Chapter 1

Introduction

1.1 Human genetic variation

Over the last century, there has been an increasing effort to understand and map genetic variation between humans, and the consequent functional effect of this. This has been accelerated in recent decades with the development of microarray and sequencing technologies, and in particular of high-throughput genetic sequencing. Initiatives such as the Human Genome Project [1], 1000 Genomes Project [2], HapMap project [3], and most recently the 100,000 Genomes Project highlight efforts within the field to chart common variation and the increasing scale at which this is being achieved.

1.1.1 The basis of genetic variation

Genetic variation stems from alteration of DNA sequences, referred to as 'variants' or 'mutations'. These events can occur as a result of endogenous processes, such as errors in DNA replication, chromosome segregation and recombination, or as a result of damage from endogenous or exogenous chemicals (Fig 1.1, [4]). While processes involving chromosome and DNA function are highly regulated, they - like any cellular function

- are not 100% efficient. In the case of DNA replication, around $6x10^9$ nucleotides must be copied in each cell division. Although the major DNA polymerases involved in this DNA synthesis have intrinsic proofreading and exonuclease capacity, allowing the ability to detect and remove incorrectly inserted bases, this does not occur in a very small proportion of cases and errors are maintained. This is often the case at regions in the genome with repeat sequences. In these areas, where there are repeats of particular nucleotide or oligonucleotide sequences, replication slippage may occur, leading to insertion or deletion of nucleotides. On a larger scale, errors in chromosome segregation and recombination may lead to variation in copy number of substantial regions of DNA or entire chromosomes. In the germline, these events often lead to embryonic lethality or developmental disorders, however they can also occur in somatic cells - a typical occurrence in cancer development.

Alongside faults in the processes described above, chemical damage to DNA can cause mutations, deriving from both endogenous and exogenous sources (Figure 1.1). Given the aqueous environment within cells, hydrolytic damage is common. This can lead to the cleavage of covalent N-glycosylic bonds between a base and its sugar, producing an abasic site, or to the deamination of some bases to leave a carbonyl group. Further elements of the cellular milieu produced by normal metabolic reactions can lead to oxidative damage, in particular reactive oxygen species (ROS). The sugar-phosphate backbone can be damaged as a result, or DNA bases can be attacked leading to the production of derivatives, many of which are mutagenic. An alternate source of endogenous damage is the erroneous methylation of adenosine. This causes distortion of the double helix and disrupts DNA-protein interactions. While the majority of chemical damage derives from these intrinsic mechanisms, exogenous agents can also play a role. Examples are the production of ROS within cells due to ionizing radiation from external sources, leading to oxidative damage as described above. Non-ionizing

ultraviolet radiation can also cause damage, resulting from the covalent bonding between adjacent pyrimidines. Finally, environmental chemicals can covalently bond to and distort the DNA helix, such as the large aromatic hydrocarbons found in the smoke of cigarettes and vehicles.

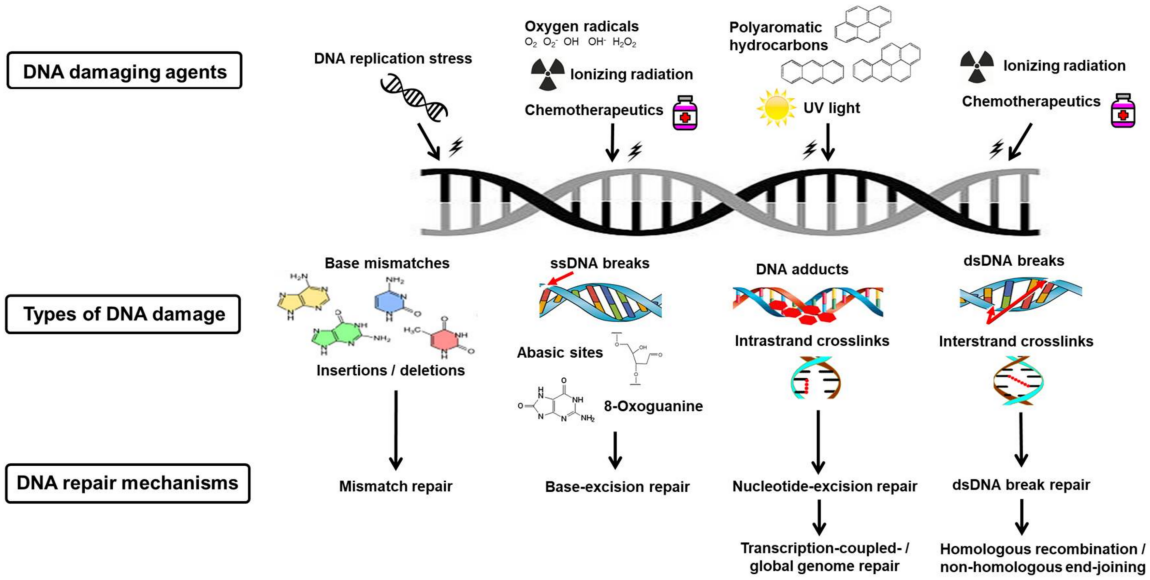

Fig. 1.1 Mechanisms of DNA damage and repair, from Helena *et al.* [4]

If not repaired, these changes in DNA sequence may have a wide range of consequences, or no discernible effect at all. A large proportion of variants do not have a functional effect for several reasons: firstly, much of the genome is non-coding and has no known function. Secondly, there is a high level of genetic redundancy, with substitutions at the third base in a codon sequence often producing the same amino acid (synonymous mutations), and also redundancy in the sequence as a whole - for example, there are hundreds of almost identical ribosomal RNA genes. Finally, even in situations where a variant in a coding gene results in a different amino acid produced (nonsynonymous), this may be functionally unimportant within the protein and therefore tolerated. Despite this, variants with phenotypic effects do arise, and while they may occasionally have a beneficial effect, and may even become positively

selected for in the population, in many cases the mutations are harmful. Variants which have become fixed in the population have been catalogued in dbSNP [5]. Many methods have been used over the years to study the effects of genetic variants, and a brief history and description of modern methods follows.

1.1.2 Approaches for studying human genetic variation

In the early decades of genetic research, familial history was used in genetic linkage studies. The basis of this approach is the increased frequency of co-inheritance of genetic markers close in genomic location than would be expected by chance. Huntington's disease was the first for which the locus - on chromosome 4 - was identified purely by linkage [6]. Following this, developments were made in mapping cystic fibrosis to chromosome 7 [7–10], While this method provided a lot of novel insight in disorders arising from a single gene and with high penetrance - the percentage of individuals with a given genotype who exhibit the associated phenotype - these approaches were more difficult for complex diseases arising from the combination of many low penetrance variants. With the evolution of technologies to assay genome sequences, however, it has become increasingly possible to understand the role of common genetic variants both in disease and healthy phenotypes.

Accelerated by these next-generation sequencing technologies, it has been possible to deeply characterise genetic variation in the population as a whole. This has been highlighted by large-scale international consortia such as the HapMap project [3] and 1000 Genomes Project [2]. The scale of these studies will continue to grow, exemplified by the 100,000 Genomes Project currently underway. This extensive work to map common genetic variation opened the door to genome wide association studies (GWAS).

The GWAS approach is to ask whether a particular variant appears more often in individuals with a phenotype of interest than expected by chance (Figure 1.2). It

is common to use a case-control set up, where two groups are compared: those with the disease/phenotype of interest, and controls without. An odds ratio is calculated, reflecting the odds of the variant in the two groups, with an OR > 1 signifying higher prevalence in the case group. The power to detect significant effects depends on the sample size, distribution of effect sizes of causal genetic variants, and the frequency of these in the population, and the linkage disequilibrium (LD) between the observed genotyped DNA variants and the unknown causal variants. GWAS approaches have also been applied to quantitative phenotypes, such as height or concentration of given biomarkers.

The first GWAS, published in 2005, focused on age-related macular degeneration (AMD). In a comparison of 96 cases and 50 controls, Klein *et al.* identified a role of the CFH gene in AMD [11]. A major breakthrough followed in 2007 with the publication of the Wellcome Trust Case Control Consortium [12], in which 3000 shared controls were compared with around 2000 patients for each of seven common disease phenotypes. Not only was this the largest study of its kind at the time, but it also set the precedent for future GWAS studies in a number of ways. Population stratification was carefully considered, HapMap data was used for genotype imputation in a novel manner [13], and significant attention was given to genotype calling.

Since then, the number of GWAS has increased year-on-year, and vast progress has been made in identifying and understanding genetic variation in the human population. However, the nature of the studies above means that the findings often do not reveal the mechanistic basis or causative role of genetic variants, as the causal variant is usually not directly genotyped but rather in linkage disequilibrium with the genotyped SNPs. This necessitates methods to move closer to an understanding of the biology underlying a process or phenotype of interest caused by observed genetic differences (Figure 1.2b).

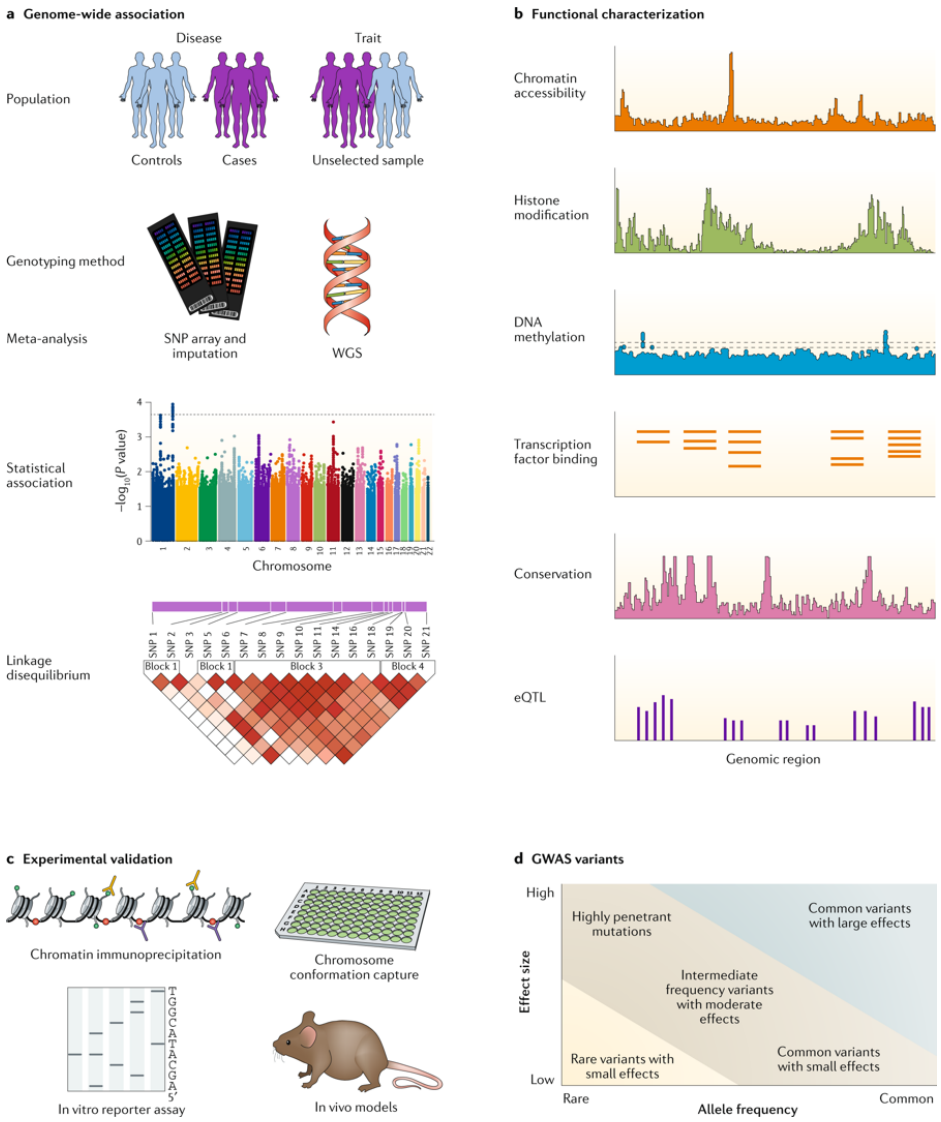

Fig. 1.2 Genome-wide association study design, from Tam *et al.* [14]: a) The aim of a genome-wide association study (GWAS) is to detect associations between allele or genotype frequency and trait status. The first step is to identify the disease or trait to be studied and select an appropriate study population. Genotyping can be performed using SNP arrays combined with imputation or whole-genome sequencing. Association tests are used to identify regions of the genome associated with the phenotype of interest at genome-wide significance, and meta-analysis is a common step to increase the statistical power to detect associations. b) Functional characterization of genetic variants is often required to move from statistical association to causal variants and genes, especially in the non-coding genome. Computational methods are used to predict the regulatory effect of non-coding variants on the basis of functional annotations. c) Target genes can be identified or confirmed using chromatin immunoprecipitation and chromosome conformation capture methods, and experimentally validated using cell-based systems and model organisms. d) Genetic variants exist along a spectrum of allele frequencies and effect sizes. Most risk variants identified by GWAS lie within the two diagonal lines. Rare variants with small effect sizes are difficult to identify using GWAS, and common variants with large effects are unusual for common complex diseases.

A key example is the expression quantitative trait loci (eQTL) approach, in which SNPs driving differences in expression levels of particular genes are identified. These eQTLs can be described as acting in *cis*, typically considered with a 1Mb window, or in *trans* from a more distant genomic location, typically 5Mb or further, or on a different chromosome entirely (Figure 1.3). By studying the transcription of genes, captured in RNA sequencing experiments, a more direct output of genetic variation can be captured. This intermediate phenotype can explain cellular events at a level closer to mechanism, uncovering novel biological insight into the disease or process of interest.

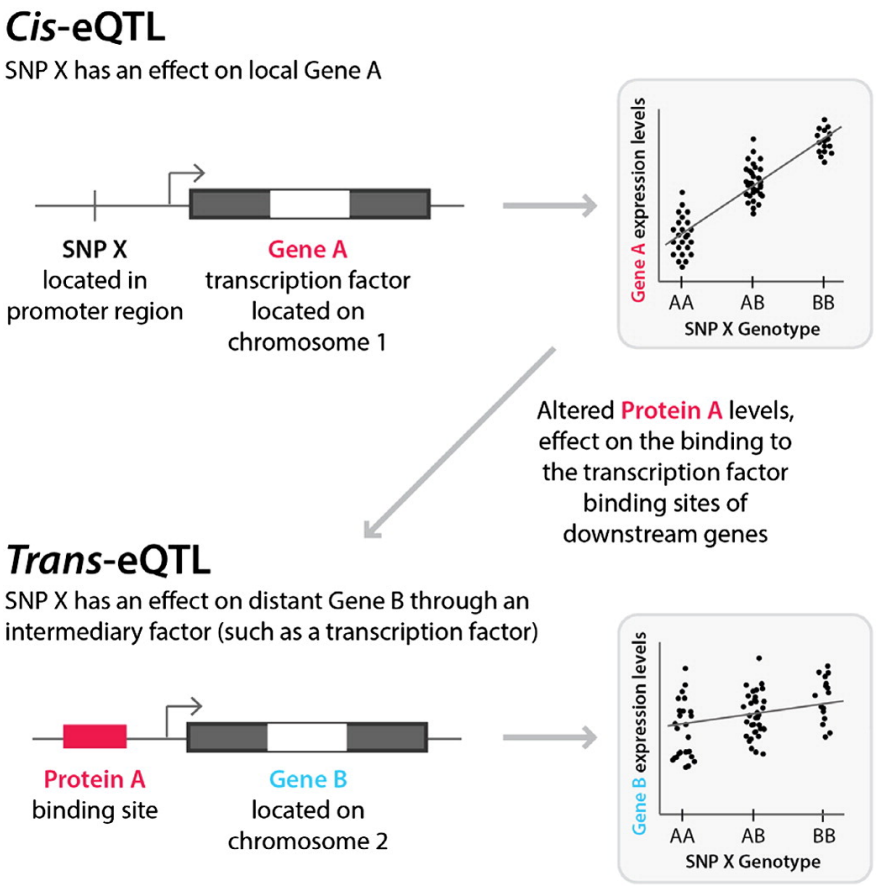

Fig. 1.3 The expression quantitative trait loci (eQTL) approach, from Westra & Franke [15]: eQTLs can be either local effects (cis-eQTLs), or distant, indirect effects (trans-eQTLs).

1.2 Single cell RNA sequencing

1.2.1 Evolution of single cell RNA sequencing technologies

While transcriptomic studies have, for many years, provided insight into mRNA expression and regulation, technological advances have allowed the quantification of transcripts at an unprecedented resolution. By sequencing the mRNA component of individual single cells, it has now become possible to study gene expression at an entirely new level, opening the door to novel biological questions which could not be addressed using population-level RNA sequencing. For example, the variability in splicing [16–20] and allelic expression [18, 21–23], between cells has been shown, along with analysis of the stochastic gene expression and transcriptional kinetics [24, 25]. Furthermore, single-cell RNA-sequencing (scRNA-seq) data have allowed fine-grained analysis of developmental trajectories [26–28] and identification of rare cell types [29, 30].

In order to obtain scRNA-seq data, cells must first be isolated individually in an accurate and rapid manner. Initially, microscopic manipulation provided a reliable method to isolate single cells through physical separation using a capillary pipette, and may still play an important role in systems where few cells are available. However, the high labour and low-throughput nature of this technique has resulted in it being surpassed by higher throughput methods. Fluorescence-activated cell sorting (FACS) provides an efficient way to isolate a large number of cells in a rapid manner, and also allows the selection of cells based on fluorescent labelling. Size or marker selection is commonly used, and through 'index sorting', the data for each cell can be recorded and used in downstream analysis. Despite the prevalence of this method, the high number of starting cells required, along with the potential damage caused by the staining and physical stress of the process, means it may be a problematic approach. More recently, microfluidic techniques have emerged as a key method for capturing

single cells, allowing isolation in small volumes within a closed system, often followed directly by amplification and downstream reactions. The small volume in which these reactions occur increases the capture efficiency and lowers the reagent cost. Finally, techniques involving the isolation of single cells in microdroplets, such as DropSeq [31] and InDrop [32], have rapidly expanded the high-throughput nature of scRNA-seq–allowing processing of tens of thousands of cells in a short space of time. The small volume of reactions, once again, decreases the cost per cell. Over time, these methods will continue to increase in speed, efficiency and reliability, further improving throughput of single-cell isolation.

Many protocols exist for the subsequent reverse transcription (RT), amplification, and library preparation prior to sequencing. Poly(T) priming is used to select polyadenylated mRNA for reverse transcription, however, only an estimated 10–20 percent of transcripts are sampled, particularly affecting lowly expressed genes [33]. Methods then differ in their approach to second-strand synthesis, either using poly(A) tailing, leading to a 3' bias, or template-switching to produce full-transcript coverage. Amplification can be achieved through two methods: linear *in vitro* transcription (IVT) or exponential PCR, each with its own advantages and drawbacks. Ziegenhain *et al.*[34] and Svensson *et al.*[35] provide a comprehensive experimental and computational comparison of most of the protocols commonly used. Following cDNA amplification, library preparation is most commonly carried out using the commercially available Nextera kit and sequencing on the Illumina platform, although other methods are available.

As a relatively new field, it is key to understand the structure and complexities of scRNA-seq data, ensuring that appropriate analytical and statistical methods are applied [36]. Particularly challenging is the high level of noise [37, 38], which derives primarily from the nature of single-cell experiments (called 'technical variation' and is mainly due to factors such as mRNA capture efficiency and cDNA amplification bias),

along with the biological heterogeneity of cells ('biological variation'). Furthermore, unlike with conventional RNA-sequencing where experimental biases are well studied [39, 40], there are biases which are still not fully understood in single-cell experiments, such as 'dropouts' due to the low amounts of starting material, leading to false negative expression.

Single-cell RNA-sequencing is a lossy technique, and it is not completely understood what causes the different failure modes for samples. Practically, this means the first step after acquiring reads from a scRNA-seq experiment is to perform quality control. Reads are processed in a similar manner to bulk RNA-seq, allowing expression quantification. There are several methods to do this, broadly split into those that use a genome reference for alignment, such as STAR [41], TopHat/TopHat2 [42, 43] and HISAT/HISAT2 [44, 45], and those that perform 'pseudoalignment', a quicker alternative, such as Kallisto [46] and Salmon [47].

It is important to check the quality of both the raw data (which can be performed using tools developed for bulk RNA-seq, such as FastQC [48] or Kraken [49]), along with the aligned output. Imperative in scRNA-seq is the cell-by-cell quality control [50], ensuring that cells of poor quality are removed from subsequent analysis. Many metrics can be used to measure cell quality, such as the number of reads or genes detected, the proportion of reads mapping to mitochondrial genes (which may signify leaking of cytoplasmic RNA or cells undergoing apoptosis), or the proportion of reads mapping to externally spiked-in RNA molecules if used in the experiment [51].

Depending on the analysis task, appropriate normalization of the data is needed. Several normalization methods have been developed, many of which adjust for differences in sequencing depth and/or make use of spike-in molecules and/or unique molecular identifiers (UMIs) when available (reviewed in detail in [52]). Once cleaned data are obtained, there are many routes of analysis depending on the biological

question under investigation (Figure 1.4). In the next section, I will consider these analysis from two viewpoints: cell-level approaches, such as the grouping of cells and trajectory ordering, along with gene-level investigations, such as gene variability and noise, co-expression, and identification of differentially expressed genes.

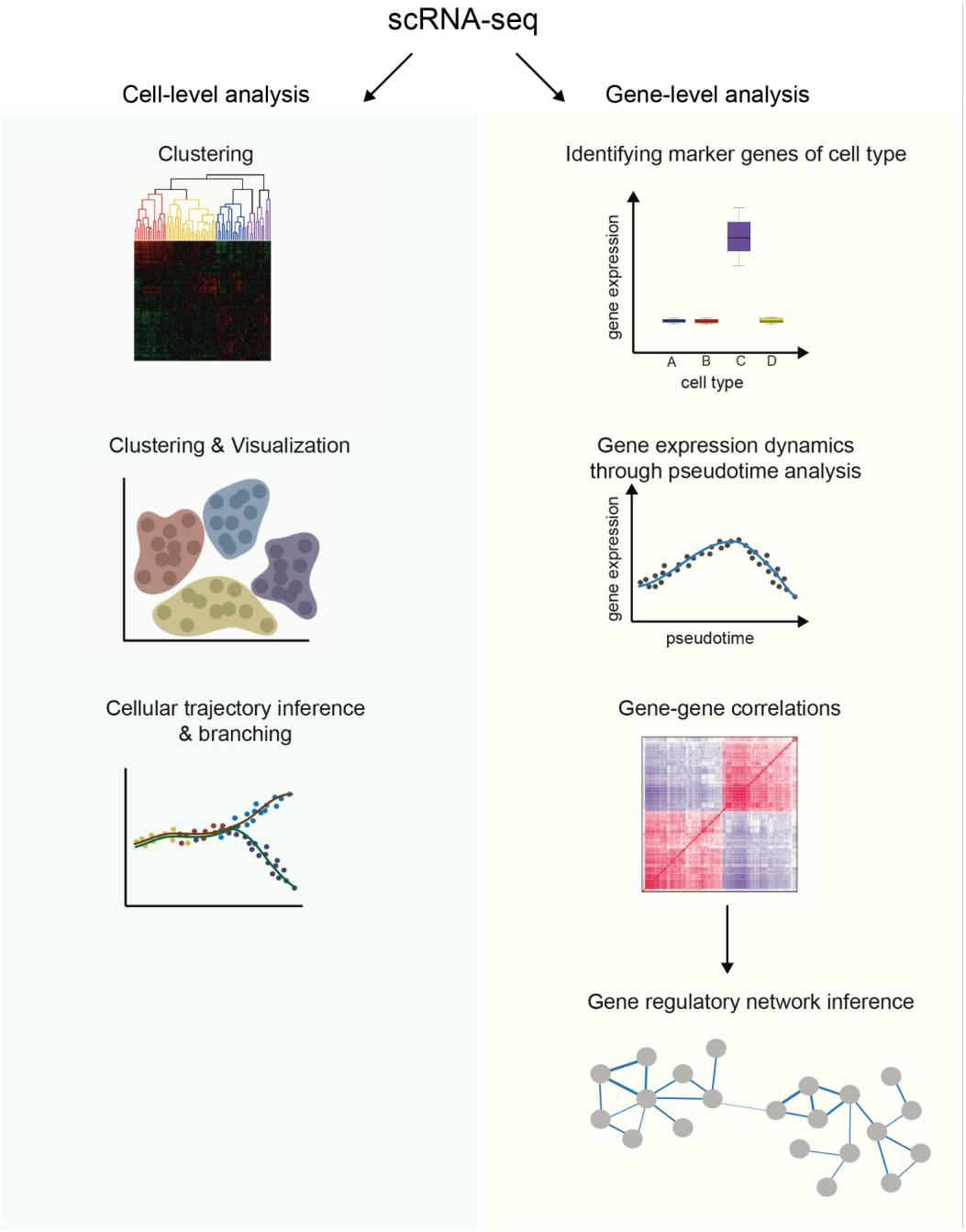

Fig. 1.4 Overview of analysis methods for the interpretation of scRNA-seq data.

1.2.2 Analysis of scRNA-sequencing data

Cell level analyses

Visualising and clustering cells

The cataloging and classification of cells is a long-standing biological challenge. Traditionally, cell types were determined morphologically or based on molecular cell surface markers. However, with the availability of genome-wide expression data, the possibility of transcriptome-based analysis of cell similarity provides an alternative indicator of cell type.

The first step in understanding the distribution of cells is often to apply dimensionality reduction techniques: this represents the thousands of dimensions (genes) found in scRNA-sequencing data with a much smaller number, attempting to maintain a representation of some variation of interest. Furthermore, by considering only a two or three dimensional space, visualisation provides a mean to qualitatively explore the data. There are hundreds of dimensionality reduction methods available which the researcher can elect to apply either to all observed genes or a selected subset of genes of interest. The most widespread is Principal Component Analysis (PCA) [53], where weighted sums of dimensions represent the data. The dimensions for each sample are known as principal components. These dimensions explain decreasing amounts of variation in the original data, with the first principal component capturing as much of the variance as possible. Another commonly applied method is t-SNE (t-Distributed Stochastic Neighbour Embedding) [54], a non-linear visualization technique which considers local distances between data points (cells) by combining dimensionality reduction with random walks on the nearest-neighbour network with the goal of separating far-apart clusters, while also ensuring all data points can be seen by eye to allow for comparisons of cluster size. This is a variation of Multidimensional Scaling (MDS), where PCA is applied on pairwise Euclidean distances to preserve pairwise distances

in a low-dimensional space. A recent, and increasingly adopted method, is uniform manifold approximation and projection (UMAP) [55], which has been shown to preserve more of the global structure within datasets, with an improvement in run time and reproducibility [56].

While powerful, and popular, these techniques can be heavily affected by the problematic abundance of zeroes in single cell data; an issue which several methods account for. ZIFA (zero-inflated factor analysis) [57] extends the linear factor analysis framework, (based on correlations in the data rather than covariances), accounting for dropout characteristics in the data. The R-package Destiny provides an alternative, non-linear method using diffusion maps [58]: distance between cells reflects the transition probability based on several paths of random walks between the cells. This assumes a smooth nature of the data, and also includes imputation of drop-outs.

Unsupervised clustering techniques provide a mechanism to group cells by similarity. While this unbiased approach has benefits, the small number of samples and absence of a way to validate if groupings are "real" poses a problem, along with prior information on the number or type of groups. The features of single cell data discussed above, such as dropouts, biases and noise, also add to the difficulty of accurate clustering. Despite these problems, several tools have been developed for use with scRNA-seq, along with traditional methods such as hierarchical clustering [59]. SNN-Cliq [60] achieves clustering by considering similarity calculated using a graph-based approach in which a shared nearest neighbour (SNN) network is constructed using rankings of similarities based on expression levels; dense clusters of nodes (cells) are then found. RaceID [29], while also using similarity in expression between cells (based on Pearson correlation), utilises a different approach: k-means clustering. In k-means clustering each sample is associated with a one of k prototypes, so that the total squared distance (inverse of similarity) from samples to prototypes is minimal. After the initial step, RaceID uses

an outlier detection algorithm and identifies cells which do not fit the model accounting for technical and biological noise. This has been used in the detection of rare cell populations. Another k-means-based tool, Single Cell Consensus Clustering (SC3) [61], uses consensus clustering [62], an ensemble strategy, to average over parameter choices in an attempt to make cluster assignments more robust. Another method, SIMLR [63], uses multiple-kernel learning to infer similarity in a gene expression matrix with a given number of cell populations. As multiple kernels are used, it is possible to learn a distance measurement between cells that is specific to the statistical properties of the scRNA-seq set under investigation. Two widely adopted strategies using a community detection approach are Louvain [64] and Leiden [65] clustering, In the first method, clusters are identified by moving nodes individually between groups until the quality of clusters can no longer be improved. The network is then aggregated, with each cluster becoming a node, and the steps of node movement and aggregation repeated. While this leads to an efficient approach, clusters may be badly connected - a problem which the Leiden method tackles by improving upon the aggregation step, allowing clusters to be split.

Cellular trajectory inference and branching analysis

Trajectory analysis is a simpler version of dimensionality reduction, where the assumption is that a 1-dimensional "time" can describe the high-dimensional expression values. The theory is that during a biological process, changes will happen gradually, so biological observations can be ordered compared to each other in terms of pairwise similarity. While clustering techniques have been used to define discrete population and states for a long time, trajectory inference is younger in the field of scRNA-seq. However, many methods have been developed in recent years, and Saelens *et al.* recently conducted a comprehensive benchmarking of 45 of these methods [66]. Here, just a subset of methods are described.

One of the initial methods for so called pseudotime analysis of single cells was Monocle [67], which used a minimum spanning tree (MST) strategy to order cells by the distance to a start cell, based on a technique for putting microarray samples on a trajectory [68]. In the updated versions of Monocle, the MST strategy has been replaced by a more sophisticated tree embedding strategy [69, 70].

Diffusion pseudotime (dpt) [27] offers an alternative, in which geodesic pairwise distances between samples on the data manifold are approximated using a diffusion map representation. Trajectory is then defined as the distance from a start cell along these distances. A different strategy for trajectory inference is to consider a generative model for the data, treating "time-points" as hidden (or latent). This leads to the probabilistic interpretation of PCA, which in turn leads to factor analysis and ZIFA. Here the expression of each gene can be described as a linear function of an unknown "time".

Non-linearity in the data, as described in [67] precludes PCA from being an effective technique for this task. The Gaussian Process Latent Variable Model (GPLVM) allows gene expression to follow any smooth (non-linear) function over time [71]. While more computationally demanding than linear versions, this allows cells to be put in the most likely ordering [71, 72]. This means that the most number of genes exhibit smooth expression curves with as little noise as possible. Being a probabilistic model, the benefit is that uninteresting structure in the data can be accounted for directly, such as batch effects or technical factors. It is also possible to incorporate more information about experimental design through priors [28].

The Ouija method [73] takes a different approach to pseudotime in a couple of ways. Firstly, it defines a generative model for gene expression in scRNA-seq data based on ZIFA, to deal with the most common types of measurement noise. Secondly, it is based on the assumption that a small number of switch-like markers for a biological process

of interest are known. The cells are then ordered according to the most likely ordering to confer with the switching genes.

A unique problem in single cell developmental data is that a set of progenitor cells can develop into multiple distinct cell types. This means the cells will not follow a single trajectory in the high-dimensional space. A couple of heuristics have been published: in Wishbone [74], cells are clustered by the pairwise detour distance relative to a reference cell, using geodesic distance. This method is reported to be correctly recovering the known stages and bifurcation point of T-cell development in mouse. Another method, that has been introduced by Haghverdi *et al.* [27], measures transition between cells using a random-walk-based distance.

More principled model based approaches have been presented with SCUBA, which considers transition of cell clusters over time [75], as well as with GPfates / OMGP [28], where multiple smooth trajectories are explicitly modeled. After inference, each cell gets assigned a posterior probability of having been sampled from a particular trajectory. This method has been shown to be efficient in reconstructing the developmental trajectories of Th1 and Tfh cell populations during Plasmodium infection in mice.

An interesting recently developed method, partition-based graph abstraction (PAGA), generates a graph-like map, estimating connectivity of partitions in the data [76]. This approach provides a way to bridge the clustering type of analysis, as discussed above, with the continuous nature of many biological processes, as modelled with conventional pseudotime approaches.

Gene level analyses

Unwanted factor removal

Uninteresting, largely technical variation can be observed in both bulk RNA-seq and scRNA-seq experiments. This variation is usually correlated with some common experimental factor, such as room temperature or stock of reagents. This form of variation are known as batch effects. It is possible to handle batch effects by having a careful balanced experimental design, such as uniformly distributing replicate conditions across batches. For statistical analysis and inference, if the samples are spread over multiple batches, this information can directly be accounted for [77]. Additionally, several statistical methods have been developed to adjust for batch effects [78, 79]. One example is ComBat, which removes known batch effects using a linear model of expression from batches where variance is based on an empirical Bayesian framework [78].

Technical variation in scRNA-seq experiments could be due to mRNA capture efficiency, cDNA amplification bias and the rate cDNAs in a library are sequenced. To estimate technical variation, several methods use spike-in molecules, which are added with each cell in the same quantity. Risso *et al.* developed a sleuth of strategies called RUVSeq that either performs factor analysis on a set of control genes such as ERCC spike-ins or samples within replicate libraries to identify technical factors which can be adjusted for [80]. Similar strategies have also been made by others [81–83].

Substantial amount of variation also results from differences in cell size or cell cycle stage of each cell. To adjust for cell cycle effects, Buettner *et al.* have developed single-cell latent variable model (scLVM), which is a two-step approach that reconstructs cell cycle state before using this information to obtain adjusted gene expression levels by linear regression [84]. They have also shown that removing cell cycle effects in T cells reveals sub-populations associated with T-cell differentiation [84]. This highlights the

importance of dissecting biological variation into interesting and uninteresting parts in correctly characterizing sub-populations.

In recent years, many further methods have been developed for the integration of discrete experimental batches. One example is canonical correlation analysis (implemented in Seurat [85], which identifies a shared gene correlation structure across datasets, using this structure to align the datasets. Haghverdi *et al.* developed a mutual nearest neighbours (MNN) approach [86] to correct expression between batches. This method uses 'landmark' cells, which are representative of cell types or clusters across all datasets to be integrated. Park *et al.* provide an alternative approach, using a batch balanced k nearest neighbour graph (BBKNN) [87] to combine batches. While these examples highlight just some of many methods available, there will undoubtedly be further work in this area, particularly given the increasing scale of scRNA-seq data generated and desire to integrate across experiments.

Identification of highly variable genes

Several methods have been developed to identify genes that show high biological variability. Brennecke *et al.* have first estimated technical noise using spike-in molecules and modeled mean-variance relationship to identify highly variable genes [37]. Kim *et al.* have presented a statistical framework to decompose the total variance into the technical and biological variance based on a generative model, which would help in identifying variable genes [22]. Another method, BASiCS, uses a Bayesian model which jointly models spike-ins and endogenous genes and provides posterior distributions for the extent of biological variability [88].

Identification of differentially expressed genes and marker genes

Identification of differentially expressed genes and marker genes of subpopulations is a simple yet important analysis in scRNA-seq studies. Although originally developed for bulk RNA-seq experiments, methods such as DESeq2 [89] and EdgeR [90] are also

widely used in scRNA-seq experiments. DESeq2 identifies differentially expressed genes by fitting a generalised linear model (GLM) for each gene, uses shrinkage estimation to stabilize variance and fold changes, and applies a Wald or likelihood ratio (LR) test for significance testing [89]. EdgeR fits a GLM with negative binomial (NB) noise for each gene, estimates dispersions by conditional maximum likelihood, and identifies differential expression using an exact test adapted for overdispersed data [90]. Monocle also fits a GLM, but dispersion is estimated directly from the data for each gene, since most single cell studies have enough samples to allow this [67]. For relative abundance data, dropouts are handled by using a tobit noise model, while using a NB noise model with imputed dropouts for count data.

One method developed for scRNA-seq experiments, called MAST, uses two-part generalized linear model that is adjusted for cellular detection rate (dropouts) [91]. Another method, M3Drop, applies Michaelis-Menten modelling of dropouts in scRNA-seq, that is used to identify genes differentially dropped out [92]. SCDE is a Bayesian method to compare two groups of single cells, taking into account variability in scRNAseq data due to drop-out and amplification biases and uses a two-component mixture for testing for differences in expression between conditions [93]. Another method, SINCERA identifies differentially expressed genes based on simple statistical tests such as Wilcoxon rank sum and t-tests [94]. In comparison to these methods, scDD identifies genes where the overall distribution of values have changed between conditions. This answers a different question which might be of interest in scRNA-seq experiments [95]. Using a Bayesian modeling framework, scDD classifies each gene into one of the four types of changes across two biological conditions: shifts in unimodal distribution, differences in the number of modes, differences in the proportion of cells within modes, or both differences in the number of modes and shifts in unimodal distribution [95].

Gene-centric expression dynamics through pseudotime analysis

Using an inferred trajectory as described above, samples can be analysed using a continuous time covariate instead of a few discrete time points. This enables the use of more sophisticated time-series based analysis techniques for modeling gene expression dynamics, and allows us to ask more complex questions from the data.

The popular scRNA-seq package Monocle provides a wrapper for the vector generalised additive model (VGAM) package to investigate how expression changes over the trajectory. Splines are used to model expression dependence on pseudotime to allow non-linear trends. The VGAM package allows for more than just expression levels to be modelled by the splines: with appropriate link functions, allelic expression balance or isoform usage can be modelled [18]. Splines require several parameters to be chosen however, and the choices greatly affect the results. A non-parametric non-linear alternative to spline regression is Gaussian Process regression, which can be used in a likelihood ratio based fashion to identify genes which are dependent on pseudotime [71, 96].

Often, we want to ask particular questions from the data, in which case parametric models are useful. In the SwitchDE method, genes which sequentially switch on or off can be identified, along with a parameter letting you learn when the switch happens [97]. Similarly, an assumption can be that genes are described as a transient pulse over the pseudotime. The package ImpulseDE identifies such genes, while providing parameters for when in pseudotime the pulse occurs [98].

Correlation analysis and network inference

One important application of scRNA-seq studies is the identification of co-regulated modules of genes and gene-regulatory networks constructed using gene-to-gene expression correlations. Here, genes with highly correlated expression levels across cells are assumed to be co-regulated. Using single-cell transcriptomic data of Th2 cells, Mahata

et al. demonstrated how gene-gene correlations can be used to reveal novel mechanistic insights; they have applied correlation analysis between steroidogenic enzyme Cyp11a1 and cell surface genes and identified Ly6c1/2 as a marker of the steroid-producing cell population in mouse [99].

One method to elucidate regulatory interactions in bulk RNA-seq studies is called the weighted gene co-expression network analysis (WGCNA) [100]. In such a network, nodes represent genes and edges represent co-expression as defined by correlation and relative interconnectedness. The method has also been applied in a scRNA-seq study where the authors have identified a number of functional modules of co-expressed genes that can describe each embryonic developmental stage in mouse [101].

Although these methods are useful, the inferred networks are undirected; that is, they do not provide direct regulatory relationships among genes. One method, SCENIC, aims to address this by constructing gene regulatory networks from scRNA-seq data [102]. SCENIC defines co-regulated modules, or 'regulons', using GENIE3 [103] to identify the targets of transcription factors, and cis-regulatory motif analysis.

1.3 The human innate immune system

The innate immune system is the body's first line of defence against damage, rapidly sensing and responding to infectious or harmful agents. Due to the diversity of potential threats, a range of mechanisms are utilised to act against infections. These are prompted by detection of pathogen-associated molecular patterns (PAMPs) - conserved structures which are predominantly expressed by large groups of pathogens, rather than the host. One major group of PAMPs is nucleic acids (Gurtler Bowie, 2013). Although there may be complexity in detecting various RNA and DNA structures due to the similarity with host nucleic acids, their essential role for pathogen survival means they are a useful signal, particularly for viruses in which there may be a limited amount of alternative distinguishing molecular features.

1.3.1 The type I interferon response

To detect these pathogenic signals, there are several classes of pattern recognition receptors (PRRs) in various cytoplasmic or membrane-bound locations. These include Toll-like receptors (TLRs), RIG-I-like receptors (RLRs) and NOD-like receptors (NLRs), among others. These receptors may function in signalling - activating downstream pathways to instigate a response - or have a direct effector function, blocking pathogenic replication and propagation [104]. In the case of viral infections, distinct sensors play a role in the recognition of viral RNA and DNA. In the case of RNA, RIG-I and MDA5 sense cytosolic non-self RNA, with specificity towards differing lengths of dsRNA [105]. In contrast, the presence of viral DNA is sensed through cGAS, leading to the production of cGAMP and consequent activation of STING [106].

Despite specific recognition pathways, activation of viral sensors converges in downstream signalling, leading to activation of NF-κB, TBK1 and IRF3 to induce

production of type I interferons (IFNs). In this thesis, the response to RNA viruses is studied, using synthetic dsRNA to mimic the presence of viral nucleic acids in host cells. The induction of the type I interferon response through dsRNA sensing is shown in Figure 1.5.

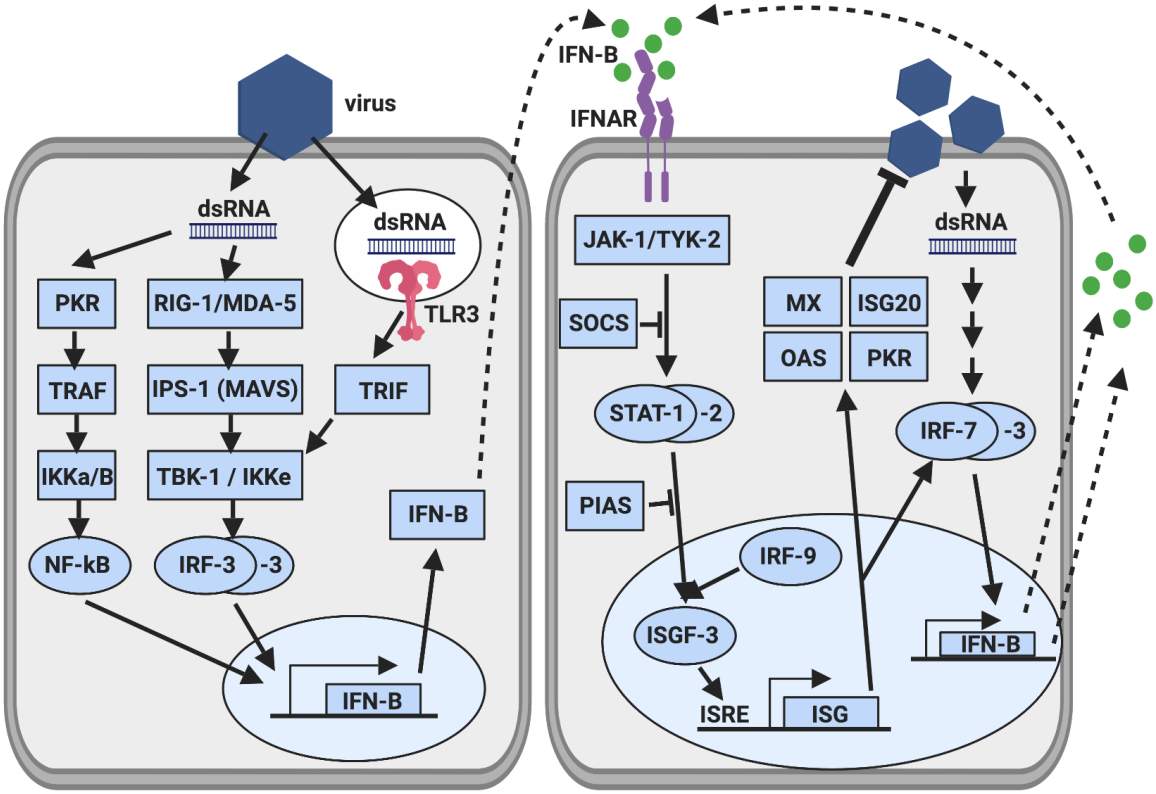

Fig. 1.5 Induction of the Type I Interferon response.

Interferons are a subset of the class of immune signalling cytokines, and can be subdivided into type I (IFN-α, IFN-β, IFN-ϵ, IFN-κ, IFN-ω), type II (IFN-γ) and type III (IFN-λ), based upon receptor specificity [107]. Of the type I IFNs, which bind the heterodimeric IFNAR1-IFNAR2 receptor, IFN-α and IFN-β are the most studied. There are 14 IFN-α genes and only one IFN-β gene in humans. When bound to type I IFNs, the IFNAR1-IFNAR2 heterodimer activates JAK1 and TYK2 [108], leading to phosphorylation of STAT1-STAT2 heterodimers [109]. Consequent migration into the

nucleus, association with IFN regulation factor 9 (IRF9) and binding to IFN-stimulated response elements instigates transcription of IFN-inducible genes.

Type I interferons play an important role in the response to viral infections (reviewed in [110]), and are able to be produced at low levels by most cell types. Certain cells have been shown to function in producing high levels of these proteins, invoking a systemic response. Plasmacytoid dendritic cells (pDCs), for example, were identified as 'natural interferon producing cells' [111]. However, fibroblasts mainly produce IFN-β, considered the central cytokine responsible for stimulating cells locally. This leads to altered gene expression, chemokine production, antigen presentation and the induction of an adaptive immune response (Figure. 1.6).

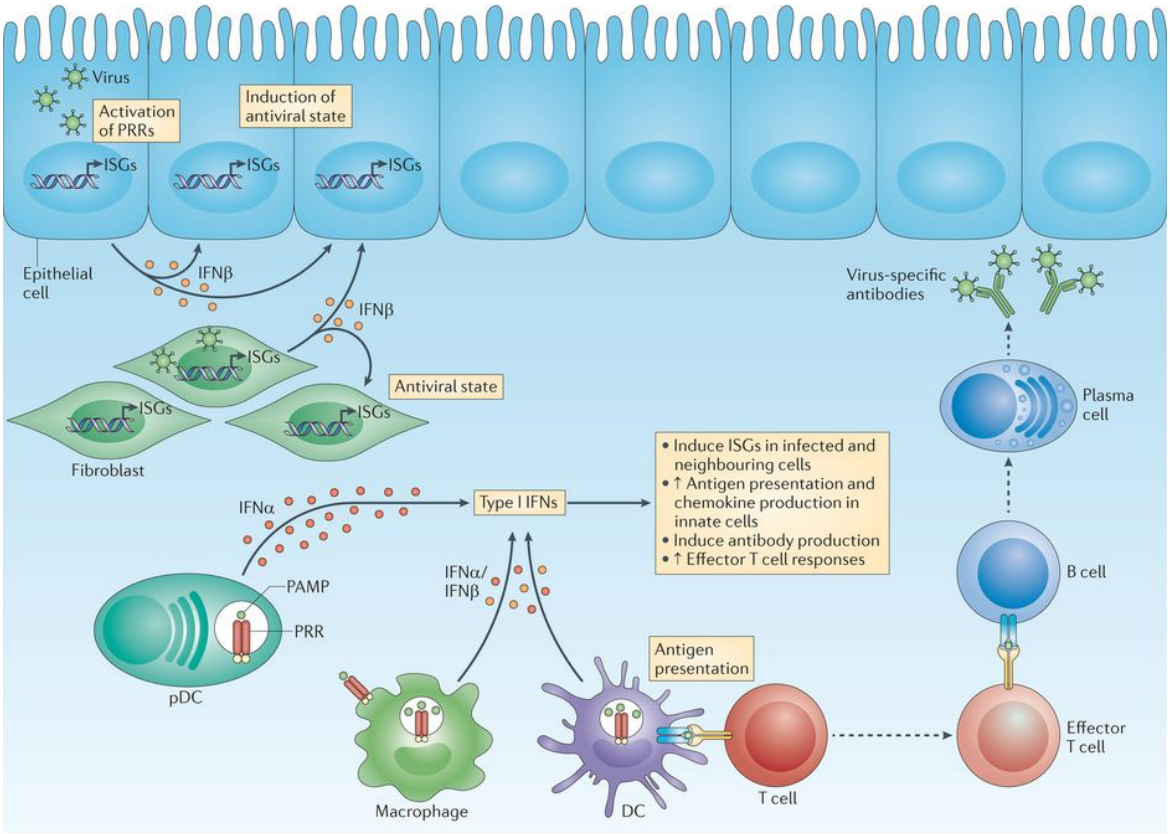

Fig. 1.6 The role of interferons and inter-cellular communication in the immune response. (Ivashkiv Donlin, 2014 [112])

1.3.2 Cell-to-cell heterogeneity in innate immunity

As with many biological processes, the innate immune response is considerably heterogenous between cells in an individual, despite cells being genetically identical. This derives both from the stochastic nature of biochemical reactions within cells (intrinsic) along with communication between cells and other environmental factors (extrinsic). Fluctuation in gene expression is one of the largest causes of variation in clonal populations, reviewed in [113]. This is due to the low molecular abundance of some key elements (such as transcription factors) along with the number of chemical reactions required to turn genetic sequence to functional product. Although modelled for many systems, a severely reduced view is often taken, with only transcription and translation included.

Within immunology, stochastic expression of many interleukin genes has been observed, such as IL-2 [114] and IL-10 [115]. In fibroblasts, large variation in the induction of IFN-β in response to viral infections has been shown [116]. Furthermore, heterogeneity in response of coarse-grained cell populations has been studied, such as macrophages [117, 118] and monocytes [**?**]. The considerable advance in single-cell technologies in recent years, however, will allow further illumination of inter-cellular variability in innate immunity. For example, scRNA-seq was recently used to reveal novel dendritic cell and monocyte sub-populations [119]. However, scRNA-seq holds exciting possibility not only for cell classification, but also in understanding the innate immune response. One example is the discovery of bimodal transcript splicing in bone-marrow derived dendritic cells in response to lipopolysaccharide (LPS) treatment [16].

1.3.3 Genetic variability in the innate immune response

The hereditary nature of some susceptibility to infectious diseases has long been known, alongside the variation between individuals in the response to particular pathogens. Early investigations involved twin studies, which showed a higher concordance in identical to non-identical twins for some infections, particular those which are chronic and have low infectivity. Examples of these findings for viral infections include poliomyelitis and hepatitis B [120]. More recently, the development of high throughput technologies, as described above, has enabled the identification of single nucleotide polymorphisms (SNPs) associated with particular infections. Just two examples of many available are Hepatitis C clearance, for which a SNP in IL28B has been identified [121], and reduced influenza virus clearance (a SNP in IFITM3; [122]).

Unlike adaptive immunity, in which receptor sequences undergo rearrangement in somatic cells, the innate immune system's pattern recognition receptors are germline encoded, along with signalling components and effector mechanisms. Therefore all aspects of innate immunity, from recognition to action, are likely to be subject to genetic variation. Genetic analysis of patients with susceptibility to particular infections has pinpointed elements of the innate immune, and more specifically type I interferon, response as playing a key role. For example, Zhang *et al.* [123] described two children with herpes simplex encephalitis, both with a heterozygous mutation in TLR3. In dermal fibroblasts from these individuals, treatment with a synthetic dsRNA (polyinosinic:polycytidylic acid; also known as poly(I:C)) did not induce expression of IFN-β, IFN-γ or IL-6. More recently, Ciancanelli *et al.* [124] characterised a patient with compound heterozygous null mutations in interferon response factor 7, who suffered a life threatening primary influenza infection. In this case, dermal fibroblasts and iPSC-derived epithelial cells from the patient produced reduced amounts of type I IFN and showed increased viral replication. There have been further studies showing

deficiency in innate immune signalling pathways in the fibroblasts of affected individuals, such as those with an IRAK1 [125] or DOCK2 [126] mutation.

Moving beyond individuals with a deficient innate response or specific susceptibility, expression quantitative trait loci (eQTL) approaches been used to characterise genetic variability within healthy populations. Some studies have identified SNPs in particular mechanisms, such as the TLR4 pathway [127]. In recent years, however, investigations have expanded from studying one pathway or pathogen to eQTL mapping in broader innate immune stimulation. Two studies in which this has been conducted are Fairfax *et al.*, 2014 [128], where primary CD14+ monocytes were treated with IFN-γ or LPS, and Lee *et al.*, 2014 [129], in which dendritic cells were stimulated with influenza virus, LPS, or IFN-β. These studies identified treatment-specific eQTLs, highlighting the importance of considering genetic variation within the biological context of interest. However, in these studies changes in expression were measured only at a cell population level and at distinct time points. Further insight is needed into the genetic effect on variability of innate immune components, gained from single-cell expression studies, along with the dynamics and regulation of the response.

1.4 Using single-cell RNA sequencing data to study genetic variation in the innate immune response

While there have been significant advances in understanding the genetic basis of variation in the innate immune response in recent years, there is still a way to go in defining the intra- and inter-individual components of this variability, particularly in healthy individuals. In order to do this, a dataset spanning a large number of donors, profiled at single cell resolution, is required. To this end, this thesis outlines the establishment of an experimental system using relatively homogenous dermal fibroblast populations of 70 human individuals obtained from the Human Induced Pluripotent Stem Cell Initiative (HipSci). Assaying these cells using two stimulation conditions - a synthetic dsRNA, and Interferon-β - over time allows us to study key questions:

(1) How does the interferon response vary between human individuals and can this variation be attributed to common genetic variants?

(2) How do different cells from the same donor respond to a danger signal that should elicit interferon, and how do they respond to a direct interferon stimulus?

Alongside this, the heterogeneity in unstimulated human fibroblasts is characterised, to understand the variation and clonality seen in genetically identical populations of fibroblasts. The single-cell resolution provides unprecedented insight into not only the human genetics of the innate immune response, but also the role of cell-to-cell variation in this response.