of interest), and pooling strategies (which increase throughput by enabling the differentiation of cells from several individuals in the same experiment), represent an excellent system to study the effect of common genetic variants on gene expression during cellular differentiation.

In particular, I have analysed two population-scale scRNA-seq datasets of differentiating human iPSCs along two different lineages, one toward definitive endoderm and the other to a midbrain neuronal fate. These represent important resources in their own right, as most current human scRNA-seq datasets only contain samples for a handful of genetically unique individuals.

Indeed, while the main objective of these studies was to identify eQTL across cell types and states during differentiation, one interesting side product of this work was the evaluation of differences in terms of differentiation outcome across several iPSC lines. In particular, full transcriptome information across hundreds of iPSC lines allowed us to assay these differences at a much larger scale than any previous study, to the best of my knowledge. In one case, we identified a set of genes whose expression at pluripotent stage can be used to predict neuronal differentiation efficiency, and which we could use to predict differentiation scores for the entire HipSci bank. This represents important progress toward understanding predictors of differentiation outcome and a useful resource for future studies using these lines.

Nevertheless, the main contributions of this thesis are in the context of eQTL mapping, specifically when using single cell RNA-seq profiles to measure gene expression. In particular, we systematically evaluated differences between mapping eQTL using bulk and single cell RNA-seq for a homogeneous cell population (human iPSCs). Additionally, we provide preliminary best-practice guidelines for single cell eQTL studies, in terms of normalisation strategies, aggregation approaches and covariate adjustment.

Furthermore, we mapped eQTL at different stages and cell types along human early development toward endoderm (mesendoderm and definitive endoderm) and along the midbrain neural lineage (floor plate progenitors, dopaminergic and serotonergic neurons, ependymal cells and astrocytes). To the best of our knowledge, these are the first eQTL maps at these stages of differentiation and thus represent an important resource for the genetics community.

Finally, work in this thesis provides insight into the importance of performing genetic analyses of gene expression in a context-specific manner, both by performing eQTL in discrete cell types and stimulation states, and by considering continuous axes of variation which modulate the genetic response.

6.1 Conclusions and discussion

The analyses we conducted have several important implications, in two main areas, which I discuss in the following sections. First, I use **section 6.1.1** to summarise and discuss our results assessing variability in the differentiation outcome of human iPSC lines and possible molecular predictors. Second, in **section 6.1.2** I discuss technical considerations and biological implications of mapping eQTL using single cell expression profiles.

6.1.1 Human iPSCs to model development and disease

In work presented in this thesis, we attempted to quantify the differentiation efficiency of different iPSC lines in two distinct protocols. In the first case, (described in **Chapter 4**) the protocol used was very short (three days) and very well understood, describing early stages of endoderm differentiation. Even so, we observed noticeable differences between lines in their ability to differentiate towards definitive endoderm. We identified a few tens of genes whose expression at iPSC stage was predictive of endoderm differentiation efficiency (**section 4.7**). These were mostly on chromosome X, confirming previous reports that the X chromosome reactivation in human iPSC lines may hamper their quality, especially with regards to their differentiation potential.

In the second study I describe in this thesis (in **Chapter 5**) the differentiation protocol used was much longer (52 days), and we differentiated significantly more lines (215). Here, we observed even more extreme differences across lines in their ability to generate neurons, with roughly one third of the lines preferentially producing non-neuronal cell types, namely ependymal- and astrocyte-like cells. Similar to previous reports [301], some batch effects were observed, but were significantly weaker than cell line effects. On the other hand, we identified an iPSC gene signature that was predictive of poor neuronal differentiation efficiency, finding around two thousand genes whose expression at pluripotent stage was significantly correlated (either positively or negatively) with a line's ability to generate neurons. We further hypothesised that this may be linked to a sub-population of iPSCs that exhibited differential expression of these genes. We speculate on possible mechanisms (**section 5.7**),

but argue that further validation would be needed to state anything conclusively. Lastly, we observe no correlation between the differentiation efficiencies defined in the two protocols, suggesting that a line's differentiation potential toward one lineage is independent, or perhaps even inversely correlated with that toward another.

Future work is required to gain a better understanding of the mechanisms and causes behind an iPSC line's differentiation potential. In particular, we note that since in both studies we only chose to differentiate one cell line per individual, we could not distinguish between cell line effects and donor effects. In future efforts, it will be important to include multiple lines per individual, to be able to effectively separate the two sources of variation. Moreover, all lines used here are skin-derived, thus the differences observed could not be driven by the somatic cell type of origin. A future area of study would involve investigating differences in the differentiation outcome of iPSC lines derived from different cell types, as well as across donor characteristics including sex, age and ethnicity. As highlighted on **page 40**, several human iPSC cohorts derived from different cell types, and for donors of different ethnicities and varying degrees of relatedness, are already available for research purposes, and could be used to address some of these aspects. Finally, in work presented here we did not have the appropriate sample size to detect genetic variants affecting differentiation efficiency. In the future, as protocols become more efficient and pooling strategies combined with single cell readouts become common-practice, it will be possible to perform *in vitro* differentiation experiments at increasingly large sample sizes. These studies will finally enable the exploration of the potential role of genetic variation on differentiation outcomes.

As more and more *in vitro* differentiation studies are conducted, across different lineages and iPSC cohorts, a systematic comparison of the outcomes can be performed, which will greatly improve our understanding of the processes involved. Indeed, such comparative studies will shed light on several unanswered questions. For example, is an iPSC line's inability to differentiate toward mature cell types simply an indication of its poor quality? And if so, is it simply not possible to use these lines for differentiation studies? Or, alternatively, are some lines more prone toward one cell fate and as a consequence less so to another? And to what extent is this dependent on the cell type of origin of those iPSCs? Importantly, is poor differentiation ability a characteristic of the cell line, or of the donor (genetic or otherwise)? This would have critical consequences, for example on the importance on deriving several iPSC lines from the same donor to maximise yield of 'good differentiating lines'. And if not, will it be harder to derive functional iPSC lines from some individuals compared to others? These and other questions remain to be investigated.

6.1.2 Bridging the genotype-phenotype gap

A large gap remains in our understanding of the functional mechanisms that link genotypes to phenotypes. eQTL studies can be used to fill some of this gap by identifying the putative regulatory role of common variants on gene expression. Indeed, when performed across tissues and contexts, eQTL maps can provide insights not only into which genes are regulated, but also in which cell types and under which conditions they are active.

The profiling of molecular traits, especially gene expression, at single cell resolution has represented a true revolution in the last ten years (**section 3.1.2**). In particular, experimental methods, and computational approaches to examine the resulting data, have become established in recent years, leading to the explosion of scRNA-seq data, with > 1,000 datasets published since 2009. Single cell expression profiling can now be deployed at population-scale and, combined with pooling strategies, permits the efficient quantification of cell-level expression across several individuals. Additionally, single cell transcriptomics can be used to estimate cell states and contexts at increased resolution [327]. For example, rare cell types and cells in different cell cycle phases can be identified unbiasedly within one experiment. Lastly, the use of single cell expression profiles allows the ordering of single cells along a continuous trajectory, without the need to discretise cells into distinct populations. Adding such cell-level context information to eQTL mapping provides one more layer to our understanding of the molecular consequences of common genetic variation, potentially making the genotype-phenotype gap one bit smaller.

In this thesis, I provide examples of how the single cell resolution of expression profiles can be leveraged to better understand the molecular machinery of gene regulation. First, single cell expression profiles can be used to unbiasedly identify pure cell populations, quantify expression within those, and then test for eQTL in such populations. Second, single cell profiles can be used to order cells along a differentiation trajectory, and used to identify dynamic eQTL, i.e. eQTL whose strength varies over time. Third, single cell transcriptomic data can be used to define other axes of variation, and thus context-specific eQTL can be identified across a plethora of cell states. From a technical standpoint, linear and linear mixed models (**Chapter 2**) are flexible frameworks that allow the user to efficiently test for associations whilst correcting for confounders and other sources of variation. Finally, colocalisation analysis between the identified eQTL and relevant GWAS trait connects the final dots to link the identified regulatory mechanisms to complex traits and diseases.

Historically, eQTL have been mapped using bulk RNA-seq profiles as a measure of expression level. The first implication of work described here is the feasibility of large-scale genetics using single cell RNA-seq data instead. Whilst this has to an extent been demonstrated before [160, 402], the small sample size of those studies only allowed the identification of tens or at most a few hundred eQTL. Moreover, these studies failed to recapitulate eQTL results obtained using bulk RNA-seq from equivalent tissues. Here, on the other hand, we identify thousands of eQTL across a range of cell types (**Tables A.2, 5.2**), and could re-discover a larger portion of bulk-discovered eQTL (**Fig. 3.7, 5.19**). Moreover, we demonstrated feasibility of single cell eQTL mapping using both plate-based (SmartSeq2, **Chapters 3, 4**) and droplet-based (10X Genomics, **Chapter 5**) scRNA-seq data.

To systematically compare the performance of using scRNA-seq as opposed to bulk RNA-seq to map eQTL, in **Chapter 3** we selected human iPSCs as a homogeneous cell type, and compared results when mapping iPSC eQTL using a common set of samples (**Fig. 3.7**). This analysis revealed an increased number of discovered eQTL when using bulk RNA-seq profiles in this well defined, pure cell population, probably due to decreased noise in expression estimates. On the other hand, we appreciated the power of the single cell transcriptomics to isolate several cell types within more heterogeneous populations, quantify expression and map eQTL within them, without the need for any gating or other experimental techniques to separate cell populations (**Fig. 4.11, 5.16**).

In addition, we provide here the first hints towards the establishment of a best-practice workflow to maximise yield of single cell eQTL studies (**section 3.8**), identifying the mean (after single cell-specific normalisation) as the optimal aggregation method, and principal component analysis as the preferable approach to capture global expression trends which should be included in the model as covariates. From a methodological perspective, linear mixed models were confirmed as the appropriate tool to identify genetic associations, given their ability to deal with confounding effects (**section 2.2.4**). In particular, LMMs can control for effects due to population structure, including replicate measurements across donors (for example across multiple differentiation experiments, **Chapters 3** and **4**). In addition, LMMs enabled the introduction of a variance term to account for number of cells across individuals, which varied widely thus rendering the expression estimates less precise (**Chapter 5**); this expedient resulted in a great boost in the number of eQTL discoveries. In future work, models which enable the incorporation of multiple random effect terms, to effectively correct for several confounders simultaneously, should be developed.

The availability of eQTL maps across cell types and stages provided the opportunity to assess the amount of eQTL signal sharing both within our studies and in comparison with existing maps. This is a notoriously complicated task, because different eQTL studies may differ in the technology used to measure expression, in the number of genes expressed and in sample size, which is in general fairly low. Here, we used two separate approaches to tackle this issue. On the one hand, we used p value thresholding to identify cell type-specific eQTL (eQTL that could only be detected in one of the cell populations considered within our study, **Fig. 4.12, 5.16**), and assess the number of eQTL identified in our study that were not discovered in eQTL maps of primary tissues (i.e. from GTEx) and viceversa (**Fig. 5.19**). On the other hand, we used a recently proposed method (MASHR [536]) to quantify genome-wide sharing across eQTL maps (**Fig. 5.18**). These approaches are complementary, representing the two ends of the spectrum: the first approach may miss signals that only just do not reach the (arbitrary) significance threshold used, whilst the second may overestimate sharing by only considering gene-SNP pairs assessed across all conditions included. To partially overcome these issues, methods exist that consider multiple eQTL datasets jointly [550, 551]. However, such methods are currently computationally too demanding for large-scale scRNA-seq data.

Next, in **Chapter 4**, we added the temporal axis, by identifying dynamic eQTL, i.e. eQTL whose strength is modulated by developmental time. This extends similar work from [470, 302], to single cell-resolved data. Indeed, in this study cells were collected at very close time points, which combined with varying differentiation rates across both cells and lines resulted in a continuous differentiation trajectory. Importantly, we observed that changes in genetic effects over time did not merely reflect changes in overall expression (**Fig. 4.14**). Moreover, we found that dynamic eQTL were enriched for epigenetic marks consistent with promoter and enhancer regions. We next used the same approach, building on allele-specific expression (similar to [491] for GxE), to test for eQTL effects that are modulated by alternative cell states, including cell cycle phase and metabolic state (**Fig. 4.16, 4.3**). This type of analysis is similar to previous work to identify 'interaction eQTL' [157, 160].

Finally, in **Chapter 5**, we assessed disease-relevance of our identified associations, by performing colocalisation analysis between eQTL maps from our neuronal cell populations and GWAS for neurological traits. Here, we uncover several colocalisation events that had not been previously identified (**Fig. 5.20**), highlighting once again the importance of studying the molecular consequences of genetic variation in relevant cell types, especially when investigating the genetic basis of disease. Indeed, some of these examples provide insight into the genetic underpinning of neurological diseases, including schizophrenia.

Overall, the work in this thesis demonstrates the feasibility of eQTL mapping using single cell expression profiles and the importance of modelling context-specific eQTL effects across cellular types and states. The methods used build on the linear mixed model framework and are extremely flexible, as demonstrated by their application across technologies and designs. Whilst extremely useful and efficient, these models assume normality of the residual phenotypes, an assumption that is often violated, as discussed (**section 2.3.4**). In particular, scRNA-seq data has has been described to follow a Poisson or a negative binomial [552–554] distribution. Future work should include the evaluation of the feasibility of integrating non-Gaussian likelihoods in the models to map eQTL using scRNA-seq data.

Moreover, in this thesis I have focused on the study of context-specific eQTL by either first discretising cells into populations and mapping eQTL in each, or by considering interactions with one single continuous cell state (or at most two, **Fig. 4.17**). In the future, it will be important to develop methods to jointly test for context-specificity of eQTL across several (continuous and discrete) cell states simultaneously. For example, a recently proposed method, Struct-LMM [555] allows the assessment of GxE interactions using larger numbers of conditions. While originally proposed in the context of population studies, the same principles could be adopted here, where one could map sc-eQTL that vary jointly across up to hundreds different cell states and types. These advanced models will enable us to leverage the rich information from single cell-resolved, transcriptome-wide population-scale datasets, to further improve our understanding of the genetic architecture of traits.

6.2 Outlook and future directions

6.2.1 More complex and realistic *in vitro* models

Work presented in this thesis demonstrated how iPSC differentiation combined with multiplexed experimental designs and single cell RNA-seq profiling unlocks population-level studies in increasingly complex, dynamic and biologically realistic cellular models. We anticipate that, in the future, uses of this model system will focus on experimental settings that are challenging or impossible with primary cells. For example, these may include single cell resolution sampling along longer differentiation times to more complex differentiation trajectories, such as cell organoids, or involve large panels of disease relevant-stimuli and drug exposures. These future efforts will greatly contribute to our understanding of the common genetic basis of complex disorders, and facilitate the development of iPSC-based approaches for modelling, and even eventually treating, these diseases.

6.2.2 More population-scale scRNA-seq datasets

The work presented in this thesis has focused on applications in iPSCs and iPSC-derived cells, however we note that the same technologies and methodologies can be applied across a variety of biological systems. For instance, scRNA-seq PBMC data across multiple conditions from a growing number of individuals (currently approximately 1,600), will be available as part of the single-cell eQTLGen (sc-eQTLGen) consortium, whose manifesto was recently published [446]. Similarly, through the UK Biobank [556], one of the largest and most deeply phenotyped cohorts of individuals in the world, blood samples as well as key biomarkers for circa 500,000 people are stored and available for the research community. Performing scRNA-seq on blood cells from these many (and well characterised) samples will provide an invaluable resource to study the effect of genetic variants across cell types and contexts. Last but not least, the human cell atlas (HCA) project [557], whose mission is "to create comprehensive reference maps of all human cells", will likely in the future be collecting samples from several donors across all human tissues, to evaluate differences across genetic backgrounds and disease states. It will be critical, when these data become available, to have robust and efficient statistical models to make use of this wealth of data.

6.2.3 Alternative single cell technologies

Finally, here we have focused on single cell transcriptomic data, which is the most well-established of the single cell sequencing technologies. Yet more recently, several alternative molecular traits have been assayed at single cell resolution, including chromatin accessibility [558, 559], DNA methylation [560–562], histone modifications [563] and chromatin 3D organisation [564]. Novel technologies even allow multiple molecular layers to be probed in parallel from the same individual cells [565–567]. The LMM-based models used here can readily be adapted to map alternative single cell molecular QTL (e.g. sc-mQTL, sc-caQTL, etc.), which could provide a much richer understanding of the molecular machinery associated with genetic regulation. Using multi-omics data, similar models can further be used to study the interplay between molecular layers (e.g. effects of methylation or accessibility on expression). Finally, standard eQTL assess the effect of naturally occurring genetic variation on gene expression. However, recent pioneering studies have used the induction of CRISPR/Cas9 perturbations, followed by scRNA-seq to identify the effect of such induced variation on gene expression [568]. Models describe here can naturally be extended for the identification of cell type- and context-specific 'crisprQTL' as well.

6.3 Genetic mapping at single cell resolution

The use of single cell omics, particularly gene expression, has revolutionised our understanding of cellular variability in several biological systems. These technologies can now be deployed across hundreds of individuals, enabling the study of the effects of common genetic variants on gene expression level (i.e. by mapping eQTL), which was once only assessed using bulk RNA-seq. In particular, the single-cell resolution can help to uncover eQTL that are only active in rare cell populations, or that change dynamically along cellular states. Taken together, the work in this thesis demonstrates the utility of mapping eQTL using single cell expression data, to reveal the function of genetic variation across cellular types and states. The models described here, combined with the increasing availability of population-scale single cell expression studies, and in the future extended to include multiple molecular layers, have the potential to greatly advance our understanding of the complex machinery that links genotype to phenotype.

References

[1] Charles Darwin's. On the origin of species. *published on*, 24, 1859.

[2] Gregor Mendel. Experiments in plant hybridization (1865). *Verhandlungen des naturforschenden Vereins Brünn) Available online*, 1996.

[3] William Bateson and Gregor Mendel. *Mendel's principles of heredity*. Courier Corporation, 2013.

[4] Walter S Sutton. The chromosomes in heredity. *The Biological Bulletin*, 4(5):231–250, 1903.

[5] Thomas H Morgan. Random segregation versus coupling in mendelian inheritance. *Science*, 34(873):384–384, 1911.

[6] Alfred H Sturtevant. The linear arrangement of six sex-linked factors in drosophila, as shown by their mode of association. *Journal of experimental zoology*, 14(1):43–59, 1913.

[7] Wilhelm Johannsen. The genotype conception of heredity. *The American Naturalist*, 45(531):129–159, 1911.

[8] Thomas Hunt Morgan, Alfred Henry Sturtevant, Calvin Blackman Bridges, and Hermann Joseph Muller. *The mechanism of Mendelian heredity*, volume 86. H. Holt, 1915.

[9] Nobel Media AB. *The Nobel Prize in Physiology or Medicine – 1933*. 1933.

[10] Erwin Schrödinger. *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, 1944.

[11] Siddartha Mukherjee. *The gene: an intimate history*. Bodley Head, 2016.

[12] Bruce Alberts. Molecular biology of the cell. 2018.

[13] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *The Journal of experimental medicine*, 79(2):137–158, 1944.

[14] Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[15] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[16] Nobel Media AB. *The Nobel Prize in Physiology or Medicine – 1962*. 1962.

[17] Francis Galton. *Hereditary genius: An inquiry into its laws and consequences*. D. Appleton, 1870.

[18] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[19] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[20] Ronald A Fisher. Xv.—the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919.

[21] Nicholas H Barton, Alison M Etheridge, and Amandine Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical population biology*, 118:50–73, 2017.

[22] Ronald A Fisher. On the interpretation of χ 2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[23] Ronald Aylmer Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

[24] Ronald Aylmer Fisher et al. 036: On a distribution yielding the error functions of several well known statistics. 1924.

[25] Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922.

[26] Anders Hald et al. On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science*, 14(2):214–222, 1999.

[27] Ronald A Fisher. The systematic location of genes by means of crossover observations. *The American Naturalist*, 56(646):406–411, 1922.

[28] Charles R Henderson. Estimation of genetic parameters. In *Biometrics*, volume 6, pages 186–187. International Biometric Soc 1441 I ST, NW, SUITE 700, WASHINGTON, DC 20005-2210, 1950.

[29] National Human Genome Research Institute. *Genetic Timeline*. 2003.

[30] Charles Yanofsky. Establishing the triplet nature of the genetic code. *Cell*, 128(5):815–818, 2007.

[31] George W Beadle and Edward L Tatum. Genetic control of biochemical reactions in neurospora. *Proceedings of the National Academy of Sciences of the United States of America*, 27(11):499, 1941.

[32] George E Palade. A small particulate component of the cytoplasm. *The Journal of biophysical and biochemical cytology*, 1(1):59, 1955.

[33] Francis HC Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[34] Sydney Brenner, François Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190(4776):576–581, 1961.

[35] Marshall W Nirenberg and J Heinrich Matthaei. The dependence of cell-free protein synthesis in e. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences*, 47(10):1588–1602, 1961.

[36] Francis HC Crick, Leslie Barnett, Sydney Brenner, and Richard J Watts-Tobin. General nature of the genetic code for proteins. *Nature*, 192(4809):1227–1232, 1961.

[37] J Heinrich Matthaei, Oliver W Jones, Robert G Martin, and Marshall W Nirenberg. Characteristics and composition of rna coding units. *Proceedings of the National Academy of Sciences of the United States of America*, 48(4):666, 1962.

[38] Fred Sanger and Alan R Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.

[39] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977.

[40] Allan M Maxam and Walter Gilbert. A new method for sequencing dna. *Proceedings of the National Academy of Sciences*, 74(2):560–564, 1977.

[41] Arthur Kornberg, IR Lehman, Maurice J Bessman, and ES Simms. Enzymic synthesis of deoxyribonucleic acid. *Biochimica et biophysica acta*, 21(1):197–198, 1956.

[42] Rodger Staden. A strategy of dna sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979.

[43] Stephen Anderson. Shotgun dna sequencing using cloned dnase i-generated fragments. *Nucleic acids research*, 9(13):3015–3027, 1981.

[44] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, 269(5223):496–512, 1995.

[45] Claire M Fraser, Jeannine D Gocayne, Owen White, Mark D Adams, Rebecca A Clayton, Robert D Fleischmann, Carol J Bult, Anthony R Kerlavage, Granger Sutton, Jenny M Kelley, et al. The minimal gene complement of mycoplasma genitalium. *Science*, 270(5235):397–404, 1995.

[46] André Goffeau, Bart G Barrell, Howard Bussey, Ronald W Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, Jörg D Hoheisel, Claude Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[47] Frederick R Blattner, Guy Plunkett, Craig A Bloch, Nicole T Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D Glasner, Christopher K Rode, George F Mayhew, et al. The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462, 1997.

[48] C. elegans Sequencing Consortium*. Genome sequence of the nematode c. elegans: a platform for investigating biology. *Science*, 282(5396):2012–2018, 1998.

[49] STea Cole, R Brosch, J Parkhill, T Garnier, C Churcher, D Harris, SV Gordon, K Eiglmeier, S Gas, CE 3rd Barry, et al. Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.

[50] Mark D Adams, Susan E Celniker, Robert A Holt, Cheryl A Evans, Jeannine D Gocayne, Peter G Amanatides, Steven E Scherer, Peter W Li, Roger A Hoskins, Richard F Galle, et al. The genome sequence of drosophila melanogaster. *Science*, 287(5461):2185–2195, 2000.

[51] Samir Kaul, Hean L Koo, Jennifer Jenkins, Michael Rizzo, Timothy Rooney, Luke J Tallon, Tamara Feldblyum, William Nierman, Maria Ines Benito, Xiaoying Lin, et al. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *nature*, 408(6814):796–815, 2000.

[52] I Dunham, AR Hunt, JE Collins, R Bruskiewich, DM Beare, M Clamp, LJ Smink, R Ainscough, JP Almeida, A Babbage, et al. The dna sequence of human chromosome 22. *Nature*, 402(6761):489–495, 1999.

[53] M Hattori, A Fujiyama, TD Taylor, H Watanabe, T Yada, H-S Park, A Toyoda, K Ishii, Y Totoki, D-K Choi, et al. The dna sequence of human chromosome 21. *Nature*, 405(6784):311–319, 2000.

[54] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001.

[55] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[56] Robert H Waterston, Kerstin Lindblad-Toh, Ewan Birney, Jane Rogers, Josep F Abril, Pankaj Agarwal, Richa Agarwala, Rachel Ainscough, Marina Alexandersson, Peter An, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[57] ArchibaldE Garrod. The incidence of alkaptonuria: a study in chemical individuality. *The Lancet*, 160(4137):1616–1620, 1902.

[58] Vernon M Ingram et al. The hemoglobins in genetics and evolution. *The hemoglobins in genetics and evolution.*, 1963.

[59] David Botstein, Raymond L White, Mark Skolnick, and Ronald W Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3):314, 1980.

[60] M Dawn Teare and Jennifer H Barrett. Genetic linkage studies. *The Lancet*, 366(9490):1036–1044, 2005.

[61] James F Gusella, Nancy S Wexler, P Michael Conneally, Susan L Naylor, Mary Anne Anderson, Rudolph E Tanzi, Paul C Watkins, Kathleen Ottina, Margaret R Wallace, Alan Y Sakaguchi, et al. A polymorphic dna marker genetically linked to huntington's disease. *Nature*, 306(5940):234–238, 1983.

[62] John R Riordan, Johanna M Rommens, Bat-sheva Kerem, Noa Alon, Richard Rozmahel, Zbyszko Grzelczak, Julian Zielenski, Si Lok, Natasa Plavsic, Jia-Ling Chou, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary dna. *Science*, 245(4922):1066–1073, 1989.

[63] David Botstein and Neil Risch. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature genetics*, 33(3):228–237, 2003.

[64] Lon R Cardon and John I Bell. Association study designs for complex diseases. *Nature Reviews Genetics*, 2(2):91–99, 2001.

[65] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[66] Inga Prokopenko, Claudia Langenberg, Jose C Florez, Richa Saxena, Nicole Soranzo, Gudmar Thorleifsson, Ruth JF Loos, Alisa K Manning, Anne U Jackson, Yurii Aulchenko, et al. Variants in mtnr1b influence fasting glucose levels. *Nature genetics*, 41(1):77–81, 2009.

[67] Sekar Kathiresan, Olle Melander, Candace Guiducci, Aarti Surti, Noël P Burtt, Mark J Rieder, Gregory M Cooper, Charlotta Roos, Benjamin F Voight, Aki S Havulinna, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature genetics*, 40(2):189–197, 2008.

[68] Eleftheria Zeggini, Laura J Scott, Richa Saxena, Benjamin F Voight, Jonathan L Marchini, Tianle Hu, Paul IW de Bakker, Gonçalo R Abecasis, Peter Almgren, Gitte Andersen, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature genetics*, 40(5):638–645, 2008.

[69] John B Harley, Marta E Alarcón-Riquelme, Lindsey A Criswell, Chaim O Jacob, Robert P Kimberly, Kathy L Moser, Betty P Tsao, Timothy J Vyse, and Carl D Langefeld. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in itgam, pxk, kiaa1542 and other loci. *Nature genetics*, 40(2):204–210, 2008.

[70] Joan E Bailey-Wilson and Alexander F Wilson. Linkage analysis in the next-generation sequencing era. *Human heredity*, 72(4):228–236, 2011.

[71] Yoshio Miki, Jeff Swensen, Donna Shattuck-Eidens, P Andrew Futreal, Keith Harshman, Sean Tavtigian, Qingyun Liu, Charles Cochran, L Michelle Bennett, Wei Ding, et al. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

[72] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Comput Biol*, 8(12):e1002822, 2012.

[73] Janine Altmüller, Lyle J Palmer, Guido Fischer, Hagen Scherb, and Matthias Wjst. Genomewide scans of complex human diseases: true linkage is hard to find. *The American Journal of Human Genetics*, 69(5):936–950, 2001.

[74] David E Reich and Eric S Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001.

[75] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[76] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews genetics*, 9(5):356–369, 2008.

[77] LB Jorde. Linkage disequilibrium and the search for complex disease genes. *Genome research*, 10(10):1435–1444, 2000.

[78] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[79] Heather J Cordell and David G Clayton. Genetic association studies. *The Lancet*, 366(9491):1121–1131, 2005.

[80] W.J. Clinton. *Remarks made by the President on the Completion of the First Survey of the Entire Human Genome Project.* The White House Office of the Press Secretary., 2000.

[81] Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, 2011.

[82] J Craig Venter, Mark D Adams, Granger G Sutton, Anthony R Kerlavage, Hamilton O Smith, and Michael Hunkapiller. Shotgun sequencing of the human genome, 1998.

[83] Jeremy Schmutz, Jeremy Wheeler, Jane Grimwood, Mark Dickson, Joan Yang, Chenier Caoile, Eva Bajorek, Stacey Black, Yee Man Chan, Mirian Denys, et al. Quality assessment of the human genome sequence. *Nature*, 429(6990):365–368, 2004.

[84] M Hattori. Finishing the euchromatic sequence of the human genome. *Tanpakushitsu kakusan koso. Protein, nucleic acid, enzyme*, 50(2):162–168, 2005.

[85] Mihaela Pertea, Alaina Shumate, Geo Pertea, Ales Varabyou, Yu-Chi Chang, Anil K Madugundu, Akhilesh Pandey, and Steven L Salzberg. Thousands of large-scale rna sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *BioRxiv*, page 332825, 2018.

[86] Nan M Laird and Christoph Lange. *The fundamentals of modern statistical genetics*. Springer Science & Business Media, 2010.

[87] David G Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998.

[88] Wen-Hsiung Li and Lori A Sadler. Low nucleotide diversity in man. *Genetics*, 129(2):513–523, 1991.

[89] Michele Cargill, David Altshuler, James Ireland, Pamela Sklar, Kristin Ardlie, Nila Patil, Charles R Lane, Esther P Lim, Nilesh Kalyanaraman, James Nemesh, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature genetics*, 22(3):231–238, 1999.

[90] International HapMap Consortium et al. The international hapmap project. *Nature*, 426(6968):789, 2003.

[91] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299, 2005.

[92] International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851, 2007.

[93] International HapMap 3 Consortium et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52, 2010.

[94] Emma Meaburn, Lee M Butcher, Leonard C Schalkwyk, and Robert Plomin. Genotyping pooled dna using 100k snp microarrays: a step towards genomewide association scans. *Nucleic acids research*, 34(4):e28–e28, 2006.

[95] Arnold Oliphant, David L Barker, John R Stuelpnagel, and Mark S Chee. Beadarray™ technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques*, 32(sup):S56–S61, 2002.

[96] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[97] Itsik Pe'er, Paul IW de Bakker, Julian Maller, Roman Yelensky, David Altshuler, and Mark J Daly. Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature genetics*, 38(6):663–667, 2006.

[98] Sarah P Otto and Thomas Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3(4):252–261, 2002.

[99] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650–654, 2002.

[100] Robert J Klein, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, Susan T Mayne, et al. Complement factor h polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

[101] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[102] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, 42(D1):D1001–D1006, 2014.

[103] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1):D896–D901, 2017.

[104] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[105] Michael D Gallagher and Alice S Chen-Plotkin. The post-gwas era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[106] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.

[107] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[108] Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[109] Eric E Schadt, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, Guy Cavet, et al. Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302, 2003.

[110] Jacob F Degner, Athma A Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J Gaffney, Joseph K Pickrell, Sherryl De Leon, Katelyn Michelini, Noah Lewellen, Gregory E Crawford, et al. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature*, 482(7385):390–394, 2012.

[111] Tom R Gaunt, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L McArdle, Karen Ho, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome biology*, 17(1):1–14, 2016.

[112] Fabian Grubert, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, Aleksandra Pekowska, et al. Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell*, 162(5):1051–1065, 2015.

[113] David Melzer, John RB Perry, Dena Hernandez, Anna-Maria Corsi, Kara Stevens, Ian Rafferty, Fulvio Lauretani, Anna Murray, J Raphael Gibbs, Giuseppe Paolisso, et al. A genome-wide association study identifies protein quantitative trait loci (pqtls). *PLoS Genet*, 4(5):e1000072, 2008.

[114] Harm-Jan Westra and Lude Franke. From genome to function by studying eqtls. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1842(10):1896–1902, 2014.

[115] Ritsert C Jansen and Jan-Peter Nap. Genetical genomics: the added value from segregation. *TRENDS in Genetics*, 17(7):388–391, 2001.

[116] Matthew V Rockman and Leonid Kruglyak. Genetics of global gene expression. *Nature Reviews Genetics*, 7(11):862–872, 2006.

[117] Rachel B Brem, Gaël Yvert, Rebecca Clinton, and Leonid Kruglyak. Genetic dissection of transcriptional regulation in budding yeast. *Science*, 296(5568):752–755, 2002.

[118] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, et al. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, 2007.

[119] Anna L Dixon, Liming Liang, Miriam F Moffatt, Wei Chen, Simon Heath, Kenny CC Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, et al. A genome-wide association study of global gene expression. *Nature genetics*, 39(10):1202–1207, 2007.

[120] Isis Ricaño-Ponce and Cisca Wijmenga. Mapping of immune-mediated disease genes. *Annual review of genomics and human genetics*, 14:325–353, 2013.

[121] Ian Dunham, Ewan Birney, Bryan R Lajoie, Amartya Sanyal, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Troy W Whitfield, Jiali Zhuang, et al. An integrated encyclopedia of dna elements in the human genome. 2012.

[122] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory dna. *Science*, 337(6099):1190–1195, 2012.

[123] Andreas PM Weber. Discovering new biology through sequencing of rna. *Plant physiology*, 169(3):1524–1531, 2015.

[124] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, 2010.

[125] Joseph K Pickrell, John C Marioni, Athma A Pai, Jacob F Degner, Barbara E Engelhardt, Everlyne Nkadori, Jean-Baptiste Veyrieras, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*, 464(7289):768–772, 2010.

[126] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter Ac't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G Ferreira, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[127] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[128] Lu Chen, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. *Cell*, 167(5):1398–1414, 2016.

[129] Marc Jan Bonder, Craig Smail, Michael J Gloudemans, Laure Frésard, David Jakubosky, Matteo D'Antonio, Xin Li, Nicole M Ferraro, Ivan Carcamo-Orive, Bogdan Mirauta, et al. Systematic assessment of regulatory effects of human disease variants in pluripotent cells. *bioRxiv*, page 784967, 2019.

[130] C Joel McManus, Joseph D Coolon, Michael O Duff, Jodi Eipper-Mains, Brenton R Graveley, and Patricia J Wittkopp. Regulatory divergence in drosophila revealed by mrna-seq. *Genome research*, 20(6):816–825, 2010.

[131] Angela Goncalves, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T Odom, and John C Marioni. Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome research*, 22(12):2376–2384, 2012.

[132] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013.

[133] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44, 2009.

[134] Amanda J Myers, J Raphael Gibbs, Jennifer A Webster, Kristen Rohrer, Alice Zhao, Lauren Marlowe, Mona Kaleem, Doris Leung, Leslie Bryden, Priti Nath, et al. A survey of genetic human cortical gene expression. *Nature genetics*, 39(12):1494–1499, 2007.

[135] Joseph E Powell, Anjali K Henders, Allan F McRae, Jinhee Kim, Gibran Hemani, Nicholas G Martin, Emmanouil T Dermitzakis, Greg Gibson, Grant W Montgomery, and Peter M Visscher. Congruence of additive and non-additive effects on gene expression estimated from pedigree and snp data. *PLoS Genet*, 9(5):e1003502, 2013.

[136] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, et al. Unraveling the polygenic architecture of complex traits using blood eqtl metaanalysis. *BioRxiv*, page 447367, 2018.

[137] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt, and Mark Lathrop. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics*, 10(3):184–194, 2009.

[138] Elin Grundberg, Kerrin S Small, Åsa K Hedman, Alexandra C Nica, Alfonso Buil, Sarah Keildson, Jordana T Bell, Tsun-Po Yang, Eshwar Meduri, Amy Barrett, et al. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics*, 44(10):1084–1089, 2012.

[139] Hua Zhong, John Beaulaurier, Pek Yee Lum, Cliona Molony, Xia Yang, Douglas J MacNeil, Drew T Weingarth, Bin Zhang, Danielle Greenawalt, Radu Dobrin, et al. Liver and adipose expression associated snps are enriched for association to type 2 diabetes. *PLoS Genet*, 6(5):e1000932, 2010.

[140] Jingyuan Fu, Marcel GM Wolfs, Patrick Deelen, Harm-Jan Westra, Rudolf SN Fehrmann, Gerard J Te Meerman, Wim A Buurman, Sander SM Rensen, Harry JM Groen, Rinse K Weersma, et al. Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet*, 8(1):e1002431, 2012.

[141] Alexandra C Nica, Leopold Parts, Daniel Glass, James Nisbet, Amy Barrett, Magdalena Sekowska, Mary Travers, Simon Potter, Elin Grundberg, Kerrin Small, et al. The architecture of gene regulatory variation across multiple human tissues: the muther study. *PLoS Genet*, 7(2):e1002003, 2011.

[142] Antigone S Dimas, Samuel Deutsch, Barbara E Stranger, Stephen B Montgomery, Christelle Borel, Homa Attar-Cohen, Catherine Ingle, Claude Beazley, Maria Gutierrez Arcelus, Magdalena Sekowska, et al. Common regulatory variation impacts gene expression in a cell type–dependent manner. *Science*, 325(5945):1246–1250, 2009.

[143] Benjamin P Fairfax, Seiko Makino, Jayachandran Radhakrishnan, Katharine Plant, Stephen Leslie, Alexander Dilthey, Peter Ellis, Cordelia Langford, Fredrik O Vannberg, and Julian C Knight. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of hla alleles. *Nature genetics*, 44(5):502, 2012.

[144] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, 8(4):e1002639, 2012.

[145] Eric E Schadt, Cliona Molony, Eugene Chudin, Ke Hao, Xia Yang, Pek Y Lum, Andrew Kasarskis, Bin Zhang, Susanna Wang, Christine Suver, et al. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, 2008.

[146] Ke Hao, Yohan Bossé, David C Nickle, Peter D Paré, Dirkje S Postma, Michel Laviolette, Andrew Sandford, Tillie L Hackett, Denise Daley, James C Hogg, et al. Lung eqtls to help reveal the molecular underpinnings of asthma. *PLoS Genet*, 8(11):e1003029, 2012.

[147] Federico Innocenti, Gregory M Cooper, Ian B Stanaway, Eric R Gamazon, Joshua D Smith, Snezana Mirkov, Jacqueline Ramirez, Wanqing Liu, Yvonne S Lin, Cliona Moloney, et al. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS Genet*, 7(5):e1002078, 2011.

[148] Andrew I Su, Michael P Cooke, Keith A Ching, Yaron Hakak, John R Walker, Tim Wiltshire, Anthony P Orth, Raquel G Vega, Lisa M Sapinoso, Aziz Moqrich, et al. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences*, 99(7):4465–4470, 2002.

[149] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6):580–585, 2013.

[150] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

[151] François Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, Yanyu Liang, Meritxell Oliva, et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *BioRxiv*, page 787903, 2019.

[152] Silva Kasela, Kai Kisand, Liina Tserel, Epp Kaleviste, Anu Remm, Krista Fischer, Tõnu Esko, Harm-Jan Westra, Benjamin P Fairfax, Seiko Makino, et al. Pathogenic implications for autoimmune mechanisms derived by comparative eqtl analysis of cd4+ versus cd8+ t cells. *PLoS genetics*, 13(3):e1006643, 2017.

[153] Vivek Naranbhai, Benjamin P Fairfax, Seiko Makino, Peter Humburg, Daniel Wong, Esther Ng, Adrian VS Hill, and Julian C Knight. Genomic modulators of gene expression in human neutrophils. *Nature communications*, 6(1):1–13, 2015.

[154] Harm-Jan Westra, Danny Arends, Tõnu Esko, Marjolein J Peters, Claudia Schurmann, Katharina Schramm, Johannes Kettunen, Hanieh Yaghootkar, Benjamin P Fairfax, Anand Kumar Andiappan, et al. Cell specific eqtl analysis without sorting cells. *PLoS genetics*, 11(5):e1005223, 2015.

[155] David Venet, F Pecasse, Carine Maenhaut, and Hugues Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl_1):S279–S287, 2001.

[156] Tongwu Zhang, Jiyeon Choi, Michael A Kovacs, Jianxin Shi, Mai Xu, Alisa M Goldstein, Adam J Trower, D Timothy Bishop, Mark M Iles, David L Duffy, et al. Cell-type–specific eqtl of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome research*, 28(11):1621–1635, 2018.

[157] Daria V Zhernakova, Patrick Deelen, Martijn Vermaat, Maarten Van Iterson, Michiel Van Galen, Wibowo Arindrarto, Peter Van't Hof, Hailiang Mei, Freerk Van Dijk, Harm-Jan Westra, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics*, 49(1):139–145, 2017.

[158] Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, et al. Single-cell rna-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, 356(6335), 2017.

[159] Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752, 2013.

[160] Monique GP van der Wijst, Harm Brugge, Dylan H de Vries, Patrick Deelen, Morris A Swertz, and Lude Franke. Single-cell rna sequencing identifies celltype-specific cis-eqtls and co-expression qtls. *Nature genetics*, 50(4):493–497, 2018.

[161] Benjamin P Fairfax, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, Chris McGee, et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343(6175):1246949, 2014.

[162] Luis B Barreiro, Ludovic Tailleux, Athma A Pai, Brigitte Gicquel, John C Marioni, and Yoav Gilad. Deciphering the genetic architecture of variation in the immune response to mycobacterium tuberculosis infection. *Proceedings of the National Academy of Sciences*, 109(4):1204–1209, 2012.

[163] Sarah Kim-Hellmuth, Matthias Bechheim, Benno Pütz, Pejman Mohammadi, Yohann Nédélec, Nicholas Giangreco, Jessica Becker, Vera Kaiser, Nadine Fricker, Esther Beier, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nature communications*, 8(1):1–10, 2017.

[164] Chen Yao, Roby Joehanes, Andrew D Johnson, Tianxiao Huan, Tõnu Esko, Saixia Ying, Jane E Freedman, Joanne Murabito, Kathryn L Lunetta, Andres Metspalu, et al. Sex-and age-interacting eqtls in human complex diseases. *Human molecular genetics*, 23(7):1947–1956, 2014.

[165] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS Genet*, 6(4):e1000888, 2010.

[166] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, 2016.

[167] Maren E Cannon and Karen L Mohlke. Deciphering the emerging complexities of molecular mechanisms at gwas loci. *The American Journal of Human Genetics*, 103(5):637–653, 2018.

[168] Xin He, Chris K Fuller, Yi Song, Qingying Meng, Bin Zhang, Xia Yang, and Hao Li. Sherlock: detecting gene-disease associations by matching patterns of expression qtl and gwas. *The American Journal of Human Genetics*, 92(5):667–680, 2013.

[169] Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature genetics*, 49(12):1676–1683, 2017.

[170] Boxiang Liu, Michael J Gloudemans, Abhiram S Rao, Erik Ingelsson, and Stephen B Montgomery. Abundant associations with gene expression complicate gwas follow-up. *Nature genetics*, 51(5):768–769, 2019.

[171] Vincent Plagnol, Deborah J Smyth, John A Todd, and David G Clayton. Statistical independence of the colocalized association signals for type 1 diabetes and rps26 gene expression on chromosome 12q13. *Biostatistics*, 10(2):327–334, 2009.

[172] Chris Wallace, Maxime Rotival, Jason D Cooper, Catherine M Rice, Jennie HM Yang, Mhairi McNeill, Deborah J Smyth, David Niblett, François Cambien, Cardiogenics Consortium, et al. Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Human molecular genetics*, 21(12):2815–2824, 2012.

[173] Alexandra C Nica, Stephen B Montgomery, Antigone S Dimas, Barbara E Stranger, Claude Beazley, Inês Barroso, and Emmanouil T Dermitzakis. Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS Genet*, 6(4):e1000895, 2010.

[174] Eddie Cano-Gamez and Gosia Trynka. From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11, 2020.

[175] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383, 2014.

[176] Hui Guo, Mary D Fortune, Oliver S Burren, Ellen Schofield, John A Todd, and Chris Wallace. Integration of disease association and eqtl data using a bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human molecular genetics*, 24(12):3305–3313, 2015.

[177] Claudia Giambartolomei, Jimmy Zhenli Liu, Wen Zhang, Mads Hauberg, Huwenbo Shi, James Boocock, Joe Pickrell, Andrew E Jaffe, CommonMind Consortium, Bogdan Pasaniuc, et al. A bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545, 2018.

[178] Sung Chun, Alexandra Casparino, Nikolaos A Patsopoulos, Damien C Croteau-Chonka, Benjamin A Raby, Philip L De Jager, Shamil R Sunyaev, and Chris Cotsapas. Limited statistical evidence for shared genetic effects of eqtls and autoimmune-disease-associated loci in three major immune-cell types. *Nature genetics*, 49(4):600–605, 2017.

[179] Farhad Hormozdiari, Martijn Van De Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of gwas and eqtl signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[180] Yangqing Deng and Wei Pan. A powerful and versatile colocalization test. *PLOS Computational Biology*, 16(4):e1007778, 2020.

[181] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[182] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9):1091, 2015.

[183] Oscar Hertwig. *Beitraege zur Kenntniss der Bildung, Befruchtung und Theilung des thierischen eies*, volume 1. W. Engelmann, 1875.

[184] Karl Ernst Von Baer. *Über Entwickelungsgeschichte der Thiere: Beobachtung und Reflexion*, volume 1. Bornträger, 1828.

[185] Shigehito Yamada, Mark Hill, and Tetsuya Takakuwa. Human embryology. *New discoveries in embryology. Rijeka, Croatia: InTech. p*, pages 97–124, 2015.

[186] Raymond F Gasser, R John Cork, Brian J Stillwell, and David T McWilliams. Rebirth of human embryology. *Developmental Dynamics*, 243(5):621–628, 2014.

[187] Marta N Shahbazi. Mechanisms of human embryo development: from cell fate to tissue shape and back. *Development*, 147(14), 2020.

[188] Franz Keibel and Franklin Paine Mall. *Manual of human embryology*, volume 1. JB Lippincott Company, 1910.

[189] Hideo Nishimura, Kiichi Takano, Takashi Tanimura, and Mineo Yasuda. Normal and abnormal development of human embryos: first report of the analysis of 1,213 intact embryos. *Teratology*, 1(3):281–290, 1968.

[190] Robert G Edwards, Barry D Bavister, and Patrick C Steptoe. Early stages of fertilization in vitro of human oocytes matured in vitro. *Nature*, 221(5181):632–635, 1969.

[191] John Rock and Miriam F Menkin. In vitro fertilization and cleavage of human ovarian eggs. *Science*, 100(2588):105–107, 1944.

[192] Landrum B Shettles. A morula stage of human ovum developed in vitro. *Fertility and sterility*, 6(4):287–289, 1955.

[193] Robert G Edwards, Patrick C Steptoe, and Jean M Purdy. Fertilization and cleavage in vitro of preovulator human oocytes. *Nature*, 227(5265):1307–1309, 1970.

[194] PC Steptoe, RG Edwards, and JM Purdy. Human blastocysts grown in culture. *Nature*, 229(5280):132–133, 1971.

[195] Kathy K Niakan and Kevin Eggan. Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental biology*, 375(1):54–64, 2013.

[196] Sophie Petropoulos, Daniel Edsgärd, Björn Reinius, Qiaolin Deng, Sarita Pauliina Panula, Simone Codeluppi, Alvaro Plaza Reyes, Sten Linnarsson, Rickard Sandberg, and Fredrik Lanner. Single-cell rna-seq reveals lineage and x chromosome dynamics in human preimplantation embryos. *Cell*, 165(4):1012–1026, 2016.

[197] Edouard Hannezo and Carl-Philipp Heisenberg. Mechanochemical feedback loops in development and disease. *Cell*, 178(1):12–25, 2019.

[198] George L Streeter. Developmental horizons in human embryos. description of age group xi, 13 to 20 somites, and age group xii, 21 to 29 somites. *Contrib. Embryol. Carnegie Inst.*, 30:211–245, 1942.

[199] Ronan O'Rahilly. *Developmental stages in human embryos, including a survey of the Carnegie collection*, volume 1. Carnegie Institution of Washington, 1973.

[200] Ronan O'rahilly and Fabiola Müller. Developmental stages in human embryos: revised and new measurements. *Cells Tissues Organs*, 192(2):73–84, 2010.

[201] Scott F Gilbert. Developmental biology 9th edition, 2008.

[202] Khan Academy. *Human embryogenesis.* 2015.

[203] Kyoko Iwata, Keitaro Yumoto, Minako Sugishima, Chizuru Mizoguchi, Yoshiteru Kai, Yumiko Iba, and Yasuyuki Mio. Analysis of compaction initiation in human embryos by using time-lapse cinematography. *Journal of assisted reproduction and genetics*, 31(4):421–426, 2014.

[204] Connie C Wong, Kevin E Loewke, Nancy L Bossert, Barry Behr, Christopher J De Jonge, Thomas M Baer, and Renee A Reijo Pera. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. *Nature biotechnology*, 28(10):1115, 2010.

[205] William James Larsen. *Human embryology*. Churchill Livingstone, 2001.

[206] Arthur T Hertig, John Rock, and Eleanor C Adams. A description of 34 human ova within the first 17 days of development. *American Journal of Anatomy*, 98(3):435–493, 1956.

[207] Guojun Sheng. Epiblast morphogenesis before gastrulation. *Developmental biology*, 401(1):17–24, 2015.

[208] Lewis Wolpert and Catarina Vicente. An interview with lewis wolpert. *Development (Cambridge, England)*, 142(15):2547–2548, 2015.

[209] Andrew J Becker, Ernest A McCulloch, and James E Till. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*, 197(4866):452–454, 1963.

[210] Martin J Evans and Matthew H Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292(5819):154–156, 1981.

[211] Gail R Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638, 1981.

[212] James A Thomson, Jennifer Kalishman, Thaddeus G Golos, Maureen Durning, Charles P Harris, Robert A Becker, and John P Hearn. Isolation of a primate embryonic stem cell line. *Proceedings of the National Academy of Sciences*, 92(17):7844–7848, 1995.

[213] James A Thomson, Jennifer Kalishman, Thaddeus G Golos, Maureen Durning, Charles P Harris, and John P Hearn. Pluripotent cell lines derived from common marmoset (callithrix jacchus) blastocysts. *Biology of reproduction*, 55(2):254–259, 1996.

[214] James A Thomson, Joseph Itskovitz-Eldor, Sander S Shapiro, Michelle A Waknitz, Jennifer J Swiergiel, Vivienne S Marshall, and Jeffrey M Jones. Embryonic stem cell lines derived from human blastocysts. *science*, 282(5391):1145–1147, 1998.

[215] Department of Health and Human Services. *Regenerative Medicine.* 2006.

[216] Krishanu Saha and Rudolf Jaenisch. Technical challenges in using human induced pluripotent stem cells to model disease. *Cell stem cell*, 5(6):584–595, 2009.

[217] Hans Clevers. Modeling development and disease with organoids. *Cell*, 165(7):1586–1597, 2016.

[218] Madeline A Lancaster, Magdalena Renner, Carol-Anne Martin, Daniel Wenzel, Louise S Bicknell, Matthew E Hurles, Tessa Homfray, Josef M Penninger, Andrew P Jackson, and Juergen A Knoblich. Cerebral organoids model human brain development and microcephaly. *Nature*, 501(7467):373–379, 2013.

[219] Susanne C van den Brink, Anna Alemany, Vincent van Batenburg, Naomi Moris, Marloes Blotenburg, Judith Vivié, Peter Baillie-Johnson, Jennifer Nichols, Katharina F Sonnen, Alfonso Martinez Arias, et al. Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids. *Nature*, pages 1–5, 2020.

[220] Erin A Kimbrel and Robert Lanza. Current status of pluripotent stem cells: moving the first therapies to the clinic. *Nature reviews Drug discovery*, 14(10):681–692, 2015.

[221] Rosario M Isasi and Bartha M Knoppers. Governing stem cell banks and registries: emerging issues. *Stem Cell Research*, 3(2-3):96–105, 2009.

[222] Shinya Yamanaka. Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell stem cell*, 1(1):39–49, 2007.

[223] Konrad Hochedlinger and Rudolf Jaenisch. Nuclear transplantation, embryonic stem cells, and the potential for cell therapy. *New England Journal of Medicine*, 349(3):275–286, 2003.

[224] John B Gurdon. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. *Development*, 10(4):622–640, 1962.

[225] Thomas J King and Robert Briggs. Changes in the nuclei of differentiating gastrula cells, as demonstrated by nuclear transplantation. *Proceedings of the National Academy of Sciences of the United States of America*, 41(5):321, 1955.

[226] Ian Wilmut, Angelika E Schnieke, Jim McWhir, Alexander J Kind, and Keith HS Campbell. Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385(6619):810–813, 1997.

[227] Masako Tada, Yousuke Takahama, Kuniya Abe, Norio Nakatsuji, and Takashi Tada. Nuclear reprogramming of somatic cells by in vitro hybridization with es cells. *Current Biology*, 11(19):1553–1558, 2001.

[228] Chad A Cowan, Jocelyn Atienza, Douglas A Melton, and Kevin Eggan. Nuclear reprogramming of somatic cells after fusion with human embryonic stem cells. *Science*, 309(5739):1369–1373, 2005.

[229] Josef Fulka, Alena Langerova, Pasqualino Loi, Grazyna Ptak, David Albertini, and Helena Fulka. The ups and downs of somatic cell nucleus transfer (scnt) in humans. *Journal of assisted reproduction and genetics*, 30(8):1055–1058, 2013.

[230] Masahito Tachibana, Paula Amato, Michelle Sparman, Nuria Marti Gutierrez, Rebecca Tippner-Hedges, Hong Ma, Eunju Kang, Alimujiang Fulati, Hyo-Sang Lee, Hathaitip Sritanaudomchai, et al. Human embryonic stem cells derived by somatic cell nuclear transfer. *Cell*, 153(6):1228–1238, 2013.

[231] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[232] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *cell*, 131(5):861–872, 2007.

[233] Junying Yu, Maxim A Vodyanik, Kim Smuga-Otto, Jessica Antosiewicz-Bourget, Jennifer L Frane, Shulan Tian, Jeff Nie, Gudrun A Jonsdottir, Victor Ruotti, Ron Stewart, et al. Induced pluripotent stem cell lines derived from human somatic cells. *science*, 318(5858):1917–1920, 2007.

[234] Nobel Media AB. *The Nobel Prize in Physiology or Medicine – 2012 Press Release*. 2012.

[235] Keisuke Okita, Tomoko Ichisaka, and Shinya Yamanaka. Generation of germline-competent induced pluripotent stem cells. *nature*, 448(7151):313–317, 2007.

[236] Adekunle Ebenezer Omole and Adegbenro Omotuyi John Fakoya. Ten years of progress and promise of induced pluripotent stem cells: historical origins, characteristics, mechanisms, limitations, and potential applications. *PeerJ*, 6:e4370, 2018.

[237] Nimet Maherali, Rupa Sridharan, Wei Xie, Jochen Utikal, Sarah Eminli, Katrin Arnold, Matthias Stadtfeld, Robin Yachechko, Jason Tchieu, Rudolf Jaenisch, et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell stem cell*, 1(1):55–70, 2007.

[238] Marius Wernig, Alexander Meissner, Ruth Foreman, Tobias Brambrink, Manching Ku, Konrad Hochedlinger, Bradley E Bernstein, and Rudolf Jaenisch. In vitro reprogramming of fibroblasts into a pluripotent es-cell-like state. *nature*, 448(7151):318–324, 2007.

[239] In-Hyun Park, Rui Zhao, Jason A West, Akiko Yabuuchi, Hongguang Huo, Tan A Ince, Paul H Lerou, M William Lensch, and George Q Daley. Reprogramming of human somatic cells to pluripotency with defined factors. *nature*, 451(7175):141–146, 2008.

[240] Matthias Stadtfeld, Kristen Brennand, and Konrad Hochedlinger. Reprogramming of pancreatic β cells into induced pluripotent stem cells. *Current Biology*, 18(12):890–894, 2008.

[241] Sarah Eminli, Jochen Utikal, Katrin Arnold, Rudolf Jaenisch, and Konrad Hochedlinger. Reprogramming of neural progenitor cells into induced pluripotent stem cells in the absence of exogenous sox2 expression. *Stem cells*, 26(10):2467–2474, 2008.

[242] Jeong Beom Kim, Holm Zaehres, Guangming Wu, Luca Gentile, Kinarm Ko, Vittorio Sebastiano, Marcos J Araúzo-Bravo, David Ruau, Dong Wook Han, Martin Zenke, et al. Pluripotent stem cells induced from adult neural stem cells by reprogramming with two factors. *Nature*, 454(7204):646–650, 2008.

[243] Takashi Aoi, Kojiro Yae, Masato Nakagawa, Tomoko Ichisaka, Keisuke Okita, Kazutoshi Takahashi, Tsutomu Chiba, and Shinya Yamanaka. Generation of pluripotent stem cells from adult mouse liver and stomach cells. *Science*, 321(5889):699–702, 2008.

[244] Jacob Hanna, Styliani Markoulaki, Patrick Schorderet, Bryce W Carey, Caroline Beard, Marius Wernig, Menno P Creyghton, Eveline J Steine, John P Cassady, Ruth Foreman, et al. Direct reprogramming of terminally differentiated mature b lymphocytes to pluripotency. *Cell*, 133(2):250–264, 2008.

[245] Jochen Utikal, Nimet Maherali, Warakorn Kulalert, and Konrad Hochedlinger. Sox2 is dispensable for the reprogramming of melanocytes and melanoma cells into induced pluripotent stem cells. *Journal of cell science*, 122(19):3502–3510, 2009.

[246] Ning Sun, Nicholas J Panetta, Deepak M Gupta, Kitchener D Wilson, Andrew Lee, Fangjun Jia, Shijun Hu, Athena M Cherry, Robert C Robbins, Michael T Longaker, et al. Feeder-free derivation of induced pluripotent stem cells from adult human adipose stem cells. *Proceedings of the National Academy of Sciences*, 106(37):15720–15725, 2009.

[247] Nimet Maherali, Tim Ahfeldt, Alessandra Rigamonti, Jochen Utikal, Chad Cowan, and Konrad Hochedlinger. A high-efficiency system for the generation and study of human induced pluripotent stem cells. *Cell stem cell*, 3(3):340–345, 2008.

[248] WE Lowry, L Richter, R Yachechko, AD Pyle, J Tchieu, R Sridharan, AT Clark, and K Plath. Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proceedings of the National Academy of Sciences*, 105(8):2883–2888, 2008.

[249] K Kim, A Doi, B Wen, K Ng, R Zhao, Patrick Cahan, J Kim, MJ Aryee, H Ji, LIR Ehrlich, et al. Epigenetic memory in induced pluripotent stem cells. *Nature*, 467(7313):285–290, 2010.

[250] Jose M Polo, Susanna Liu, Maria Eugenia Figueroa, Warakorn Kulalert, Sarah Eminli, Kah Yong Tan, Effie Apostolou, Matthias Stadtfeld, Yushan Li, Toshi Shioda, et al. Cell type of origin influences the molecular and functional properties of mouse induced pluripotent stem cells. *Nature biotechnology*, 28(8):848–855, 2010.

[251] Kazutoshi Takahashi and Shinya Yamanaka. A decade of transcription factor-mediated reprogramming to pluripotency. *Nature reviews Molecular cell biology*, 17(3):183, 2016.

[252] Tobias Brambrink, Ruth Foreman, G Grant Welstead, Christopher J Lengner, Marius Wernig, Heikyung Suh, and Rudolf Jaenisch. Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell stem cell*, 2(2):151–159, 2008.

[253] Matthias Stadtfeld, Masaki Nagaya, Jochen Utikal, Gordon Weir, and Konrad Hochedlinger. Induced pluripotent stem cells generated without viral integration. *Science*, 322(5903):945–949, 2008.

[254] Jose M Polo, Endre Anderssen, Ryan M Walsh, Benjamin A Schwarz, Christian M Nefzger, Sue Mei Lim, Marti Borkent, Effie Apostolou, Sara Alaei, Jennifer Cloutier, et al. A molecular roadmap of reprogramming somatic cells into ips cells. *Cell*, 151(7):1617–1632, 2012.

[255] Jenny Hansson, Mahmoud Reza Rafiee, Sonja Reiland, Jose M Polo, Julian Gehring, Satoshi Okawa, Wolfgang Huber, Konrad Hochedlinger, and Jeroen Krijgsveld. Highly coordinated proteome dynamics during reprogramming of somatic cells to pluripotency. *Cell reports*, 2(6):1579–1592, 2012.

[256] Yosef Buganim, Dina A Faddah, Albert W Cheng, Elena Itskovich, Styliani Markoulaki, Kibibi Ganz, Sandy L Klemm, Alexander van Oudenaarden, and Rudolf Jaenisch. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell*, 150(6):1209–1222, 2012.

[257] Marinka Brouwer, Huiqing Zhou, and Nael Nadif Kasri. Choices for induction of pluripotency: recent developments in human induced pluripotent stem cell reprogramming strategies. *Stem Cell Reviews and Reports*, 12(1):54–72, 2016.

[258] Yosef Buganim, Dina A Faddah, and Rudolf Jaenisch. Mechanisms and models of somatic cell reprogramming. *Nature Reviews Genetics*, 14(6):427–439, 2013.

[259] Robert Blelloch, Monica Venere, Jonathan Yen, and Miguel Ramalho-Santos. Generation of induced pluripotent stem cells in the absence of drug selection. *Cell stem cell*, 1(3):245–247, 2007.

[260] Frank Soldner, Dirk Hockemeyer, Caroline Beard, Qing Gao, George W Bell, Elizabeth G Cook, Gunnar Hargus, Alexandra Blak, Oliver Cooper, Maisam Mitalipova, et al. Parkinson's disease patient-derived induced pluripotent stem cells free of viral reprogramming factors. *Cell*, 136(5):964–977, 2009.

[261] Noemi Fusaki, Hiroshi Ban, Akiyo Nishiyama, Koichi Saeki, and Mamoru Hasegawa. Efficient induction of transgene-free human pluripotent stem cells using a vector based on sendai virus, an rna virus that does not integrate into the host genome. *Proceedings of the Japan Academy, Series B*, 85(8):348–362, 2009.

[262] Ken Nishimura, Masayuki Sano, Manami Ohtaka, Birei Furuta, Yoko Umemura, Yoshiro Nakajima, Yuzuru Ikehara, Toshihiro Kobayashi, Hiroaki Segawa, Satoko Takayasu, et al. Development of defective and persistent sendai virus vector a unique gene delivery/expression system ideal for cell reprogramming. *Journal of Biological Chemistry*, 286(6):4760–4771, 2011.

[263] Junying Yu, Kejin Hu, Kim Smuga-Otto, Shulan Tian, Ron Stewart, Igor I Slukvin, and James A Thomson. Human induced pluripotent stem cells free of vector and transgene sequences. *Science*, 324(5928):797–801, 2009.

[264] Keisuke Okita, Masato Nakagawa, Hong Hyenjong, Tomoko Ichisaka, and Shinya Yamanaka. Generation of mouse induced pluripotent stem cells without viral vectors. *Science*, 322(5903):949–953, 2008.

[265] Keisuke Okita, Yasuko Matsumura, Yoshiko Sato, Aki Okada, Asuka Morizane, Satoshi Okamoto, Hyenjong Hong, Masato Nakagawa, Koji Tanabe, Ken-ichi Tezuka, et al. A more efficient method to generate integration-free human ips cells. *Nature methods*, 8(5):409–412, 2011.

[266] Fangjun Jia, Kitchener D Wilson, Ning Sun, Deepak M Gupta, Mei Huang, Zongjin Li, Nicholas J Panetta, Zhi Ying Chen, Robert C Robbins, Mark A Kay, et al. A nonviral minicircle vector for deriving human ips cells. *Nature methods*, 7(3):197–199, 2010.

[267] Keisuke Kaji, Katherine Norrby, Agnieszka Paca, Maria Mileikovsky, Paria Mohseni, and Knut Woltjen. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature*, 458(7239):771–775, 2009.

[268] Knut Woltjen, Iacovos P Michael, Paria Mohseni, Ridham Desai, Maria Mileikovsky, Riikka Hämäläinen, Rebecca Cowling, Wei Wang, Pentao Liu, Marina Gertsenstein, et al. piggybac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*, 458(7239):766–770, 2009.

[269] Luigi Warren, Philip D Manos, Tim Ahfeldt, Yuin-Han Loh, Hu Li, Frank Lau, Wataru Ebina, Pankaj K Mandal, Zachary D Smith, Alexander Meissner, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mrna. *Cell stem cell*, 7(5):618–630, 2010.

[270] Dohoon Kim, Chun-Hyung Kim, Jung-Il Moon, Young-Gie Chung, Mi-Yoon Chang, Baek-Soo Han, Sanghyeok Ko, Eungi Yang, Kwang Yul Cha, Robert Lanza, et al. Generation of human induced pluripotent stem cells by direct delivery of reprogramming proteins. *Cell stem cell*, 4(6):472, 2009.

[271] Trond Aasen, Angel Raya, Maria J Barrero, Elena Garreta, Antonella Consiglio, Federico Gonzalez, Rita Vassena, Josipa Bilić, Vladimir Pekarik, Gustavo Tiscornia, et al. Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nature biotechnology*, 26(11):1276–1284, 2008.

[272] Michael Xavier Doss and Agapios Sachinidis. Current challenges of ipsc-based disease modeling and therapeutic implications. *Cells*, 8(5):403, 2019.

[273] Shigeki Sugii, Yasuyuki Kida, Teruhisa Kawamura, Jotaro Suzuki, Rita Vassena, Yun-Qiang Yin, Margaret K Lutz, W Travis Berggren, Juan Carlos Izpisúa Belmonte, and Ronald M Evans. Human and mouse adipose-derived cells support feeder-independent induction of pluripotent stem cells. *Proceedings of the National Academy of Sciences*, 107(8):3558–3563, 2010.

[274] Yuin-Han Loh, Suneet Agarwal, In-Hyun Park, Achia Urbach, Hongguang Huo, Garrett C Heffner, Kitai Kim, Justine D Miller, Kitwa Ng, and George Q Daley. Generation of induced pluripotent stem cells from human blood. *Blood, The Journal of the American Society of Hematology*, 113(22):5476–5479, 2009.

[275] Tomohisa Seki, Shinsuke Yuasa, Mayumi Oda, Toru Egashira, Kojiro Yae, Dai Kusumoto, Hikari Nakata, Shugo Tohyama, Hisayuki Hashimoto, Masaki Kodaira, et al. Generation of induced pluripotent stem cells from human terminally differentiated circulating t cells. *Cell stem cell*, 7(1):11–14, 2010.

[276] Xing Yan, Haiyan Qin, Cunye Qu, Rocky S Tuan, Songtao Shi, and George T-J Huang. ips cells reprogrammed from human mesenchymal-like stem/progenitor cells of dental tissue origin. *Stem cells and development*, 19(4):469–480, 2010.

[277] Yinyin Cao, Jin Xu, Junxiang Wen, Xiaojing Ma, Fang Liu, Yang Li, Weicheng Chen, Liqun Sun, Yao Wu, Shuolin Li, et al. Generation of a urine-derived ips cell line from a patient with a ventricular septal defect and heart failure and the robust differentiation of these cells to cardiomyocytes via small molecules. *Cellular Physiology and Biochemistry*, 50(2):538–551, 2018.

[278] Zhumur Ghosh, Kitchener D Wilson, Yi Wu, Shijun Hu, Thomas Quertermous, and Joseph C Wu. Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PloS one*, 5(2):e8975, 2010.

[279] Gabriella L Boulting, Evangelos Kiskinis, Gist F Croft, Mackenzie W Amoroso, Derek H Oakley, Brian J Wainger, Damian J Williams, David J Kahler, Mariko Yamaki, Lance Davidow, et al. A functionally characterized test set of human induced pluripotent stem cells. *Nature biotechnology*, 29(3):279–286, 2011.

[280] Franz-Josef Müller, Bernhard M Schuldt, Roy Williams, Dylan Mason, Gulsah Altun, Eirini P Papapetrou, Sandra Danner, Johanna E Goldmann, Arne Herbst, Nils O Schmidt, et al. A bioinformatic assay for pluripotency in human cells. *Nature methods*, 8(4):315, 2011.

[281] Foad Rouhani, Natsuhiko Kumasaka, Miguel Cardoso de Brito, Allan Bradley, Ludovic Vallier, and Daniel Gaffney. Genetic background drives transcriptional variation in human induced pluripotent stem cells. *PLoS Genet*, 10(6):e1004432, 2014.

[282] Leonardo D'Aiuto, Yun Zhi, Dhanjit Kumar Das, Madeleine R Wilcox, Jon W Johnson, Lora McClain, Matthew L MacDonald, Roberto Di Maio, Mark E Schurdak, Paolo Piazza, et al. Large-scale generation of human ipsc-derived neural stem cells/early neural progenitor cells and their neuronal differentiation. *Organogenesis*, 10(4):365–377, 2014.

[283] Yichen Shi, Peter Kirwan, James Smith, Hugh PC Robinson, and Frederick J Livesey. Human cerebral cortex development from pluripotent stem cells to functional excitatory synapses. *Nature neuroscience*, 15(3):477–486, 2012.

[284] Sonja Kriks, Jae-Won Shim, Jinghua Piao, Yosif M Ganat, Dustin R Wakeman, Zhong Xie, Luis Carrillo-Reid, Gordon Auyeung, Chris Antonacci, Amanda Buch, et al. Dopamine neurons derived from human es cells efficiently engraft in animal models of parkinson's disease. *Nature*, 480(7378):547–551, 2011.

[285] Saravanan Karumbayaram, Bennett G Novitch, Michaela Patterson, Joy A Umbach, Laura Richter, Anne Lindgren, Anne E Conway, Amander T Clark, Steve A Goldman, Kathrin Plath, et al. Directed differentiation of human-induced pluripotent stem cells generates active motor neurons. *Stem cells*, 27(4):806–811, 2009.

[286] Atossa Shaltouki, Jun Peng, Qiuyue Liu, Mahendra S Rao, and Xianmin Zeng. Efficient generation of astrocytes from human pluripotent stem cells in defined conditions. *Stem cells*, 31(5):941–952, 2013.

[287] Panagiotis Douvaras, Jing Wang, Matthew Zimmer, Stephanie Hanchuk, Melanie A O'Bara, Saud Sadiq, Fraser J Sim, James Goldman, and Valentina Fossati. Efficient generation of myelinating oligodendrocytes from primary progressive multiple sclerosis patients by induced pluripotent stem cells. *Stem cell reports*, 3(2):250–259, 2014.

[288] Paul W Burridge, Elena Matsa, Praveen Shukla, Ziliang C Lin, Jared M Churko, Antje D Ebert, Feng Lan, Sebastian Diecke, Bruno Huber, Nicholas M Mordwinkin, et al. Chemically defined generation of human cardiomyocytes. *Nature methods*, 11(8):855–860, 2014.

[289] Sara M Maffioletti, Mattia FM Gerli, Martina Ragazzi, Sumitava Dastidar, Sara Benedetti, Mariana Loperfido, Thierry VandenDriessche, Marinee K Chuah, and Francesco Saverio Tedesco. Efficient derivation and inducible differentiation of expandable skeletal myogenic cells from human es and patient-specific ips cells. *Nature protocols*, 10(7):941–958, 2015.

[290] Christoph Patsch, Ludivine Challet-Meylan, Eva C Thoma, Eduard Urich, Tobias Heckel, John F O'Sullivan, Stephanie J Grainger, Friedrich G Kapp, Lin Sun, Klaus Christensen, et al. Generation of vascular endothelial and smooth muscle cells from human pluripotent stem cells. *Nature cell biology*, 17(8):994–1003, 2015.

[291] Karim Si-Tayeb, Fallon K Noto, Masato Nagaoka, Jixuan Li, Michele A Battle, Christine Duris, Paula E North, Stephen Dalton, and Stephen A Duncan. Highly efficient generation of human hepatocyte–like cells from induced pluripotent stem cells. *Hepatology*, 51(1):297–305, 2010.

[292] Donghui Zhang, Wei Jiang, Meng Liu, Xin Sui, Xiaolei Yin, Song Chen, Yan Shi, and Hongkui Deng. Highly efficient differentiation of human es cells and ips cells into mature pancreatic insulin-producing cells. *Cell research*, 19(4):429–438, 2009.

[293] Sarah XL Huang, Mohammad Naimul Islam, John O'neill, Zheng Hu, Yong-Guang Yang, Ya-Wen Chen, Melanie Mumau, Michael D Green, Gordana Vunjak-Novakovic, Jahar Bhattacharya, et al. Efficient generation of lung and airway epithelial cells from human pluripotent stem cells. *Nature biotechnology*, 32(1):84–91, 2014.

[294] Helena Kilpinen, Angela Goncalves, Andreas Leha, Vackar Afzal, Kaur Alasoo, Sofie Ashford, Sendu Bala, Dalila Bensaddek, Francesco Paolo Casale, Oliver J Culley, et al. Common genetic variation drives molecular heterogeneity in human ipscs. *Nature*, 546(7658):370–375, 2017.

[295] Athanasia D Panopoulos, Matteo D'Antonio, Paola Benaglio, Roy Williams, Sherin I Hashem, Bernhard M Schuldt, Christopher DeBoever, Angelo D Arias, Melvin Garcia, Bradley C Nelson, et al. ipscore: a resource of 222 ipsc lines enabling functional characterization of genetic variation across a variety of cell types. *Stem cell reports*, 8(4):1086–1100, 2017.

[296] Ivan Carcamo-Orive, Gabriel E Hoffman, Paige Cundiff, Noam D Beckmann, Sunita L D'Souza, Joshua W Knowles, Achchhe Patel, Dimitri Papatsenko, Fahim Abbasi, Gerald M Reaven, et al. Analysis of transcriptional variability in a large human ipsc

library reveals genetic and non-genetic determinants of heterogeneity. *Cell stem cell*, 20(4):518–532, 2017.

[297] Evanthia E Pashos, Yoson Park, Xiao Wang, Avanthi Raghavan, Wenli Yang, Deepti Abbey, Derek T Peters, Juan Arbelaez, Mayda Hernandez, Nicolas Kuperwasser, et al. Large, diverse population cohorts of hipscs and derived hepatocyte-like cells reveal functional genetic variation at blood lipid-associated loci. *Cell stem cell*, 20(4):558–570, 2017.

[298] Nicholas E Banovich, Yang I Li, Anil Raj, Michelle C Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, Jonathan E Burnett, Marsha Myrthil, Samantha M Thomas, et al. Impact of regulatory variation across human ipscs and differentiated cells. *Genome research*, 28(1):122–131, 2018.

[299] Christopher DeBoever, He Li, David Jakubosky, Paola Benaglio, Joaquin Reyna, Katrina M Olson, Hui Huang, William Biggs, Efren Sandoval, Matteo D'Antonio, et al. Large-scale profiling reveals the influence of genetic variation on gene expression in human induced pluripotent stem cells. *Cell stem cell*, 20(4):533–546, 2017.

[300] Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J Knights, Alice L Mann, Kousik Kundu, Christine Hale, Gordon Dougan, and Daniel J Gaffney. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature genetics*, 50(3):424–431, 2018.

[301] Jeremy Schwartzentruber, Stefanie Foskolou, Helena Kilpinen, Julia Rodrigues, Kaur Alasoo, Andrew J Knights, Minal Patel, Angela Goncalves, Rita Ferreira, Caroline Louise Benn, et al. Molecular and functional variation in ipsc-derived sensory neurons. *Nature genetics*, 50(1):54–61, 2018.

[302] BJ Strober, Reem Elorbany, K Rhodes, Nirmal Krishnan, Karl Tayeb, Alexis Battle, and Yoav Gilad. Dynamic genetic regulation of gene expression during cellular differentiation. *Science*, 364(6447):1287–1290, 2019.

[303] Fumio Hayashi. Econometrics. 2000. *Princeton University Press. Section*, 1:60–69, 2000.

[304] H Desmond Patterson and Robin Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971.

[305] Wei-Min Chen and Hong-Wen Deng. A general and accurate approach for computing the statistical power of the transmission disequilibrium test for complex disease genes. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 21(1):53–67, 2001.

[306] David Clayton and Michael Hills. *Statistical models in epidemiology*. OUP Oxford, 2013.

[307] Abraham Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical society*, 54(3):426–482, 1943.

[308] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

[309] C Radhakrishna Rao. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 44, pages 50–57. Cambridge University Press, 1948.

[310] Robert F Engle. Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826, 1984.

[311] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[312] Robert B Davies. Algorithm as 155: The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(3):323–333, 1980.

[313] Diego Kuonen. Miscellanea. saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935, 1999.

[314] Huan Liu, Yongqiang Tang, and Hao Helen Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

[315] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[316] Peter H Sudmant, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, Markus Hsi-Yang Fritz, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015.

[317] Jae Hoon Sul, Towfique Raj, Simone de Jong, Paul IW De Bakker, Soumya Raychaudhuri, Roel A Ophoff, Barbara E Stranger, Eleazar Eskin, and Buhm Han. Accurate and fast multiple-testing correction in eqtl studies. *The American Journal of Human Genetics*, 96(6):857–868, 2015.

[318] Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient qtl mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

[319] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[320] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[321] Daniel Yekutieli and Yoav Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82(1-2):171–196, 1999.

[322] John D Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[323] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[324] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):e161, 2007.

[325] Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS computational biology*, 6(5), 2010.

[326] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500, 2012.

[327] Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell rna-seq. *Genome biology*, 18(1):212, 2017.

[328] Ghislain Durif, Laurent Modolo, Jeff E Mold, Sophie Lambert-Lacroix, and Franck Picard. Probabilistic count matrix factorization for single cell expression data analysis. *Bioinformatics*, 35(20):4011–4019, 2019.

[329] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

[330] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

[331] Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, Florian Buettner, Wolfgang Huber, and Oliver Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6), 2018.

[332] Hugues Aschard, Vincent Guillemot, Bjarni Vilhjalmsson, Chirag J Patel, David Skurnik, J Ye Chun, Brian Wolpin, Peter Kraft, and Noah Zaitlen. Covariate selection for association screening in multiphenotype genetic studies. *Nature genetics*, 49(12):1789–1795, 2017.

[333] Paul R Burton, Martin D Tobin, and John L Hopper. Key concepts in genetic epidemiology. *The Lancet*, 366(9489):941–951, 2005.

[334] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999.

[335] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, 2004.

[336] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[337] Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[338] Jun Z Li, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, Luigi L Cavalli-Sforza, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *science*, 319(5866):1100–1104, 2008.

[339] Chao Tian, Robert M Plenge, Michael Ransom, Annette Lee, Pablo Villoslada, Carlo Selmi, Lars Klareskog, Ann E Pulver, Lihong Qi, Peter K Gregersen, et al. Analysis and application of european genetic substructure using 300 k snp information. *PLoS Genet*, 4(1):e4, 2008.

[340] Alkes L Price, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L Pascali, Francesca Scarnicci, Andres Ruiz-Linares, Leif Groop, Angelica A Saetta, Penelope Korkolopoulou, et al. Discerning the ancestry of european americans in genetic association studies. *PLoS Genet*, 4(1):e236, 2008.

[341] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646–649, 2008.

[342] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics*, 38(2):203–208, 2006.

[343] Hyun Min Kang, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–1723, 2008.

[344] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[345] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[346] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

[347] Chaeyoung Lee. Genome-wide expression quantitative trait loci analysis using mixed models. *Frontiers in Genetics*, 9:341, 2018.

[348] Gabriel E Hoffman. Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PloS one*, 8(10):e75707, 2013.

[349] Kenneth Lange, Joan Westlake, and M ANNE Spence. Extensions to pedigree analysis. iii. variance components by the scoring method. *Annals of human genetics*, 39(4):485, 1976.

[350] Peter M Visscher, Sarah E Medland, Manuel AR Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet*, 2(3):e41, 2006.

[351] Peter M Visscher, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, Dorret I Boomsma, et al. Genome partitioning of genetic variation for height from 11,214 sibling pairs. *The American Journal of Human Genetics*, 81(5):1104–1110, 2007.

[352] Ben John Hayes, Peter M Visscher, and Michael E Goddard. Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics research*, 91(1):47–60, 2009.

[353] Sang Hong Lee, Michael E Goddard, Peter M Visscher, and Julius HJ Van Der Werf. Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. *Genetics Selection Evolution*, 42(1):22, 2010.

[354] Pieter A Oliehoek, Jack J Windig, Johan AM Van Arendonk, and Piter Bijma. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. *Genetics*, 173(1):483–496, 2006.

[355] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[356] Paul M VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.

[357] Joseph E Powell, Peter M Visscher, and Michael E Goddard. Reconciling the analysis of ibd and ibs in complex trait studies. *Nature Reviews Genetics*, 11(11):800–805, 2010.

[358] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. Gcta: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1):76–82, 2011.

[359] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833, 2011.

[360] Christoph Lippert, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle. Limix: genetic analysis of multiple traits. *BioRxiv*, 2014.

[361] Francesco Paolo Casale, Barbara Rakitsch, Christoph Lippert, and Oliver Stegle. Efficient set tests for the genetic analysis of correlated traits. *Nature methods*, 12(8):755–758, 2015.

[362] Michael E Goddard, Naomi R Wray, Klara Verbyla, Peter M Visscher, et al. Estimating effects and making predictions from genome-wide marker data. *Statistical science*, 24(4):517–529, 2009.

[363] Ali Pazokitoroudi, Yue Wu, Kathryn S Burch, Kangcheng Hou, Aaron Zhou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes. *bioRxiv*, page 522003, 2020.

[364] Charles E McCulloch and John M Neuhaus. Generalized linear mixed models. *Wiley StatsRef: Statistics Reference Online*, 2014.

[365] George EP Box and David R Cox. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243, 1964.

[366] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, 2014.

[367] Nicolo Fusi, Christoph Lippert, Neil D Lawrence, and Oliver Stegle. Warped linear mixed models for the genetic analysis of transformed phenotypes. *Nature communications*, 5(1):1–8, 2014.

[368] Sang Hong Lee, Naomi R Wray, Michael E Goddard, and Peter M Visscher. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, 88(3):294–305, 2011.

[369] Wen-Hua Wei, Gibran Hemani, and Chris S Haley. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11):722, 2014.

[370] James C Alwine, David J Kemp, and George R Stark. Method for detection of specific rnas in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with dna probes. *Proceedings of the National Academy of Sciences*, 74(12):5350–5354, 1977.

[371] Ursula E Gibson, Christian A Heid, and P Mickey Williams. A novel method for real time quantitative rt-pcr. *Genome research*, 6(10):995–1001, 1996.

[372] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[373] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nature methods*, 5(7):621–628, 2008.

[374] John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

[375] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377, 2009.

[376] 10x Genomics. *Our 1.3 million single cell dataset is ready to download*. 2016.

[377] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.

[378] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. *BioRxiv*, page 742304, 2019.

[379] Valentine Svensson. *Single cell studies database*. 2020.

[380] Saiful Islam, Una Kjällquist, Annalena Moliner, Pawel Zajac, Jian-Bing Fan, Peter Lönnerberg, and Sten Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome research*, 21(7):1160–1167, 2011.

[381] Tamar Hashimshony, Florian Wagner, Noa Sher, and Itai Yanai. Cel-seq: single-cell rna-seq by multiplexed linear amplification. *Cell reports*, 2(3):666–673, 2012.

[382] Daniel Ramsköld, Shujun Luo, Yu-Chieh Wang, Robin Li, Qiaolin Deng, Omid R Faridani, Gregory A Daniels, Irina Khrebtukova, Jeanne F Loring, Louise C Laurent, et al. Full-length mrna-seq from single-cell levels of rna and individual circulating tumor cells. *Nature biotechnology*, 30(8):777, 2012.

[383] Simone Picelli, Åsa K Björklund, Omid R Faridani, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, 10(11):1096–1098, 2013.

[384] Yohei Sasagawa, Itoshi Nikaido, Tetsutaro Hayashi, Hiroki Danno, Kenichiro D Uno, Takeshi Imai, and Hiroki R Ueda. Quartz-seq: a highly reproducible and sensitive single-cell rna sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14(4):1–17, 2013.

[385] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, et al. Massively parallel single-cell rna-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, 2014.

[386] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[387] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[388] Todd M Gierahn, Marc H Wadsworth II, Travis K Hughes, Bryan D Bryson, Andrew Butler, Rahul Satija, Sarah Fortune, J Christopher Love, and Alex K Shalek. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature methods*, 14(4):395–398, 2017.

[389] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.

[390] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton JM Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, pages 1–7, 2020.

[391] Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell rna sequencing methods. *Molecular cell*, 65(4):631–643, 2017.

[392] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

[393] Jennifer Westoby, Marcela Sjöberg Herrera, Anne C Ferguson-Smith, and Martin Hemberg. Simulation-based benchmarking of isoform quantification in single-cell rna-seq. *Genome biology*, 19(1):1–14, 2018.

[394] Yuchao Jiang, Nancy R Zhang, and Mingyao Li. Scale: modeling allele-specific gene expression by single-cell rna sequencing. *Genome biology*, 18(1):1–15, 2017.

[395] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[396] Jonathan A Griffiths, Antonio Scialdone, and John C Marioni. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology*, 14(4):e8046, 2018.

[397] Oliver Stegle, Sarah A Teichmann, and John C Marioni. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, 16(3):133–145, 2015.

[398] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLoS computational biology*, 14(6):e1006245, 2018.

[399] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[400] Elham Azizi, Ambrose J Carr, George Plitas, Andrew E Cornish, Catherine Konopacki, Sandhya Prabhakaran, Juozas Nainys, Kenmin Wu, Vaidotas Kiseliovas, Manu Setty, et al. Single-cell map of diverse immune phenotypes in the breast tumor microenvironment. *Cell*, 174(5):1293–1308, 2018.

[401] Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard, and Ines Hellmann. zumis-a fast and flexible pipeline to process rna sequencing data with umis. *Gigascience*, 2018.

[402] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

[403] Davis J McCarthy, Raghd Rostom, Yuanhua Huang, Daniel J Kunz, Petr Danecek, Marc Jan Bonder, Tzachi Hagai, Ruqian Lyu, Wenyi Wang, Daniel J Gaffney, et al. Cardelino: computational integration of somatic clonal substructure and single-cell transcriptomes. *Nature Methods*, 17(4):414–421, 2020.

[404] Aaron TL Lun, Samantha Riesenfeld, Tallulah Andrews, Tomas Gomes, John C Marioni, et al. Emptydrops: distinguishing cells from empty droplets in droplet-based single-cell rna sequencing data. *Genome biology*, 20(1):1–9, 2019.

[405] Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4):281–291, 2019.

[406] Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell systems*, 8(4):329–337, 2019.

[407] Erica AK DePasquale, Daniel J Schnell, Íñigo Valiente-Alandí, Burns C Blaxall, H Leighton Grimes, Harinder Singh, and Nathan Salomonis. Doubletdecon: cell-state aware removal of single-cell rna-seq doublets. *BioRxiv*, page 364810, 2018.

[408] Matthew D Young and Sam Behjati. Soupx removes ambient rna contamination from droplet based single cell rna sequencing data. *BioRxiv*, page 303727, 2020.

[409] Aaron TL Lun, Karsten Bach, and John C Marioni. Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, 17(1):75, 2016.

[410] Catalina A Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C Marioni. Normalizing single-cell rna sequencing data: challenges and opportunities. *Nature methods*, 14(6):565, 2017.

[411] Caleb Weinreb, Samuel Wolock, and Allon M Klein. Spring: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics*, 34(7):1246–1248, 2018.

[412] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

[413] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[414] Sonja Nowotschin, Manu Setty, Ying-Yi Kuo, Vincent Liu, Vidur Garg, Roshan Sharma, Claire S Simon, Nestor Saiz, Rui Gardner, Stéphane C Boutet, et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. *Nature*, 569(7756):361–367, 2019.

[415] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[416] Joshua D Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

[417] Krzysztof Polański, Matthew D Young, Zhichao Miao, Kerstin B Meyer, Sarah A Teichmann, and Jong-Eun Park. Bbknn: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, 2020.

[418] Antonio Scialdone, Kedar N Natarajan, Luis R Saraiva, Valentina Proserpio, Sarah A Teichmann, Oliver Stegle, John C Marioni, and Florian Buettner. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods*, 85:54–61, 2015.

[419] Andrew McDavid, Greg Finak, and Raphael Gottardo. Reply to the contribution of cell cycle to heterogeneity in single-cell rna-seq data. *Nature biotechnology*, 34(6):593–595, 2016.

[420] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature methods*, 10(11):1093–1095, 2013.

[421] Shun H Yip, Pak Chung Sham, and Junwen Wang. Evaluation of tools for highly variable gene discovery from single-cell rna-seq data. *Briefings in bioinformatics*, 20(4):1583–1589, 2019.

[422] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[423] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[424] Kevin R Moon, Jay S Stanley III, Daniel Burkhardt, David van Dijk, Guy Wolf, and Smita Krishnaswamy. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

[425] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, et al. Sc3: consensus clustering of single-cell rna-seq data. *Nature methods*, 14(5):483–486, 2017.

[426] Vincent A Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[427] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 15(5):359–362, 2018.

[428] Laleh Haghverdi, Maren Büttner, F Alexander Wolf, Florian Buettner, and Fabian J Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods*, 13(10):845, 2016.

[429] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.

[430] Sean C Bendall, Kara L Davis, El-ad David Amir, Michelle D Tadmor, Erin F Simonds, Tiffany J Chen, Daniel K Shenfeld, Garry P Nolan, and Dana Pe'er. Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725, 2014.

[431] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20(1):1–9, 2019.

[432] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K Shalek, Chloe K Slichter, Hannah W Miller, M Juliana McElrath, Martin Prlic, et al. others.(2015). mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. *Genome Biology*, 16(1):278.

[433] Hirotaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru SH Ko, Shigeru BH Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. Scode: an efficient regulatory network inference algorithm from single-cell rna-seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 2017.

[434] Thalia E Chan, Michael PH Stumpf, and Ann C Babtie. Gene regulatory network inference from single-cell data using multivariate information measures. *Cell systems*, 5(3):251–267, 2017.

[435] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

[436] Aaron TL Lun, Davis J McCarthy, and John C Marioni. A step-by-step workflow for low-level analysis of single-cell rna-seq data with bioconductor. *F1000Research*, 5, 2016.

[437] D Risso and M Cole. scrnaseq: A collection of public single-cell rna-seq datasets. *R package version*, 1(0), 2016.

[438] Davis J McCarthy, Kieran R Campbell, Aaron TL Lun, and Quin F Wills. Scater: pre-processing, quality control, normalization and visualization of single-cell rna-seq data in r. *Bioinformatics*, 33(8):1179–1186, 2017.

[439] Aaron Lun, Davide Risso, and K Korthauer. Singlecellexperiment: S4 classes for single cell data. *R package version*, 1(1), 2019.

[440] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[441] Minzhe Guo, Hui Wang, S Steven Potter, Jeffrey A Whitsett, and Yan Xu. Sincera: a pipeline for single-cell rna-seq profiling analysis. *PLoS computational biology*, 11(11):e1004575, 2015.

[442] Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620, 2015.

[443] Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25(10):1491–1498, 2015.

[444] Abhishek K Sarkar, Po-Yuan Tung, John D Blischak, Jonathan E Burnett, Yang I Li, Matthew Stephens, and Yoav Gilad. Discovery and characterization of variance qtls in human induced pluripotent stem cells. *PLoS genetics*, 15(4):e1008045, 2019.

[445] Julie Jerber, Daniel D Seaton, Anna SE Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, Daniel Pearce, Malin Andersson, Marc Jan Bonder, et al. Population-scale single-cell rna-seq profiling across dopaminergic neuron differentiation. *bioRxiv*, 2020.

[446] Monique GP van der Wijst, Dylan H de Vries, Hilde E Groot, Gosia Trynka, Chung-Chau Hon, Marc-Jan Bonder, Oliver Stegle, MC Nawijn, Youssef Idaghdour, Pim van der Harst, et al. The single-cell eqtlgen consortium. *Elife*, 9, 2020.

[447] Anna SE Cuomo, Daniel D Seaton, Davis J McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, et al. Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. *Nature communications*, 11(1):1–14, 2020.

[448] Vincent Piras and Kumar Selvarajoo. The reduction of gene expression variability from single cells to populations follows simple statistical laws. *Genomics*, 105(3):137–144, 2015.

[449] Bogdan A Mirauta, Daniel D Seaton, Dalila Bensaddek, Alejandro Brenes, Marc J Bonder, Helena Kilpinen, Oliver Stegle, Angus I Lamond, HipSci Consortium, et al. Population-scale proteome variation in human induced pluripotent stem cells. *BioRxiv*, page 439216, 2018.

[450] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

[451] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417, 2017.

[452] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, et al. Ensembl 2020. *Nucleic acids research*, 48(D1):D682–D688, 2020.

[453] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

[454] Wenhao Tang, François Bertaux, Philipp Thomas, Claire Stefanelli, Malika Saint, Samuel Marguerat, and Vahid Shahrezaei. baynorm: Bayesian gene expression recovery, imputation and normalization for single-cell rna-sequencing data. *Bioinformatics*, 36(4):1174–1181, 2020.

[455] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[456] Nurlan Kerimov, James D Hayhurst, Jonathan R Manning, Peter Walter, Liis Kolberg, Kateryna Peikova, Marija Samoviča, Tony Burdett, Simon Jupp, Helen Parkinson, et al. eqtl catalogue: a compendium of uniformly processed human gene expression and splicing qtls. *BioRxiv*, 2020.

[457] NM Editorial. Method of the year 2013. *Nat Methods*, 11:1, 2014.

[458] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21(1):1–35, 2020.

[459] Benedict Anchang, Tom DP Hart, Sean C Bendall, Peng Qiu, Zach Bjornson, Michael Linderman, Garry P Nolan, and Sylvia K Plevritis. Visualization and cellular hierarchy inference of single-cell data using spade. *Nature protocols*, 11(7):1264–1279, 2016.

[460] Matthew D Young, Thomas J Mitchell, Felipe A Vieira Braga, Maxine GB Tran, Benjamin J Stewart, John R Ferdinand, Grace Collord, Rachel A Botting, Dorin-Mirel Popescu, Kevin W Loudon, et al. Single-cell transcriptomes from human kidneys reveal the cellular identity of renal tumors. *science*, 361(6402):594–599, 2018.

[461] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A Engelse, Francoise Carlotti, Eelco JP de Koning, et al. A single-cell transcriptome atlas of the human pancreas. *Cell systems*, 3(4):385–394, 2016.

[462] Christina Ernst, Nils Eling, Celia P Martinez-Jimenez, John C Marioni, and Duncan T Odom. Staged developmental mapping and x chromosome transcriptional dynamics during mouse spermatogenesis. *Nature communications*, 10(1):1–20, 2019.

[463] Blanca Pijuan-Sala, Jonathan A Griffiths, Carolina Guibentif, Tom W Hiscock, Wajid Jawaid, Fernando J Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard CV Tyser, Debbie Lee Lian Ho, et al. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019.

[464] Lars Velten, Simon F Haas, Simon Raffel, Sandra Blaszkiewicz, Saiful Islam, Bianca P Hennig, Christoph Hirche, Christoph Lutz, Eike C Buss, Daniel Nowak, et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature cell biology*, 19(4):271–281, 2017.

[465] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni, and Oliver Stegle. Computational analysis of cell-to-cell heterogeneity in single-cell rna-sequencing data reveals hidden subpopulations of cells. *Nature biotechnology*, 33(2):155, 2015.

[466] Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology*, 33(3):269–276, 2015.

[467] Anna SE Cuomo, Giordano Alvari, Christina B Azodi, Davis J McCarthy, Marc Jan Bonder, et al. Optimising expression quantitative trait locus mapping workflows for single-cell studies. *bioRxiv*, 2021.

[468] Julien F Ayroles, Sean M Buchanan, Chelsea O'Leary, Kyobi Skutt-Kakaria, Jennifer K Grenier, Andrew G Clark, Daniel L Hartl, and Benjamin L De Bivort. Behavioral idiosyncrasy reveals genetic control of phenotypic variability. *Proceedings of the National Academy of Sciences*, 112(21):6706–6711, 2015.

[469] Catalina A Vallejos, Sylvia Richardson, and John C Marioni. Beyond comparisons of means: understanding changes in gene expression at the single-cell level. *Genome biology*, 17(1):1–14, 2016.

[470] Mirko Francesconi and Ben Lehner. The effects of genetic variation on gene expression dynamics during development. *Nature*, 505(7482):208–211, 2014.

[471] Rolando Rivera-Pomar and Herbert Jackle. From gradients to stripes in drosophila embryogenesis: filling in the gaps. *Trends in Genetics*, 12(11):478–483, 1996.

[472] Pavel Tomancak, Amy Beaton, Richard Weiszmann, Elaine Kwan, ShengQiang Shu, Suzanna E Lewis, Stephen Richards, Michael Ashburner, Volker Hartenstein, Susan E Celniker, et al. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome biology*, 3(12):research0088–1, 2002.

[473] Darren A Cusanovich, James P Reddington, David A Garfield, Riza M Daza, Delasa Aghamirzaie, Raquel Marco-Ferreres, Hannah A Pliner, Lena Christiansen, Xiaojie Qiu, Frank J Steemers, et al. The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*, 555(7697):538–542, 2018.

[474] David Jonah Grunwald and Judith S Eisen. Headwaters of the zebrafish—emergence of a new model vertebrate. *Nature reviews genetics*, 3(9):717–724, 2002.

[475] Elke A Ober, Holly A Field, and Didier YR Stainier. From endoderm formation to liver and pancreas development in zebrafish. *Mechanisms of development*, 120(1):5–18, 2003.

[476] David Traver, Barry H Paw, Kenneth D Poss, W Todd Penberthy, Shuo Lin, and Leonard I Zon. Transplantation and in vivo imaging of multilineage engraftment in zebrafish bloodless mutants. *Nature immunology*, 4(12):1238–1246, 2003.

[477] Laura Beth Corson, Yojiro Yamanaka, Ka-Man Venus Lai, and Janet Rossant. Spatial and temporal patterns of erk signaling during mouse embryogenesis. *Development*, 130(19):4527–4537, 2003.

[478] Patrick PL Tam and David AF Loebel. Gene function in mouse embryogenesis: get set for gastrulation. *Nature Reviews Genetics*, 8(5):368–381, 2007.

[479] Christoph Bock, Evangelos Kiskinis, Griet Verstappen, Hongcang Gu, Gabriella Boulting, Zachary D Smith, Michael Ziller, Gist F Croft, Mackenzie W Amoroso, Derek H Oakley, et al. Reference maps of human es and ips cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144(3):439–452, 2011.

[480] Nicholas RF Hannan, Charis-Patricia Segeritz, Thomas Touboul, and Ludovic Vallier. Production of hepatocyte-like cells from human pluripotent stem cells. *Nature protocols*, 8(2):430, 2013.

[481] Greg Finak, Jacob Frelinger, Wenxin Jiang, Evan W Newell, John Ramey, Mark M Davis, Spyros A Kalams, Stephen C De Rosa, and Raphael Gottardo. Opencyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol*, 10(8):e1003806, 2014.

[482] Krueger F Trim Galore. A wrapper tool around cutadapt and fastqc to consistently apply quality and adapter trimming to fastq files. 2015.

[483] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):10–12, 2011.

[484] Simon Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

[485] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[486] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[487] Tom Smith, Andreas Heger, and Ian Sudbery. Umi-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, 27(3):491–499, 2017.

[488] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature methods*, 11(2):163, 2014.

[489] Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

[490] Trevor Hastie and Werner Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[491] David A Knowles, Joe R Davis, Hilary Edgington, Anil Raj, Marie-Julie Favé, Xiaowei Zhu, James B Potash, Myrna M Weissman, Jianxin Shi, Douglas F Levinson, et al. Allele-specific expression reveals interactions between genetic variation and environment. *Nature methods*, 14(7):699–702, 2017.

[492] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, 13(3):241, 2016.

[493] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *science*, 315(5814):972–976, 2007.

[494] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: machine learning in python*. Packt Publishing Ltd, 2013.

[495] DV Klopfenstein, Liangsheng Zhang, Brent S Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J Mungall, Jeffrey M Yunes, Olga Botvinnik, Mark Weigel, et al. Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):1–17, 2018.

[496] Xiuling Xu, Shunlei Duan, Fei Yi, Alejandro Ocampo, Guang-Hui Liu, and Juan Carlos Izpisua Belmonte. Mitochondrial regulation in pluripotent stem cells. *Cell metabolism*, 18(3):325–332, 2013.

[497] Gregory A Moyerbrailean, Allison L Richards, Daniel Kurtz, Cynthia A Kalita, Gordon O Davis, Chris T Harvey, Adnan Alazizi, Donovan Watza, Yoram Sorokin, Nancy Hauff, et al. High-throughput allele-specific expression across 250 environmental conditions. *Genome research*, 26(12):1627–1638, 2016.

[498] Harvind S Chahal, Wenting Wu, Katherine J Ransohoff, Lingyao Yang, Haley Hedlin, Manisha Desai, Yuan Lin, Hong-Ji Dai, Abrar A Qureshi, Wen-Qing Li, et al. Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nature communications*, 7(1):1–10, 2016.

[499] Fredrick R Schumacher, Ali Amin Al Olama, Sonja I Berndt, Sara Benlloch, Mahbubl Ahmed, Edward J Saunders, Tokhir Dadaev, Daniel Leongamornlert, Ezequiel Anokian, Clara Cieza-Borrella, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nature genetics*, 50(7):928–936, 2018.

[500] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[501] Montserrat C Anguera, Ruslan Sadreyev, Zhaoqing Zhang, Attila Szanto, Bernhard Payer, Steven D Sheridan, Showming Kwok, Stephen J Haggarty, Mriganka Sur, Jason Alvarez, et al. Molecular signatures of human induced pluripotent stem cells highlight sex differences and cancer genes. *Cell stem cell*, 11(1):75–90, 2012.

[502] Sanjeet Patel, Giancarlo Bonora, Anna Sahakyan, Rachel Kim, Constantinos Chronis, Justin Langerman, Sorel Fitz-Gibbon, Liudmilla Rubbi, Rhys JP Skelton, Reza Ardehali, et al. Human embryonic stem cells do not change their x inactivation status during differentiation. *Cell reports*, 18(1):54–67, 2017.

[503] Yunlong Tao and Su-Chun Zhang. Neural subtype specification from human pluripotent stem cells. *Cell stem cell*, 19(5):573–586, 2016.

[504] Eike Müller, Weijia Wang, Wenlian Qiao, Martin Bornhäuser, Peter W Zandstra, Carsten Werner, and Tilo Pompe. Distinguishing autocrine and paracrine signals in hematopoietic stem cell culture using a biofunctional microcavity platform. *Scientific reports*, 6(1):1–12, 2016.

[505] Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, and K Frazer. In vitro differentiation of human ipsc-derived retinal pigment epithelium cells (ipsc-rpe)., 2019.

[506] Viola Volpato, James Smith, Cynthia Sandor, Janina S Ried, Anna Baud, Adam Handel, Sarah E Newey, Frank Wessely, Moustafa Attar, Emma Whiteley, et al. Reproducibility of molecular phenotypes after long-term differentiation to human ipsc-derived neurons: a multi-site omics study. *Stem cell reports*, 11(4):897–911, 2018.

[507] Quan H Nguyen, Samuel W Lukowski, Han Sheng Chiu, Anne Senabouth, Timothy JC Bruxner, Angelika N Christ, Nathan J Palpant, and Joseph E Powell. Single-cell rna-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. *Genome research*, 28(7):1053–1066, 2018.

[508] Andrew Lees, John Hardy, and Tamas Revesz. Parkinson's disease. *Lancet*, 373(9680):2055–2066, 2009.

[509] Teresia Osborn and Penelope J Hallett. Seq-ing markers of midbrain dopamine neurons. *Cell stem cell*, 20(1):11–12, 2017.

[510] Theo Stoddard-Bennett and Renee Reijo Pera. Stem cell therapy for parkinson's disease: safety and modeling. *Neural regeneration research*, 15(1):36, 2020.

[511] Nian Xiong, Xi Long, Jing Xiong, Min Jia, Chunnuan Chen, Jinsha Huang, Devina Ghoorah, Xiangquan Kong, Zhicheng Lin, and Tao Wang. Mitochondrial complex i inhibitor rotenone-induced toxicity and its potential mechanisms in parkinson's disease models. *Critical reviews in toxicology*, 42(7):613–632, 2012.

[512] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature methods*, pages 1–8, 2019.

[513] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[514] Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E Borm, Simon RW Stott, Enrique M Toledo, J Carlos Villaescusa, et al. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell*, 167(2):566–580, 2016.

[515] Nicolas Bertrand, Diogo S Castro, and François Guillemot. Proneural genes and the specification of neural cell types. *Nature Reviews Neuroscience*, 3(7):517–530, 2002.

[516] Marine Lacomme, Laurence Liaubet, Fabienne Pituello, and Sophie Bel-Vialar. Neurog2 drives cell cycle exit of neuronal precursors by specifically repressing a subset of cyclins acting at the g1 and s phases of the cell cycle. *Molecular and cellular biology*, 32(13):2596–2607, 2012.

[517] Ernest Arenas, Mark Denham, and J Carlos Villaescusa. How to make a midbrain dopaminergic neuron. *Development*, 142(11):1918–1936, 2015.

[518] Chang-Hwan Park, Jin Sun Kang, Yeon Ho Shin, Mi-Yoon Chang, Seungsoo Chung, Hyun-Chul Koh, Mei Hong Zhu, Seog Bae Oh, Yong-Sung Lee, Georgia Panagiotakos, et al. Acquisition of in vitro and in vivo functionality of nurr1-induced dopamine neurons. *The FASEB journal*, 20(14):2553–2555, 2006.

[519] David Ramonet, Agata Podhajska, Klodjan Stafa, Sarah Sonnay, Alzbeta Trancikova, Elpida Tsika, Olga Pletnikova, Juan C Troncoso, Liliane Glauser, and Darren J Moore. Park9-associated atp13a2 localizes to intracellular acidic vesicles and regulates cation homeostasis and neuronal integrity. *Human molecular genetics*, 21(8):1725–1743, 2012.

[520] Kevin J Cummings and Matthew R Hodges. The serotonergic system and the control of breathing during development. *Respiratory physiology & neurobiology*, 270:103255, 2019.

[521] John N Campbell, Evan Z Macosko, Henning Fenselau, Tune H Pers, Anna Lyubetskaya, Danielle Tenen, Melissa Goldman, Anne MJ Verstegen, Jon M Resch, Steven A McCarroll, et al. A molecular census of arcuate hypothalamus and median eminence cell types. *Nature neuroscience*, 20(3):484–496, 2017.

[522] Steven A Sloan, Spyros Darmanis, Nina Huber, Themasap A Khan, Fikri Birey, Christine Caneda, Richard Reimer, Stephen R Quake, Ben A Barres, and Sergiu P Paşca. Human astrocyte maturation captured in 3d cerebral cortical spheroids derived from pluripotent stem cells. *Neuron*, 95(4):779–790, 2017.

[523] Ye Zhang, Steven A Sloan, Laura E Clarke, Christine Caneda, Colton A Plaza, Paul D Blumenthal, Hannes Vogel, Gary K Steinberg, Michael SB Edwards, Gordon Li, et al. Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. *Neuron*, 89(1):37–53, 2016.

[524] Todd B Sherer, Ranjita Betarbet, Claudia M Testa, Byoung Boo Seo, Jason R Richardson, Jin Ho Kim, Gary W Miller, Takao Yagi, Akemi Matsuno-Yagi, and J Timothy Greenamyre. Mechanism of toxicity in rotenone models of parkinson's disease. *Journal of Neuroscience*, 23(34):10756–10764, 2003.

[525] H Knönagel and U Karmann. Autologous blood transfusions in interventions of the pelvis using the cell saver. *Helvetica Chirurgica Acta*, 59(3):485–488, 1992.

[526] Jason R Cannon, Victor Tapias, Hye Mee Na, Anthony S Honick, Robert E Drolet, and J Timothy Greenamyre. A highly reproducible rotenone model of parkinson's disease. *Neurobiology of disease*, 34(2):279–290, 2009.

[527] Agnieszka D'Antonio-Chronowska, Margaret KR Donovan, William W Young Greenwald, Jennifer Phuong Nguyen, Kyohei Fujita, Sherin Hashem, Hiroko Matsui, Francesca Soncin, Mana Parast, Michelle C Ward, et al. Association of human ipsc gene signatures and x chromosome dosage with two distinct cardiac differentiation trajectories. *Stem cell reports*, 13(5):924–938, 2019.

[528] Weilan Ye, Kenji Shimamura, John LR Rubenstein, Mary A Hynes, and Arnon Rosenthal. Fgf and shh signals control dopaminergic and serotonergic cell fate in the anterior neural plate. *Cell*, 93(5):755–766, 1998.

[529] Lining Cao, Rui Hu, Ting Xu, Zhen-Ning Zhang, Weida Li, and Jianfeng Lu. Characterization of induced pluripotent stem cell-derived human serotonergic neurons. *Frontiers in Cellular Neuroscience*, 11:131, 2017.

[530] Madeline A Lancaster, Nina S Corsini, Simone Wolfinger, E Hilary Gustafson, Alex W Phillips, Thomas R Burkard, Tomoki Otani, Frederick J Livesey, and Juergen A Knoblich. Guided self-organization and cortical plate formation in human brain organoids. *Nature biotechnology*, 35(7):659–666, 2017.

[531] David J Miller and Patrice E Fort. Heat shock proteins regulatory role in neurodevelopment. *Frontiers in neuroscience*, 12:821, 2018.

[532] Britta Bartelt-Kirbach, Margarethe Moron, Maximilian Glomb, Clara-Maria Beck, Marie-Pascale Weller, and Nikola Golenhofen. Hspb5/αb-crystallin increases dendritic complexity and protects the dendritic arbor during heat shock in cultured rat hippocampal neurons. *Cellular and Molecular Life Sciences*, 73(19):3761–3775, 2016.

[533] Hideki Shimura, Yuko Miura-Shimura, and Kenneth S Kosik. Binding of tau to heat shock protein 27 leads to decreased concentration of hyperphosphorylated tau and enhanced cell survival. *Journal of Biological Chemistry*, 279(17):17957–17962, 2004.

[534] Micha MM Wilhelmus, Wilbert C Boelens, Irene Otte-Höller, Bram Kamps, Robert MW de Waal, and Marcel M Verbeek. Small heat shock proteins inhibit amyloid-β protein aggregation and cerebrovascular amyloid-β protein toxicity. *Brain research*, 1089(1):67–78, 2006.

[535] Sara Tucci. Brain metabolism and neurological symptoms in combined malonic and methylmalonic aciduria. *Orphanet Journal of Rare Diseases*, 15(1):27, 2020.

[536] Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, 2019.

[537] Matthew Stephens. *eQTL analysis outline.* 2020.

[538] Graham Palmer, Douglas J Horgan, Howard Tisdale, Thomas P Singer, and Helmut Beinert. Studies on the respiratory chain-linked reduced nicotinamide adenine dinucleotide dehydrogenase xiv. location of the sites of inhibition of rotenone, barbiturates, and piericidin by means of electron paramagnetic resonance spectroscopy. *Journal of Biological Chemistry*, 243(4):844–847, 1968.

[539] Ranjita Betarbet, Todd B Sherer, Gillian MacKenzie, Monica Garcia-Osuna, Alexander V Panov, and J Timothy Greenamyre. Chronic systemic pesticide exposure reproduces features of parkinson's disease. *Nature neuroscience*, 3(12):1301–1306, 2000.

[540] Dengke K Ma, Karthikeyan Ponnusamy, Mi-Ryoung Song, Guo-li Ming, and Hongjun Song. Molecular genetic analysis of fgfr1 signalling reveals distinct roles of mapk and plcγ1 activation for self-renewal of adult neural stem cells. *Molecular brain*, 2(1):16, 2009.

[541] EK Stachowiak, CA Benson, ST Narla, A Dimitri, LE Bayona Chuye, S Dhiman, K Harikrishnan, S Elahi, D Freedman, KJ Brennand, et al. Cerebral organoids reveal early cortical maldevelopment in schizophrenia—computational anatomy and genomics, role of fgfr1. *Translational psychiatry*, 7(11):1–24, 2017.

[542] International Stem Cell Initiative et al. Assessment of established techniques to determine developmental and malignant potential of human pluripotent stem cells. *Nature communications*, 9, 2018.

[543] Alexander M Tsankov, Veronika Akopian, Ramona Pop, Sundari Chetty, Casey A Gifford, Laurence Daheron, Nadejda M Tsankova, and Alexander Meissner. A qpcr scorecard quantifies the differentiation potential of human pluripotent stem cells. *Nature biotechnology*, 33(11):1182–1192, 2015.

[544] Shijun Hu, Ming-Tao Zhao, Fereshteh Jahanbani, Ning-Yi Shao, Won Hee Lee, Haodong Chen, Michael P Snyder, and Joseph C Wu. Effects of cellular origin on differentiation of human induced pluripotent stem cell–derived endothelial cells. *JCI insight*, 1(8), 2016.

[545] Nathalie Spassky, Florian T Merkle, Nuria Flames, Anthony D Tramontin, José Manuel García-Verdugo, and Arturo Alvarez-Buylla. Adult ependymal cells are postmitotic and are derived from radial glial cells during embryogenesis. *Journal of Neuroscience*, 25(1):10–18, 2005.

[546] Alfonso Lavado and Guillermo Oliver. Six3 is required for ependymal cell maturation. *Development*, 138(24):5291–5300, 2011.

[547] Gregory J Michael, Sharmin Esmailzadeh, Linda B Moran, Lynne Christian, Ronald KB Pearce, and Manuel B Graeber. Up-regulation of metallothionein gene expression in parkinsonian astrocytes. *Neurogenetics*, 12(4):295–305, 2011.

[548] Benoit V Jacquet, Raul Salinas-Mondragon, Huixuan Liang, Blair Therit, Justin D Buie, Michael Dykstra, Kenneth Campbell, Lawrence E Ostrowski, Steven L Brody, and H Troy Ghashghaei. Foxj1-dependent gene expression is required for differentiation of radial glia into ependymal cells and a subset of astrocytes in the postnatal brain. *Development*, 136(23):4021–4031, 2009.

[549] Greg Gibson and Bruce Weir. The quantitative genetics of transcription. *TRENDS in Genetics*, 21(11):616–623, 2005.

[550] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS Genet*, 9(5):e1003486, 2013.

[551] Jae Hoon Sul, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin. Effectively identifying eqtls from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet*, 9(6):e1003491, 2013.

[552] Dominic Grün, Lennart Kester, and Alexander Van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6):637–640, 2014.

[553] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome biology*, 20(1):1–15, 2019.

[554] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147–150, 2020.

[555] Rachel Moore, Francesco Paolo Casale, Marc Jan Bonder, Danilo Horta, Lude Franke, Inês Barroso, and Oliver Stegle. A linear mixed-model approach to study multivariate gene–environment interactions. *Nature genetics*, 51(1):180–186, 2019.

[556] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[557] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *Elife*, 6:e27041, 2017.

[558] Jason D Buenrostro, M Ryan Corces, Caleb A Lareau, Beijing Wu, Alicia N Schep, Martin J Aryee, Ravindra Majeti, Howard Y Chang, and William J Greenleaf. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell*, 173(6):1535–1548, 2018.

[559] M Ryan Corces, Jason D Buenrostro, Beijing Wu, Peyton G Greenside, Steven M Chan, Julie L Koenig, Michael P Snyder, Jonathan K Pritchard, Anshul Kundaje, William J Greenleaf, et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature genetics*, 48(10):1193–1203, 2016.

[560] Hongshan Guo, Ping Zhu, Xinglong Wu, Xianlong Li, Lu Wen, and Fuchou Tang. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. *Genome research*, 23(12):2126–2135, 2013.

[561] Sébastien A Smallwood, Heather J Lee, Christof Angermueller, Felix Krueger, Heba Saadeh, Julian Peat, Simon R Andrews, Oliver Stegle, Wolf Reik, and Gavin Kelsey. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nature methods*, 11(8):817–820, 2014.

[562] Matthias Farlik, Nathan C Sheffield, Angelo Nuzzo, Paul Datlinger, Andreas Schönegger, Johanna Klughammer, and Christoph Bock. Single-cell dna methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell reports*, 10(8):1386–1397, 2015.

[563] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. Single-cell chip-seq reveals cell subpopulations defined by chromatin state. *Nature biotechnology*, 33(11):1165–1172, 2015.

[564] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[565] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865, 2017.

[566] Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409):1380–1385, 2018.

[567] Stephen J Clark, Ricard Argelaguet, Chantriolnt-Andreas Kapourani, Thomas M Stubbs, Heather J Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C Marioni, et al. scnmt-seq enables joint profiling of chromatin accessibility dna methylation and transcription in single cells. *Nature communications*, 9(1):1–9, 2018.

[568] Molly Gasperini, Andrew J Hill, José L McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S Noble, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, 176(1-2):377–390, 2019.

[569] Jean-Charles Lambert, Carla A Ibrahim-Verbaas, Denise Harold, Adam C Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L DeStefano, Joshua C Bis, Gary W Beecham, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nature genetics*, 45(12):1452–1458, 2013.

[570] Mike A Nalls, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, et al. Expanding parkinson's disease genetics: novel risk loci, genomic context, causal insights and heritable risk. *BioRxiv*, page 388165, 2019.

[571] Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511(7510):421–427, 2014.

[572] Liping Hou, Sarah E Bergen, Nirmala Akula, Jie Song, Christina M Hultman, Mikael Landén, Mazda Adli, Martin Alda, Raffaella Ardau, Bárbara Arias, et al. Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Human molecular genetics*, 25(15):3383–3394, 2016.

[573] Aysu Okbay, Bart ML Baselmans, Jan-Emmanuel De Neve, Patrick Turley, Michel G Nivard, Mark Alan Fontana, S Fleur W Meddens, Richard Karlsson Linnér, Cornelius A Rietveld, Jaime Derringer, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature genetics*, 48(6):624–633, 2016.

[574] Michelle Luciano, Saskia P Hagenaars, Gail Davies, W David Hill, Toni-Kim Clarke, Masoud Shirali, Sarah E Harris, Riccardo E Marioni, David C Liewald, Chloe Fawns-Ritchie, et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. *Nature genetics*, 50(1):6–11, 2018.

[575] Patrick Turley, Raymond K Walters, Omeed Maghzian, Aysu Okbay, James J Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, Robbee Wedow, Meghan Zacher, Nicholas A Furlotte, et al. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature genetics*, 50(2):229–237, 2018.

[576] David M Howard, Mark J Adams, Masoud Shirali, Toni-Kim Clarke, Riccardo E Marioni, Gail Davies, Jonathan RI Coleman, Clara Alloza, Xueyi Shen, Miruna C Barbu, et al. Genome-wide association study of depression phenotypes in uk biobank identifies variants in excitatory synaptic pathways. *Nature communications*, 9(1):1–10, 2018.

[577] Gail Davies, Riccardo E Marioni, David C Liewald, W David Hill, Saskia P Hagenaars, Sarah E Harris, Stuart J Ritchie, Michelle Luciano, Chloe Fawns-Ritchie, Donald Lyall, et al. Genome-wide association study of cognitive functions and educational attainment in uk biobank (n= 112 151). *Molecular psychiatry*, 21(6):758–767, 2016.

[578] Riccardo E Marioni, Sarah E Harris, Qian Zhang, Allan F McRae, Saskia P Hagenaars, W David Hill, Gail Davies, Craig W Ritchie, Catharine R Gale, John M Starr, et al. Gwas on family history of alzheimer's disease. *Translational psychiatry*, 8(1):1–7, 2018.

[579] Mats Nagel, Philip R Jansen, Sven Stringer, Kyoko Watanabe, Christiaan A de Leeuw, Julien Bryois, Jeanne E Savage, Anke R Hammerschlag, Nathan G Skene, Ana B Muñoz-Manchado, et al. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics*, 50(7):920–927, 2018.

[580] James J Lee, Robbee Wedow, Aysu Okbay, Edward Kong, Omeed Maghzian, Meghan Zacher, Tuan Anh Nguyen-Viet, Peter Bowers, Julia Sidorenko, Richard Karlsson Linnér, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics*, 50(8):1112–1121, 2018.

[581] Neale Lab. *GWAS analysis of 7,221 phenotypes across 6 continental ancestry groups in the UK Biobank.* 2018.

[582] Thomas Touboul, Nicholas RF Hannan, Sébastien Corbineau, Amélie Martinez, Clémence Martinet, Sophie Branchereau, Sylvie Mainot, Hélène Strick-Marchand, Roger Pedersen, James Di Santo, et al. Generation of functional hepatocytes from human embryonic stem cells under chemically defined conditions that recapitulate liver development. *Hepatology*, 51(5):1754–1765, 2010.

[583] Loukia Yiangou, Alexander DB Ross, Kim Jee Goh, and Ludovic Vallier. Human pluripotent stem cell-derived endoderm for modeling development and clinical applications. *Cell Stem Cell*, 22(4):485–499, 2018.

[584] I Gabrielle M Brons, Lucy E Smithers, Matthew WB Trotter, Peter Rugg-Gunn, Bowen Sun, Susana M Chuva de Sousa Lopes, Sarah K Howlett, Amanda Clarkson, Lars Ahrlund-Richter, Roger A Pedersen, et al. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature*, 448(7150):191–195, 2007.

[585] Siim Pauklin and Ludovic Vallier. The cell-cycle state of stem cells determines cell fate propensity. *Cell*, 155(1):135–147, 2013.

[586] Asako Sakaue-Sawano, Hiroshi Kurokawa, Toshifumi Morimura, Aki Hanyu, Hiroshi Hama, Hatsuki Osawa, Saori Kashiwagi, Kiyoko Fukami, Takaki Miyata, Hiroyuki Miyoshi, et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132(3):487–498, 2008.

[587] Ludovic Vallier, Thomas Touboul, Zhenzhi Chng, Minodora Brimpari, Nicholas Hannan, Enrique Millan, Lucy E Smithers, Matthew Trotter, Peter Rugg-Gunn, Anne

Weber, et al. Early cell fate decisions of human embryonic stem cells and mouse epiblast stem cells are controlled by the same signalling pathways. *PloS one*, 4(6):e6082, 2009.

[588] Siim Pauklin, Pedro Madrigal, Alessandro Bertero, and Ludovic Vallier. Initiation of stem cell differentiation involves cell cycle-dependent regulation of developmental genes by cyclin d. *Genes & development*, 30(4):421–433, 2016.

[589] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

[590] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS Comput Biol*, 9(11):e1003326, 2013.

[591] Xin Feng, Robert Grossman, and Lincoln Stein. Peakranger: a cloud-enabled peak caller for chip-seq data. *BMC bioinformatics*, 12(1):139, 2011.

[592] Aaron R Quinlan and Ira M Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[593] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

Appendix A

Supplementary Tables

A.1 Additional results Chapter 3

	0	5	10	15	20	25
PCA	41	77	78	88	86	84
PEER	41	70	70	71	63	87
MOFA	41	64	76	75	77	73
MOFA-ns	41	74	78	73	74	72
LDVAE	41	58	54	71	74	73

Table A.1: Number of m-bulk-replicated eQTL across covariates.
Equivalent to **Table 3.4** (i.e. mean as aggregation method, 1,421 chromosome 2 genes tested only, FDR<10% and same direction of effect as replication), except considering replication using bulk with matched samples only (m-bulk; n=88).

A.2 Additional results for Chapter 4

	eGenes	genes tested	cells	unique donors	unique samples*
iPSC (day0)	1,833	10,840	9,661	111	136
mesendo	1,702	10,924	9,809	123	224
defendo	1,342	10,901	10,924	116	238
transitioning	227	10,924	6,387	118	313
day1	1,181	10,787	9,443	111	138
day2	718	10,788	8,455	105	116
day3	631	10,765	8,485	108	127

Table A.2: Summary of the type and number of eQTL.
Including all eQTL discovered based on single cell RNA traits, both at the level of our computationally-defined stages (iPSC, mesendo and defendo) and the time points of collection by design (day1, day2 and day3). We note that in the stage definition a number of cells ($\sim$20%) were excluded and considered to not belong to any well defined stage (see **page 110**), exclusively for the purposes of stage-level eQTL mapping. We included here an eQTL map of such transitioning population, which as expected was a very underpowered analysis. Additionally, we added results for day2 cells, for completeness. Shown are the number of genes that were considered for eQTL mapping, as well as the number of genes for which a eQTL was detected. *Samples are donor-experiment(-time point) combinations which were effectively used for testing (see **section 4.4**).

Antibody raised against	Catalog number	Company
Histone H3	ab1791	Abcam
Histone H3 (tri methyl K4)	ab8580	Abcam
Histone H3 (tri methyl K27)	C15200181 (MAb-181-050)	Diagenode
Histone H3 (mono methyl K4)	ab8895	Abcam
Histone H3 (acetyl K27)	ab4729	Abcam
Histone H3 (tri methyl K36)	ab9050	Abcam

Table A.3: Antibodies used for the ChIP-seq experiments.
Experimental methods for this analysis are described in **section C.1.3**, and these data were used for analysis shown in **Fig. 4.12** and **4.15**.

A.3 Additional information for Chapter 5

trait	category	study	n	n (replication)
Alzheimer's Disease (late onset)	neurodegenerative	[569]	55,134	19,884
Parkinson's Disease	neurodegenerative	[570]	442,271	-
Schizophrenia	neurodevelopmental	[571]	150,064	-
Bipolar disorder	neurodevelopmental	[572]	34,950	5,305
Neuroticism	personality	[573]	170,911	-
Neuroticism	personality	[574]	329,821	122,867
Neuroticism	personality	[575]	168,105	-
Depression (broad)	behavioural	[576]	322,580	-
Educational attainment	intelligence	[577]	111,114	-
Paternal history of Alzheimer's disease	neurodegenerative	[578]	260,279	-
Family history of Alzheimer's disease	neurodegenerative	[578]	314,278	-
Maternal history of Alzheimer's disease	neurodegenerative	[578]	288,676	-
Depressed affect	behavioural	[579]	357,957	-
Cognitive performance	intelligence	[580]	257,841	-
Sleeplessness / insomnia	personality	[581]	360,738	-
Nervous feelings	behavioural	[581]	351,829	-
Worrier / anxious feelings	behavioural	[581]	351,833	-
Tense / 'highly strung'	behavioural	[581]	350,159	-
Suffer from 'nerves'	behavioural	[581]	348,082	-
Neuroticism score	personality	[581]	293,006	-
Intelligence questions[1]	intelligence	[581]	117,131	-
Risk taking	behavioural	[581]	348,549	-
College or University degree	intelligence	[581]	357,549	-
A levels/AS levels or equivalent	intelligence	[581]	357,549	-
Other professional qualifications	intelligence	[581]	357,549	-

Table A.4: Neurological traits used for the colocalisation analysis.
Table compiled by Natsuhiko Kumasaka. Traits used for the colocalisation analysis in **Chapter 5** (**section 5.6**). The corresponding publication and sample size (both for the initial study and, when available for a replication study) are indicated. Additionally, traits are divided into broad categories.

[1] Number of fluid intelligence questions attempted within time limit.

Appendix B

Supplementary Figures

B.1 Additional results for Chapter 3

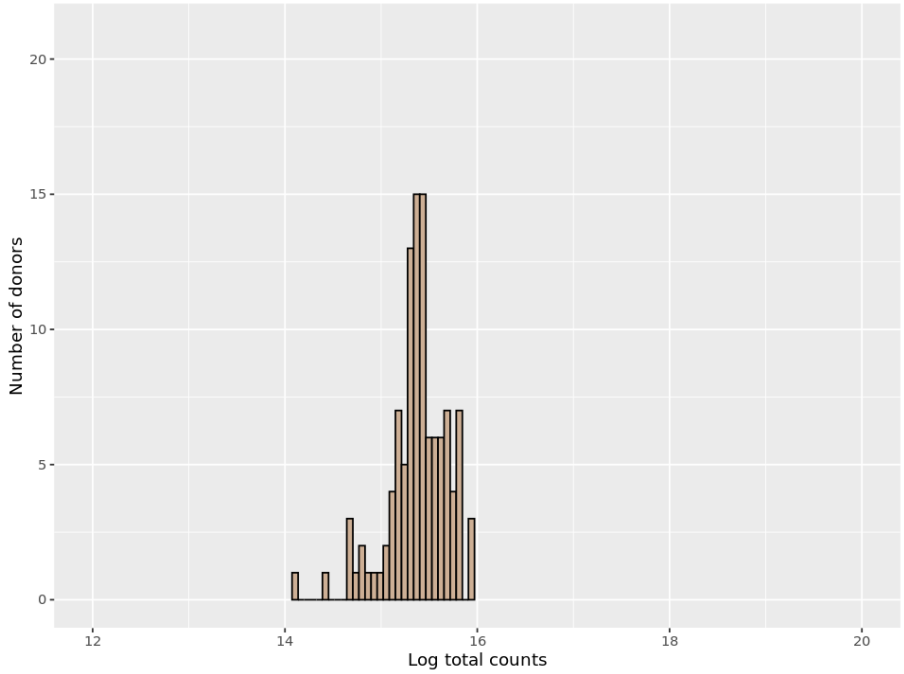


Fig. B.1: Distribution of total reads from a scRNA-seq dataset, when considering the same number of cells for each donor. Same as the left panel from **Fig. 3.4**, but downsampling to the same number of cells (n=10) from each donor.

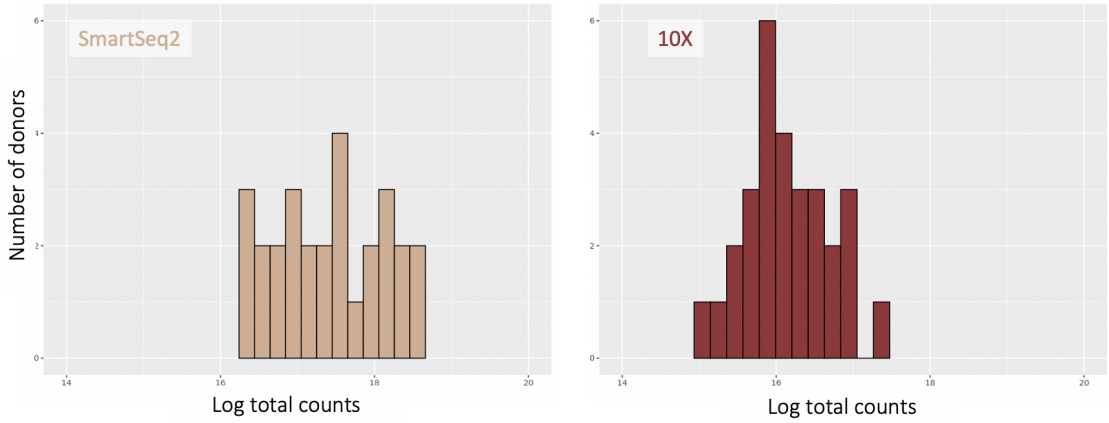


Fig. B.2: Distribution of total reads between scRNA-seq technologies, when considering the same 29 lines. Similar to **Fig. 3.4**, but considering SmartSeq2 and 10x data from the same 29 cell lines.

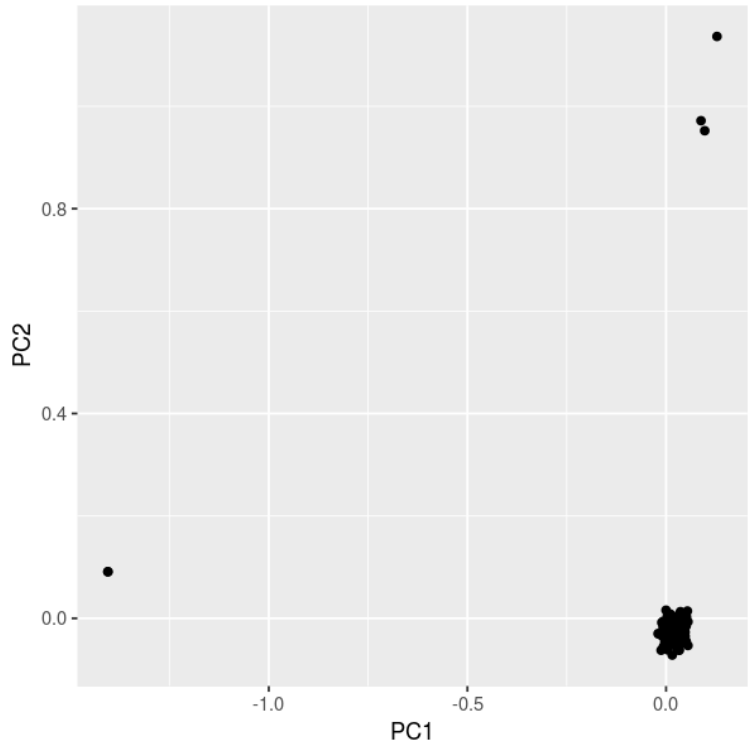


Fig. B.3: Population structure of donors included in the study.
Principal component (PC) decomposition of the kinship matrix (calculated using PLINK [355]) across all cell lines included in our study. The four outlier cell lines were excluded from the analyses described in **section 3.8**.

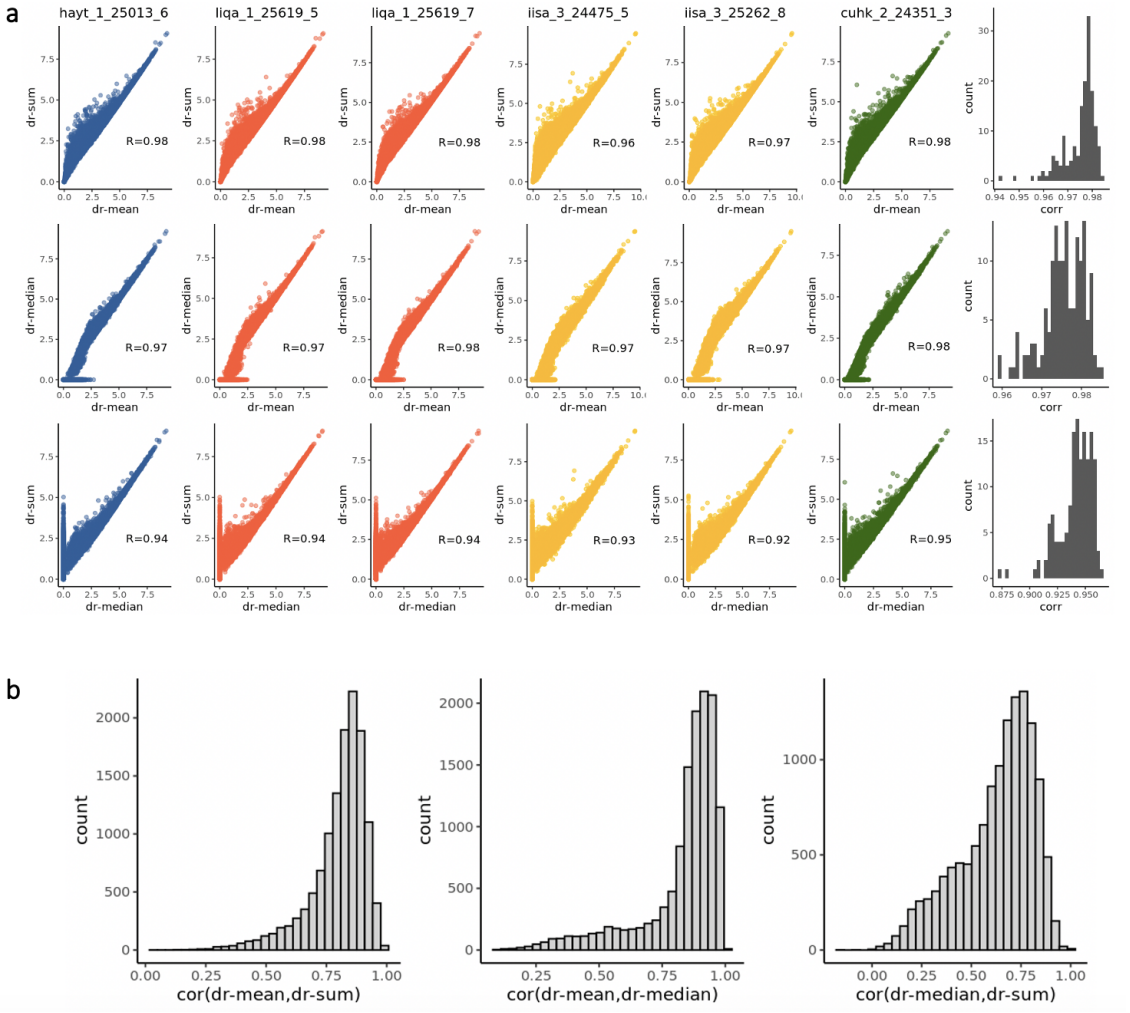


Fig. B.4: Comparison of 'dr' aggregated measures.
(a) For a random selection of 4 donors (two of which present in two sequencing runs, resulting in 6 donor-run combinations, or samples), scatter plots between aggregation metrics, across the set of common genes (n=12,720). First row is dr-mean vs dr-sum, second dr-mean vs dr-median, third dr-median vs dr-sum. The last column represents, for each of the comparisons a histogram of the distribution of correlations, across donors. (b) Histograms representing the distribution of correlations across donors, for each genes, for the same three comparisons as in (a).

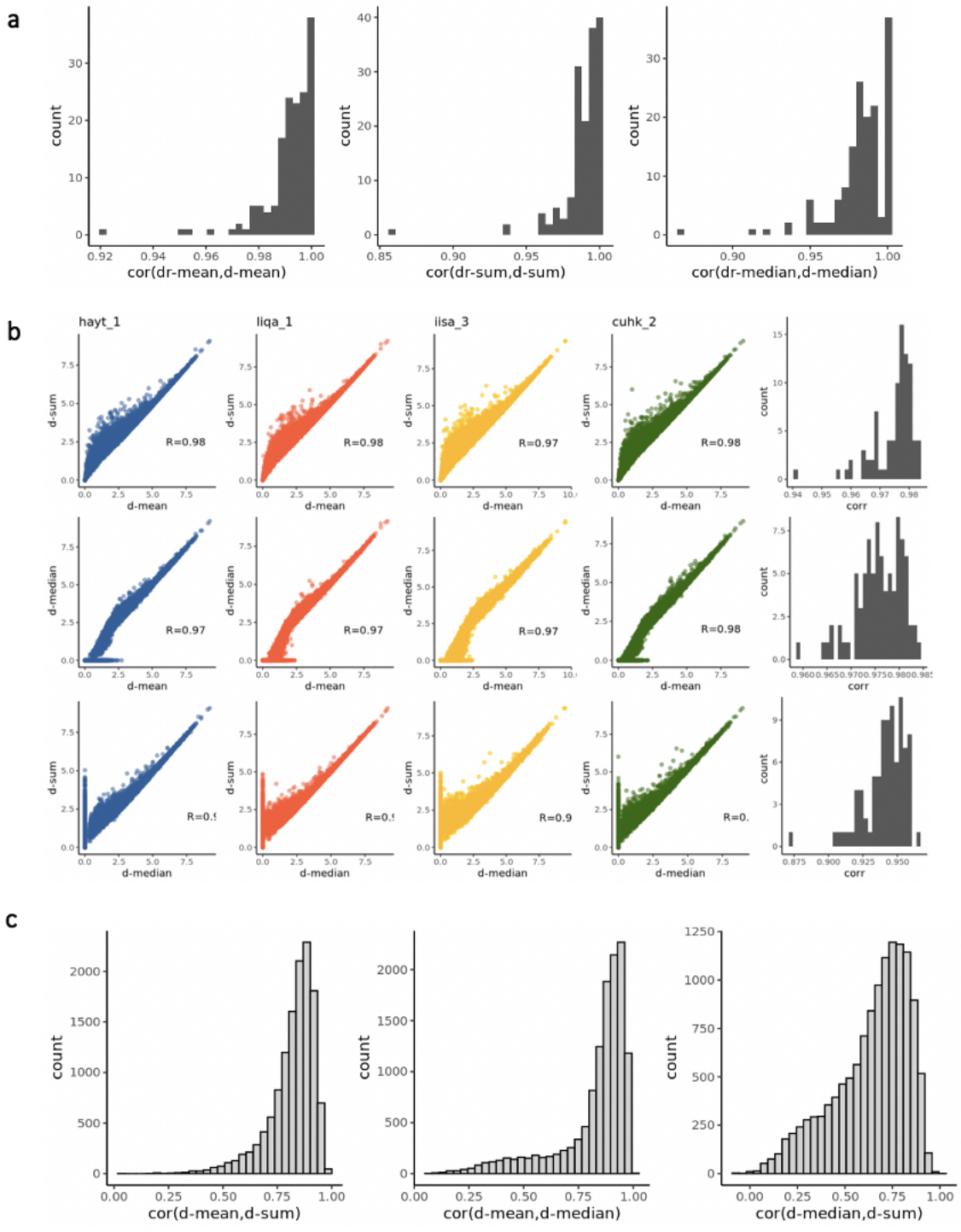


Fig. B.5: Comparison of 'd' aggregated measures.
(a) Histograms of correlations between 'dr' and 'd' aggregation measures, for each of mean, sum, median. (b,c) Similar to **Fig. B.4**, panels a and b, but across 'd' aggregation methods (instead of 'dr'; the same donors are considered).

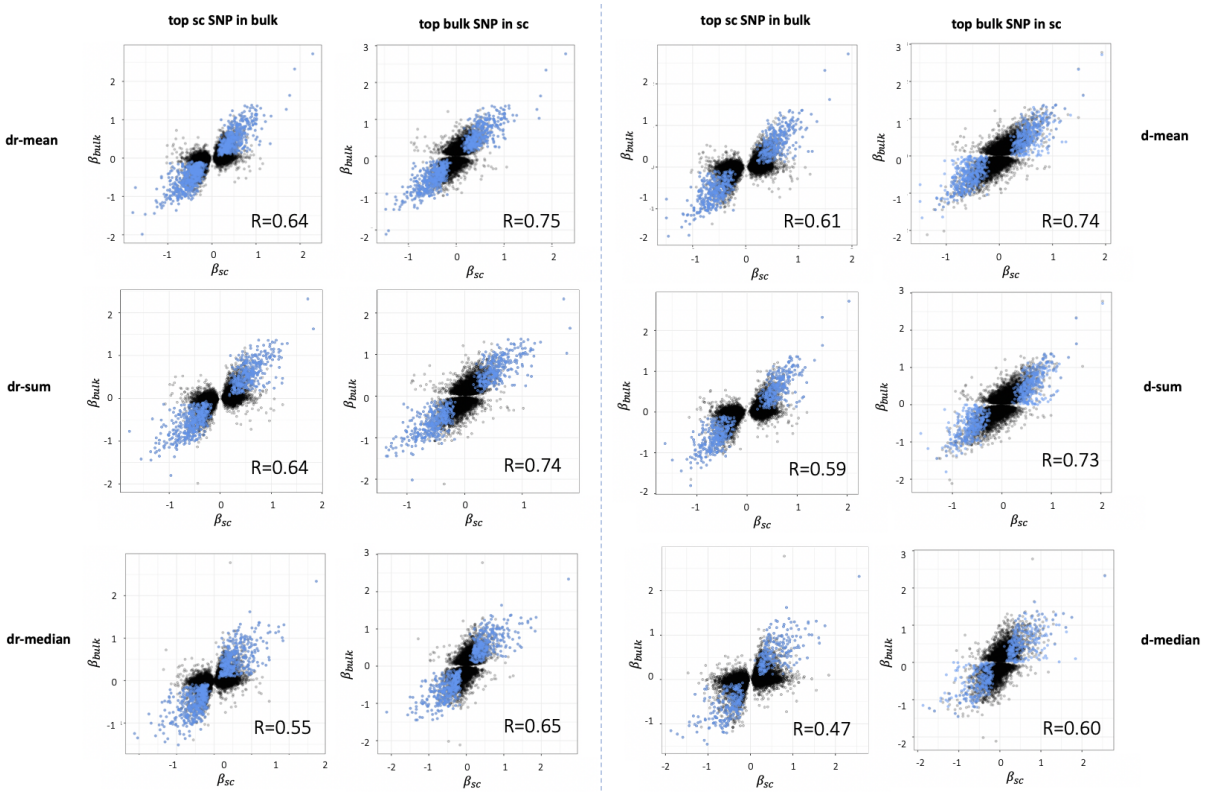


Fig. B.6: Comparison of results between single cell and bulk iPSC eQTL.
Scatter plots of eQTL effect sizes obtained when testing association of iPSC eQTL discovered using single cell (SmartSeq2) and bulk RNA-seq. The bulk eQTL results considered are obtained using all samples available ('a-bulk') The same set of n=12,720 genes that were assessed in all maps are considered. First two columns, 'dr'-aggregations (dr-mean, dr-sum, dr-median; aggregated at donor and sequencing run). Left, top SNP per gene (such that there are exactly as many points as there are genes) from the single cell data ('discovery set'; x axis) in bulk results ('replication set'; y axis). Right, vice versa (i.e. top SNP per gene in bulk results (now 'discovery set' y axis), in single cell results, now 'replication set', x axis). In blue, for each pair of sets of results are the 'replicated' eQTL, i.e. eQTL at FDR<5% in the discovery set, replicated in the replication set (FDR<10% and same direction of effect; consistently with the results presented in **section 3.8**). Third and fourth columns, same as first two but for 'd' methods, i.e. d-mean, d-sum, d-median (aggregated at donor level).

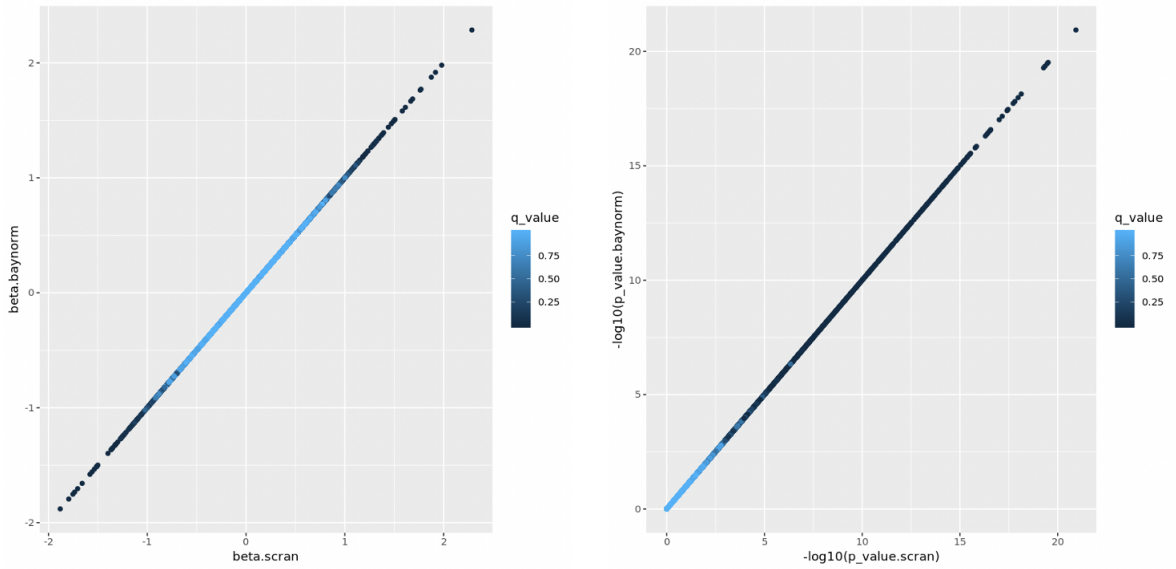


Fig. B.7: Correlation of results between scran and bayNorm normalisation.
Related to results from **Table 3.2**. Scatter plots of eQTL effect sizes (left) and p values (right) obtained when testing association of iPSC single cell eQTL discovered using dr-mean as aggregation method and after normalising counts using two alternative methods: scran/scater [438] (x axis) and bayNorm [454] (y axis).

B.2 Additional results for Chapter 4

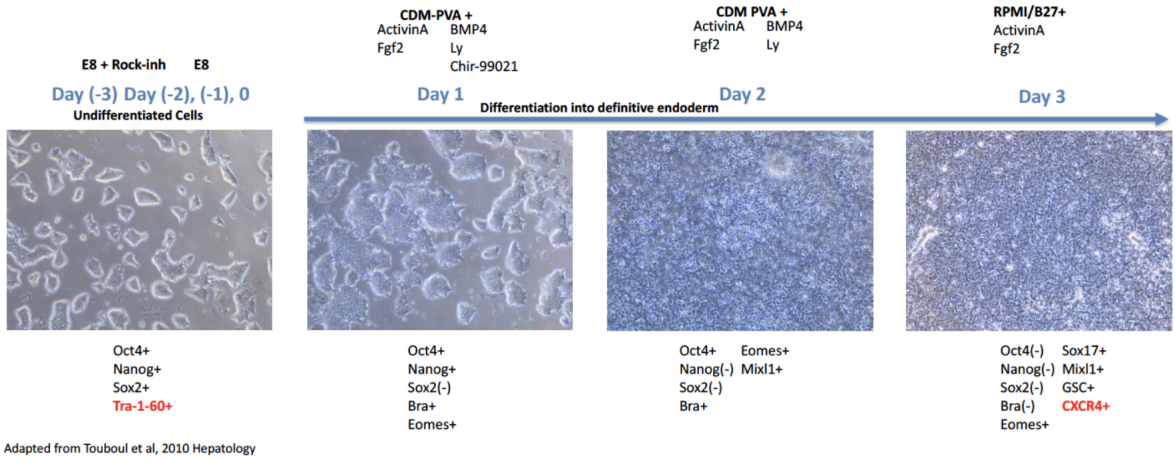


Fig. B.8: Endoderm differentiation protocol.
Adapted from Touboul *et al.* [582]. Related to **Chapter 4, section 4.2**. Schematic representation of the chemically defined protocol used to initiate differentiation towards definitive endoderm. Tra-1-60 and CXCR4 are canonical cell surface markers of pluripotent and definitive endoderm stages, respectively, which were used to sort live cells by differentiation stage.

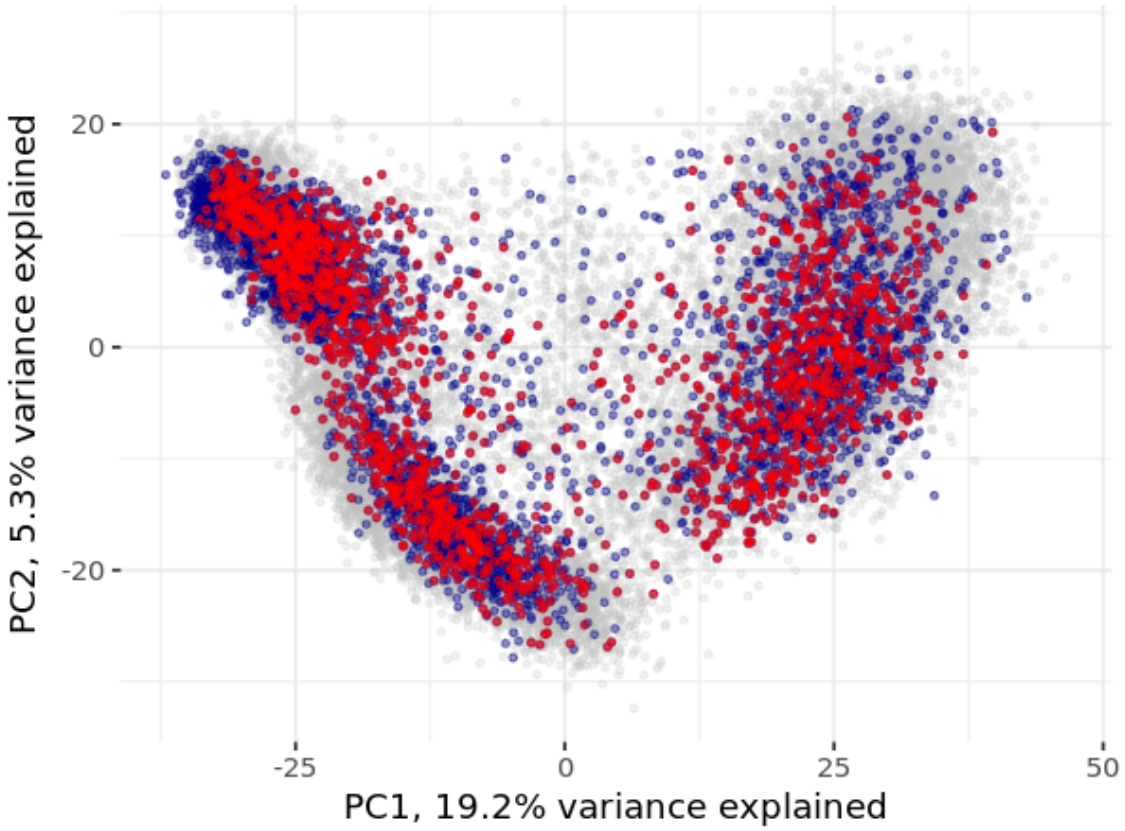


Fig. B.9: Comparison of expression patterns between healthy and diseased cell lines.
PCA plot (similar to **Fig. 4.8**) of cells from cell lines derived from monogenic diabetes donors (red), cells from healthy donors from the same seven experiments (dark blue), against the background of all cells (grey).

B.3 Additional results for Chapter 5

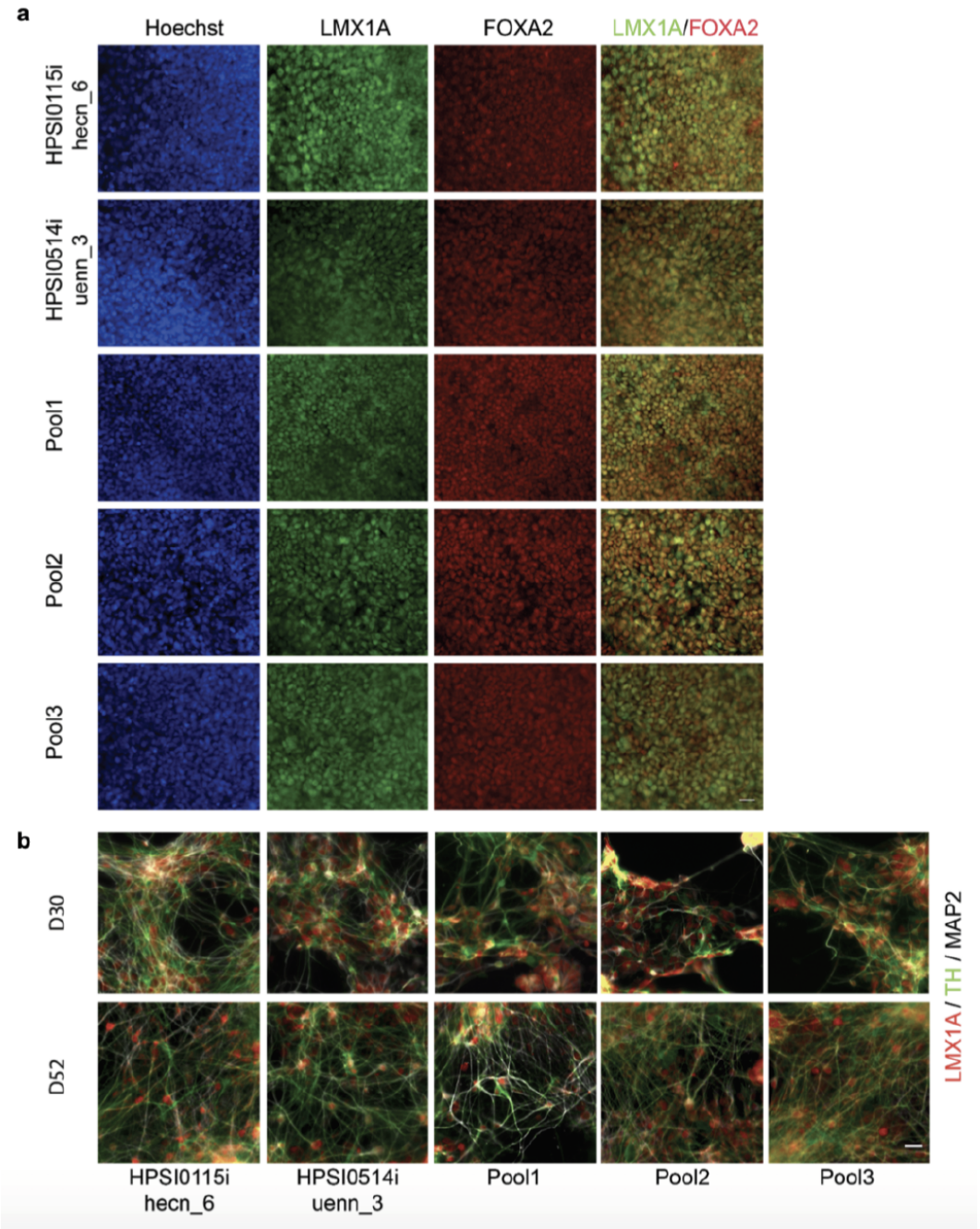


Fig. B.10: Immunostaining of midbrain neural progenitors and DA neurons (Full legend on next page).

Fig. B.10: Immunostaining of midbrain neural progenitors and DA neurons (continued).
Figure by Julie Jerber. (a) Immunostaining for known midbrain progenitor markers LMX1A and FOXA2 at day 11. Nuclei were counterstained with Hoechst. Scale bar: 25µm. (b) Immunostaining of differentiated dopaminergic neurons for the neuronal marker MAPT2 (white) and the dopaminergic neuronal markers TH and LMX1A. Scale bar: 25µm. Data is shown for two example individual cell lines (HPSI0155i-hecn_6 and HPSI0514i-uenn_3) as well as three entire differentiation pools (Pools 1,2,3).

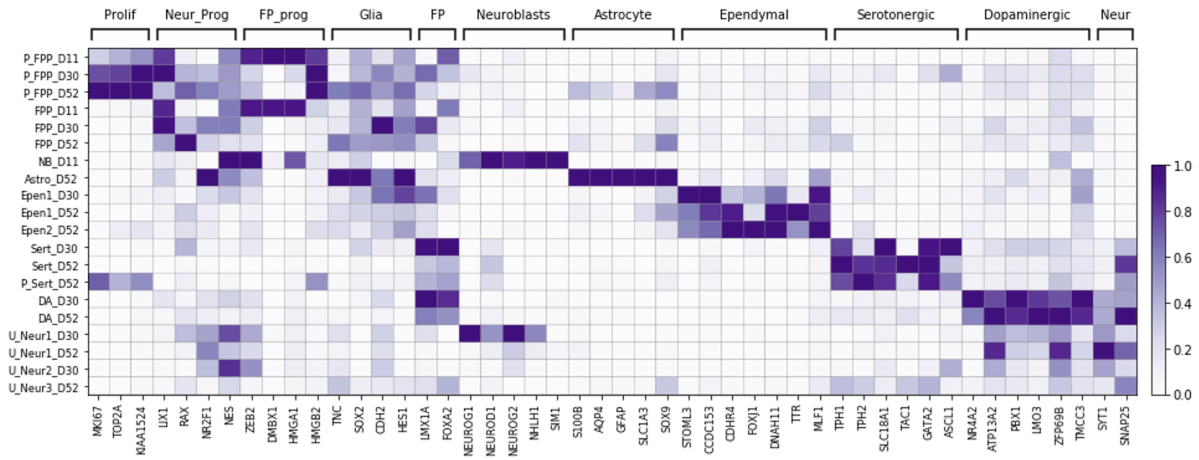


Fig. B.11: Neuronal cell type markers.
Heat map showing the average expression of the 48 literature-curated neuronal marker genes used to annotate the identified clusters as cell types (columns) by the annotated cell types at different time points (rows, as in **Fig. 5.2**).

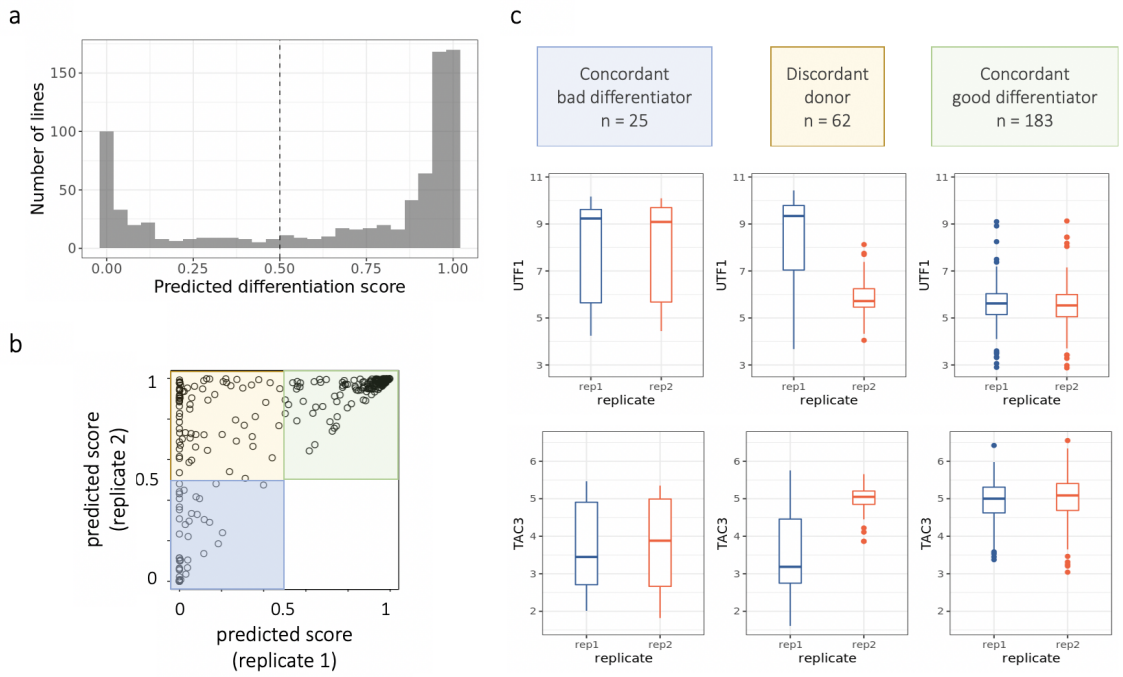


Fig. B.12: Predicted differentiation scores.
Results related to **section 5.4.1**. (a) Histogram of predicted differentiation scores across all HipSci cell lines. The bimodal distribution is especially extreme in this case, and 0.5 was used as the threshold to split bad from good differentiator lines. (b) Scatter plot of predicted differentiation scores for donor for which we have data for two different cell lines. Replicate1 (rep1) is chosen as the line with lower predicted score (n=270). Colours indicate three categories of donors, according to whether both lines from the same donor are predicted to fail differentiation (blue), both are predicted to succeed (green), the two lines are discordant (one is predicted to successfully differentiate, but not the other, yellow). (c) Bulk RNA-seq expression of *UTF1, TAC3* for the two replicate lines per donor stratified by the categorization described in (b).

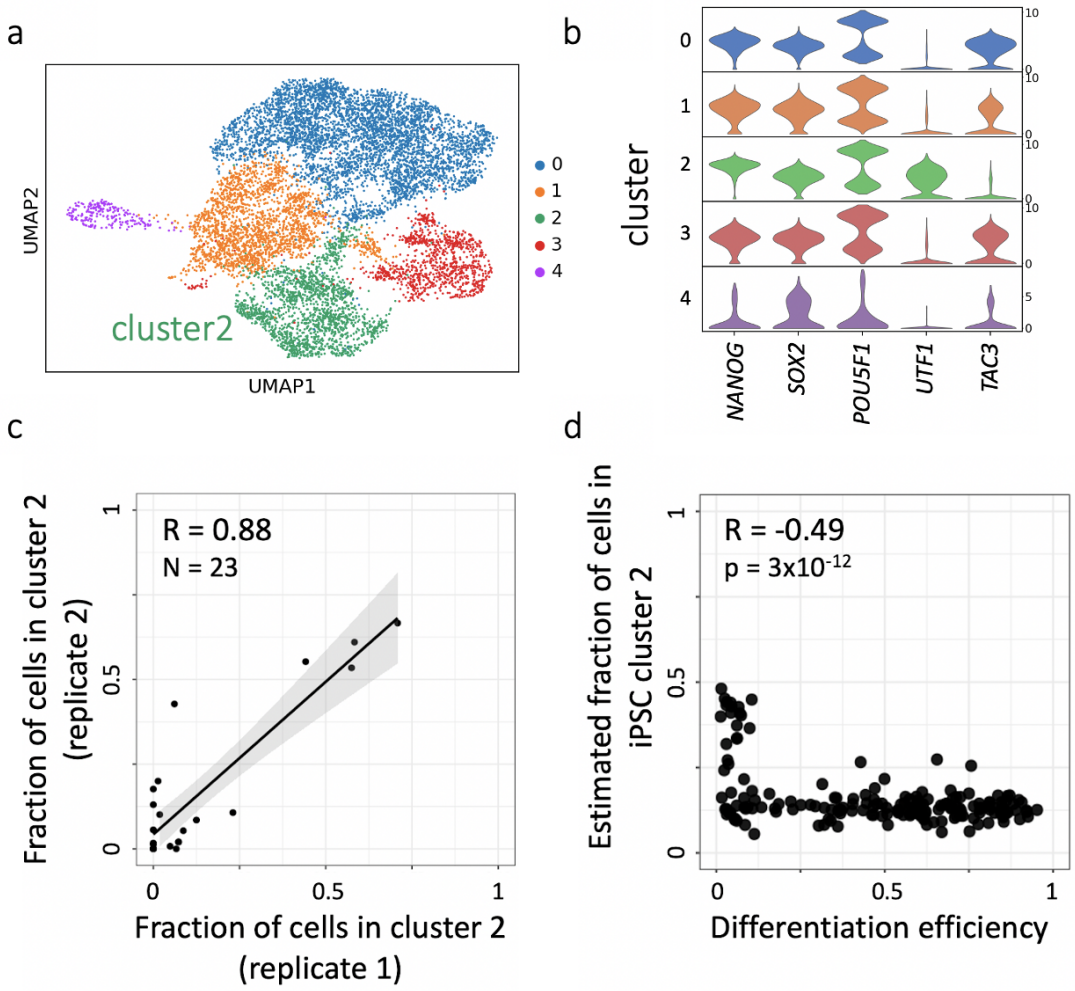


Fig. B.13: Re-analysis of iPSC scRNA-seq data reveals a subpopulation characterised by expression of predictive marker genes associated with lower neuronal differentiation efficiency. (a) UMAP overview of the dataset. iPSC scRNA-seq data from [447] were re-analysed following the same batch correction and clustering steps applied to our neural differentiation data, identifying 5 clusters. (b) Violin plots of gene expression for genes related to pluripotency (*NANOG, SOX2, POU5F1*) and two gene markers that are respectively upregulated and downregulated in cluster 2 (*UTF1, TAC3*, from **Fig. 5.9**). (c) Scatter plot showing the proportion of cells assigned to cluster 2 between replicates (n=23). (d) Scatter plot between the proportion of cells assigned to cluster 2 (y axis) and differentiation efficiency (x axis) similar to **Fig 5.11**, panel c, but where we use imputed proportions of cluster 2 cells from bulk RNA-seq available for most cell lines (n=182, out of 199 lines for which we have day 52 data and thus a measure of neuronal differentiation efficiency).

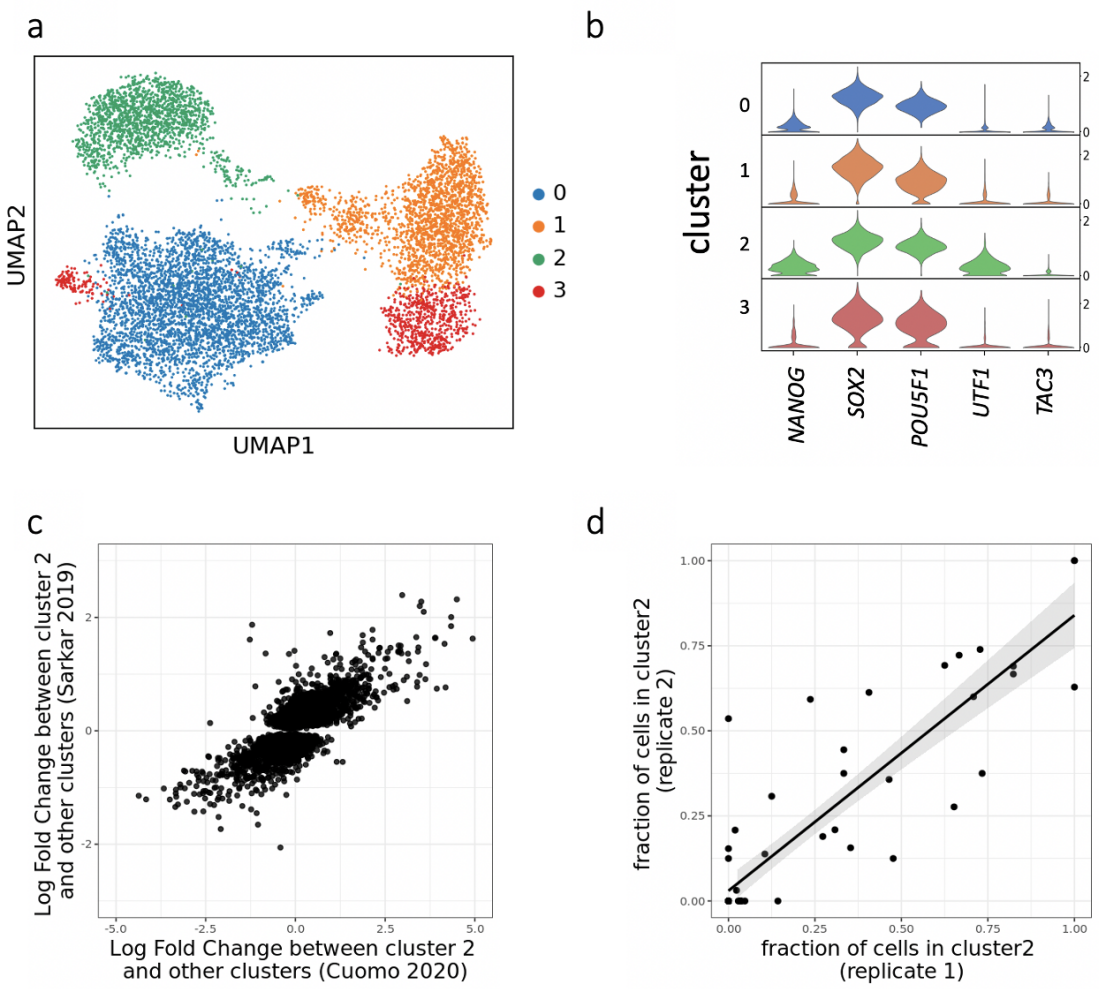


Fig. B.14: Analysis of a single cell iPSC dataset from Sarkar *et al.*
Similar to **Fig. B.13**. (a) UMAP overview of the dataset. iPSC scRNA-seq data from [444] were re-analysed following the same data normalisation and clustering steps applied to our neural differentiation data, identifying 4 clusters. (b) Violin plots of gene expression for genes related to pluripotency (*NANOG, SOX2, POU5F1*) and a subset of markers of the cluster 2 population (*UTF1, TAC3*). (c) Scatter plot showing the proportion of cells assigned to cluster 2 between replicates (n=59). (d) Expression log fold change between cluster 2 and all other clusters from [447] compared to the same between cluster 2 and the rest from [444]. Shown are all 5,397 DE genes between cluster 2 and all other clusters from Sarkar *et al.* (FDR < 0.05).

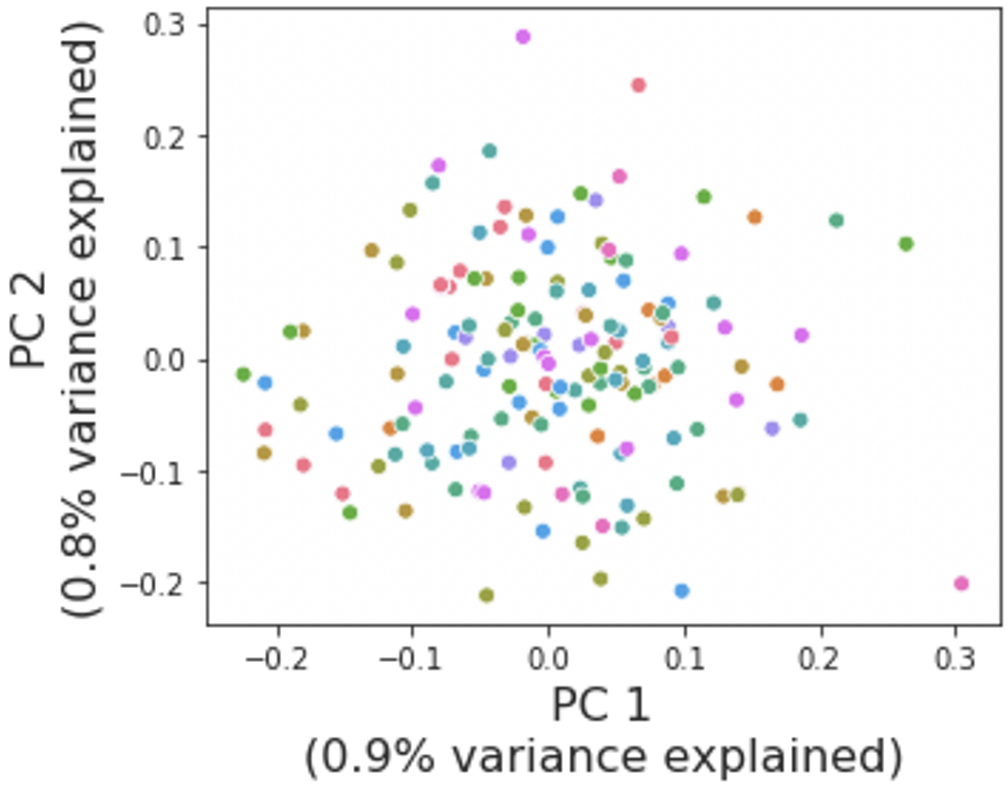


Fig. B.15: Absence of population structure.
Principal component (PC) decomposition of the kinship matrix (calculated using PLINK [355]) across all cell lines included in the study described in **Chapter 5**, coloured by batch.

Appendix C

Experimental Methods

C.1 Experimental methods for Chapter 4

Experimental methods compiled by Mariya Chhatriwala, as in [447].

C.1.1 Cell culture for maintenance and differentiation

Human iPSC lines were thawed for differentiation and maintained in Essential 8 (E8) media (LifeTech) on vitronectin (StemCell Technologies, #07180) coated Corning plates according to the manufacturer's instructions. Cells were passaged at least twice after thawing and always 3 - 4 days before plating for differentiation to ensure all the cell lines in each experiment were growing at a similar rate prior to differentiation. Gelatine/MEF coated plates were prepared 24 – 48 hours before plating for differentiation by incubating plates with 0.1% gelatine for 20 minutes at room temp. The gelatine was then aspirated and plates were incubated in MEF medium overnight at 37°C. Immediately prior to plating cells, plates were washed once with D-PBS to remove any residual MEF medium. To plate for endoderm differentiation, cells were washed once with D-PBS and dissociated using StemPro Accutase (Life Technologies, A1110501) at 37°C for 3 - 5 min. Colonies were fully dissociated through gentle pipetting. Cells were resuspended in MEF medium, passed through a 40 μm cell strainer, and pelleted gently by centrifuging at 300 x g for 5 min. Cells were re-suspended in E8 media and plated at a density of 15,000 cells per cm2 on gelatin/MEF coated plates [480, 583] in the presence of 10 μM Rock inhibitor – Y27632 (Sigma, #Y0503 - 5 mg). Media was replaced with fresh E8 free of Rock inhibitor every 24 hours post plating. Differentiation into definitive endoderm commenced 72 hours post plating. Cells were washed 1x gently with D-PBS to remove residual E8. Cells were then incubated in CDM-PVA containing 100 ng/mL ActivinA (made in house), 80 ng/mL FGF2 (made in house), 10 ng/mL BMP4

(R&D systems, #314-BP-050), 10 µm Ly294002 (Promega, #V1201), and 3 µM CHIR99201 (Selleckchem, #S1263) for 24 hours (day 1). After 24 hours, the day 1 media was replaced with CDM-PVA containing 100 ng/mL ActivinA, 80 ng/mL FGF2, 10 ng/mL BMP4, and 10 µm Ly294002 for another 24 hours (day 2). Day 2 media was then replaced with RPMI/B27 containing 100 ng/mL ActivinA and 80 ng/mL FGF2 for another 24 hours (day 3) [480]. The overall efficiency of the differentiation protocol was validated using reference lines with good and poor differentiation capacity, respectively. All media was filtered through 0.22 µm filters prior to use.

C.1.2 Single cell preparation and sorting for scRNAseq

Cells were dissociated into single cells using Accutase and washed once with MEF medium as described above when plating cells for differentiation. For all subsequent steps, cells were kept on ice to avoid degradation. Approximately 1 x 106 cells were re-suspended in PBS + 2% BSA + 2 mM EDTA (FACS buffer); BSA and PBS were nuclease-free. For staining of cell surface markers, 1 x 106 cells were re-suspended in 100 µL of ice-cold FACS buffer containing 20 µL anti-Tra-1-60 antibody (BD Pharmingen, BD560380) and 5 µL of anti-CXCR4 antibody (eBioscience 12-9999-42), and were placed on ice for 30 min. Cells were protected from light during staining and all subsequent steps. Cells were washed with 5 mL of FACS buffer, passed through a 35 µm filter to remove clumps, and re-suspended in 300 µL of FACS buffer for live cell sorting on the BD Influx Cell Sorter (BD Biosciences). Live/dead marker 7AAD (eBioscience 00-6993) was added immediately prior to analysis at a concentration of 2 µL/mL and only living cells were considered when determining differentiation capacities. Living cells stained with 7AAD but not TRA-1-60 or CXCR4 were used as gating controls. Data for TRA-1-60 and CXCR4 staining were available for 31,724 cells, of the total 36,044. Single-cell transcriptomes of sorted cells were assayed as follows: reverse transcription and cDNA amplification was performed according to the SmartSeq2 protocol [383], and library preparation was performed using an Illumina Nextera kit. Samples were sequenced using paired-end 75bp reads on an Illumina HiSeq 2500 machine (one lane of sequencing per 384 well plate).

C.1.3 ChIP-seq experiments and data processing

ChIP-seq was performed using FUCCI-Human Embryonic Stem Cells (FUCCI-hESCs, H9 from WiCell) in a modified endoderm differentiation protocol to that used for the iPSC differentiations (see details below). Cells were grown in defined culture conditions as described previously [584]. Pluripotent cells were maintained in Chemically Defined Media

with BSA (CDM-BSA) supplemented with 10ng/ml recombinant Activin A and 12ng/ml recombinant FGF2 (both from Dr. Marko Hyvonen, Dept. of Biochemistry, University of Cambridge) on 0.1% Gelatin and MEF media coated plates. Cells were passaged every 4-6 days with collagenase IV as clumps of 50-100 cells. The culture media was replaced 48 hours after the split and then every 24 hours.

The generation of FUCCI-hESC lines has been described in [585] and are based on the FUCCI system described in [586]. hESCs were differentiated into endoderm as previously described [587]. Following FACS sorting, Early G1 (EG1) cells were collected and immediately placed into the endoderm differentiation media and time-points were collected every 24h up to 72h. Endoderm specification was performed in CDM with Polyvynilic acid (CDM-PVA) supplemented with 20ng/ml FGF2, 10μM Ly-294002 (Promega), 100ng/ml Activin A, and 10ng/ml BMP4 (R&D).

We performed ChIP as described previously [588]. For ChIP-sequencing, ChIP for various histone marks (H3K4me3, H3K27me3, H3K4me1, H3K27ac, H3K36me3) (see **Table A.3** for antibodies) was performed on two biological replicates per condition. At the end of the ChIP protocol, fragments between 100bp and 400bp were used to prepare barcoded sequencing libraries. 10ng of input material for each condition were also used for library preparation and later used as a control during peak calls. The libraries were generated by performing 8 PCR cycles for all samples. Equimolar amounts of each library were pooled and this multiplexed library was diluted to 8pM before sequencing using an Illumina HiSeq 2000 with 75bp paired-end reads.

Reads were mapped to GRCh38 reference assembly using BWA [589]. Only reads with mapping quality score $\geq$ 10 and aligned to autosomal and sex chromosomes were kept for further processing. Peak calling analysis [590] was performed using PeakRanger [591], and only the peaks that were reproducible at an FDR of $\leq$ 0.05 in two biological replicates were selected for further processing. Peak calling was done using appropriate controls with the tool peakranger 1.18 in modes ranger (H3K4me3, H3K27ac; '-l 316 -b 200 -q 0.05'), ccat (H3K27me3; '-l 316 –win_size 1000 –win_step 100 –min_count 70 –min_score 7 -q 0.05') and bcp (H3K4me1, H3K36me3; '-l 316'). Adjacent peak regions closer than 40 bp were merged using the BEDTools suite [592], and those overlapping ENCODE blacklisted regions were filtered out (ENCODE Excludable Mappability Regions [593]). Finally, peaks were converted to GRCh37 coordinates using UCSC LiftOver.

C.2 Experimental methods for Chapter 5

Experimental methods written by Julie Jerber, as in [445].

C.2.1 Human iPSC culture

Feeder-free human iPSCs were obtained from the HipSci project [294]. Lines were thawed onto tissue culture-treated plates (Corning, 3516) coated with 10 µg/mL VitronectinXF (StemCell Technologies, 07180) using complete Essential 8 (E8) medium (Thermo Fisher, A1517001) and 10 µM Rock inhibitor (Sigma, Y0503). Cells were expanded in E8 medium for 2 passages using 0.5 µM EDTA pH 8.0 (Thermo Fisher, 15575-020) for cell dissociation.

C.2.2 Pooling and differentiation of midbrain dopaminergic neurons

iPSC colonies were dissociated into a single-cell suspension using Accutase (Thermo Fisher, A11105-01) and resuspended in E8 medium containing 10 µM Rock inhibitor. Cells were counted using an automated cell counter (Chemometec NC-200) and a cell suspension containing an equal amount of each iPSC line was prepared in E8 medium containing 10 µM Rock inhibitor and seeded at 2 x 105 cells per cm2 on 0.01% Geltrex- (Thermo Fisher, A1413202) coated plates. Each pool of lines contained between 7 to 24 donors. 24h after plating, neuronal differentiation of the pooled lines to a midbrain lineage was performed as described by12 with minor modifications: 1. SHH C25II was replaced by 100nM SAG (Tocris, 6390) in the neuronal induction phase. 2. On day 20, the cells were passaged with Accutase containing 20 units/mL of papain (Worthington, LK00031765) and plated at 3.5 x 105 cells per cm2 on 0.01% Geltrex-coated plates for final maturation.

C.2.3 Rotenone stimulation

On day 51 of differentiation, the cells were exposed for 24h to freshly prepared 0.1 µM rotenone (Sigma, R8875, purity HPLC $\geq$ 95%) diluted in neuronal maturation medium [284]. The final DMSO concentration was 0.01% in all exposure conditions. Unstimulated control samples (i.e. DMSO only) were taken concurrently.

C.2.4 Generation of cerebral organoids

Cerebral organoids were generated according to the enCOR method as previously described by [530]. Briefly, one pool of 18 iPSC lines was thawed and expanded for 1 passage before seeding 18,000 cells onto PLGA microfilaments prepared from Vicryl sutures. STEMdiff

Cerebral Organoid kit (Stem Cell Technologies, 08570) was used for organoid culture with timing according to manufacturer's suggestion and Matrigel embedding as previously described57. From day 35 onward the medium was supplemented with 2% dissolved Matrigel basement membrane (Corning, 354234), and processed for scRNA-seq after 113 days of culture.

C.2.5 Generation of single cell suspensions for sequencing

On harvesting days, the cells were washed once with 1X DPBS (Thermo Fisher, 14190-144) before adding either Accutase (day 11) or Accutase containing 20 units/mL of papain (days 30 and 52). The cells were incubated at 37°C for up to 20 min (day 11) or up to 35 min (days 30 and 52) before adding DMEM:F12 (Thermo Fisher Scientific, 10565-018) supplemented with 10 µM Rock inhibitor and 33 µg/mL DNase I (Worthington, LK003170, only for days 30 and 52). The cells were dissociated using a P1000 and collected in a 15 mL tube capped with a 40 µm cell strainer. After centrifugation, the cells were resuspended in 1X DPBS containing 0.04% BSA (Sigma, A0281) and washed 3 additional times in 1X DPBS containing 0.04% BSA. Single-cell suspensions were counted using an automated cell counter (Chemometec NC-200) and concentrations adjusted to 5 x 105 cells/mL.

Organoids were washed twice in 1X DPBS before adding EBSS (Worthington, LK003188) dissociation buffer containing 19 U/mL of papain, 50 µg/mL of DNase I and 22.5X of Accutase. Organoids were incubated in a shaking block (750 rpm) at 37°C for 30 min. Every 10 min, the organoids were triturated using a P1000 and BSA-coated pipette tips until large clumps were dissociated. Dissociated organoids were transferred into a new tube capped with a 40 µm cell strainer and pelleted for 4 min at 300g. After centrifugation, the cells were resuspended in EBSS containing 50µg/mL of DNase I and 2 mg/mL ovomucoid (Worthington, LK003150). 0.5 volume of EBSS, followed by 0.5 volume of 20 mg/mL ovomucoid were added to the top of the cell suspension and the cells were mixed by flicking the tube. After centrifugation, the cells were resuspended in 1X DPBS containing 0.04% BSA. Single-cell suspensions were counted using an automated cell counter and concentrations adjusted to 5 x105 cells/mL.

C.2.6 Immunohistochemistry

Cells were fixed in 4% paraformaldehyde (Thermo Fisher Scientific, 28908) for 15 min, rinsed 3 times with PBS1X (Sigma, D8662) and blocked with 5% normal donkey serum (NDS; AbD Serotec, C06SBZ) in PBST (PBS1X + 0.1% Triton X-100, Sigma, 93420) for

2h at room temperature. Primary antibodies were diluted in PBST containing 1% NDS and incubated overnight at 4°C. Cells were washed 5 times with PBS1X and incubated with secondary antibodies diluted in PBS1X for 45 min at room temperature. Cells were washed 3 more times with PBS1X and Hoechst (Thermo Fisher Scientific, H3569) was used to visualize cell nuclei. Image acquisition was performed using Cellomics array scan VTI (Thermo Fisher Scientific). The following antibodies were used: FOXA2 (Santa Cruz, sc101060 - 1/100) LMX1A (Millipore, AB10533 - 1/500) TH (Santa Cruz, sc-25269 - 1/200) MAP2 (Abcam, 5392 - 1/2000) Donkey anti-chicken AF647 (Thermo Fisher Scientific, A21449) Donkey anti-mouse AF488 (Thermo Fisher Scientific, A11008) Donkey anti-mouse AF555 (Thermo Fisher Scientific, A31570) Donkey anti-rabbit AF488 (Thermo Fisher Scientific, A21206) Donkey anti-rabbit AF555 (Thermo Fisher Scientific, A27039)

C.2.7 Chromium 10x Genomics library and sequencing

Single cell suspensions were processed by the Chromium Controller (10x Genomics) using Chromium Single Cell 3' Reagent Kit v2 (PN-120237). On average, 15,000 cells from each 10x reaction were directly loaded into one inlet of the 10x Genomics chip. All the steps were performed according to the manufacturer's specifications. Barcoded libraries were sequenced using HiSeq4000 (Illumina, one lane per 10x chip position) with 50bp or 75bp paired end reads to an average depth of 40,000-60,000 reads per cell.

Appendix D

List of Publications

D.1 Published papers

Anna S.E. Cuomo*, Daniel D. Seaton*, Davis J. McCarthy*, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, Andrew Knights, Kedar Nath Natarajan, the Hipsci Consortium, Ludovic Vallier, John C. Marioni, Mariya Chhatriwala, Oliver Stegle. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, 2020 (* equal contribution).

Julie Jerber*, Daniel D. Seaton*, Anna S.E. Cuomo*, Natsuhiko Kumasaka, James Haldane, Juliette Steer, M Patel, D Pearce, M Andersson, Marc Jan Bonder, Ed Mountjoy, Maya Ghoussaini, Madeline A. Lancaster, the HipSci Consortium, John C. Marioni, Florian T. Merkle, Oliver Stegle, Daniel J. Gaffney. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature Genetics*, 2021 (* equal contributions).

Christoph Muus et al. (including Anna S.E. Cuomo), Single-cell meta-analysis of SARS-CoV-2 entry genes across tissues and demographics. *Nature Medicine*, 2021.

D.2 Accepted manuscripts

Ricard Argelaguet, Anna S.E. Cuomo, Oliver Stegle, John C. Marioni. Computational principles and challenges for the integration of single-cell multi-modal data. *Nature Biotechnology*, 2021, in press.

D.3 Submitted manuscripts

Anna S.E. Cuomo*, Giordano Alvari*, Christina B. Azodi*, single cell eQTLGen Consortium, Davis J. McCarthy, Marc Jan Bonder. Optimising expression quantitative trait locus mapping workflows for single-cell studies. *bioRxiv*, 2021 (* equal contributions).

D.4 Manuscripts in preparation

Anna S.E. Cuomo, Danilo Horta, Danai Vagiaki, John C. Marioni, Oliver Stegle. An integrated framework to map complex context-specific eQTL using single cell RNA-sequencing.