From The Department of Cell and Molecular Biology
Karolinska Institutet, Stockholm, Sweden

SINGLE-CELL RNA SEQUENCING FOR SUBTYPE DISCOVERY IN PLASMODIUM FALCIPARUM AND MAMMALIAN CELLS

Mtakai Ngara

Stockholm 2019

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Eprint AB 2019

ISBN 978-91-7831-403-4
The drawing is a depiction of cytological stages of malaria parasite in tSNE subspace

Single-cell RNA sequencing for subtype discovery in Plasmodium falciparum and mammalian cells
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Mtakai Ngara

Principal Supervisor:

Professor Rickard Sandberg

Karolinska Institutet

Department of Cell and Molecular Biology

Co-supervisor(s):
Professor Björn Andersson
Karolinska Institutet
Department of Cell and Molecular Biology

Opponent:
Professor Bart Deplancke
The Ecole Spéciale de Lausanne (EPFL)
School of Life Sciences

Examination Board:
Dr. Goncalo Castelo-Branco
Karolinska Institutet
Department of Medical Biochemistry and Biophysics

Professor Staffan Svärd
Uppsala University
Department of Cell and Molecular Biology

Dr. Ola Larsson
Karolinska Institutet
Department of Oncology-Pathology

To my loving family and friends

Since new developments are the products of a creative mind, we must therefore stimulate and encourage that mind in every way possible.

George Washington Carver

ABSTRACT

Since the dawn of massively parallel sequencing technologies in mid-2000s their utility in profiling the expression of genes in a genome-wide fashion has matured and progressed from cell populations to individual cells. In particular, single-cell RNA sequencing (scRNA-seq) has impacted numerous domains in life sciences and hold immense promise in biology and medicine. Indeed, it has become realistic to chart the complete set of cell types and states in multicellular organisms, and projects have started to map out cell types in humans (i.e. the Human Cell Atlas project) and model organsims. In this thesis, I present the application of scRNA-seq to infectious disease and cancer as well as a computational assessment of the general possibilities and limitations of scRNA-seq for enumerating cell types and states de novo. In Paper I, we describe the ability of scRNA-seq to profile transcriptomes from individual malaria-causing *P. falciparum* parasites. We reveal heterogeneity even among synchronized cultures of parasites during their red blood cell life cycle. Moreover, we identify a subset of sexually differentiated *P. falciparum* with a distinct gene signature, likely important for parasite transmission that may be exploited for the design of transmission-blocking drugs and/or vaccines. In Paper II, I present a computational strategy to identify the magnitude of biological gene expression differences needed for accurate inference of cell identities using scRNA-seq. Interestingly, rather large differences are needed for proper cell state discrimination, irrespective of scRNA-seq protocol, implying that large number of cell states may escape detection. In Paper III, we used scRNA-seq and bulk RNA-seq to characterize the molecular programs during the later stages of lung metastasis. We demonstrate that a transition from epithelial to mesenchymal cell characteristics occurs in cancer cells during metastasis, and that the mesenchymal properties are maintained during metastasis growth extending over a week. In Paper IV we performed transcriptome analyses on stem and progenitor populations in myelodysplastic syndrome (MDS) patients. We provide evidence that the MDS stem cells and the progenitors have distinct transcriptome. Altogether, this thesis expands the applications of scRNA-seq towards parasite biology and cancer metastasis and we provide valuable insights into the abilities of current scRNA-seq technologies in mapping cell states in an unbiased fashion.

LIST OF SCIENTIFIC PAPERS

I. **Mtakai Ngara***, Mia Palmkvist*, Sven Sagasser*, Daisy Hjelmqvist, Åsa K. Björklund, Mats Wahlgren, Johan Ankarklev* and Rickard Sandberg*
Exploring parasite heterogeneity using single-cell RNA-seq reveals a gene signature among sexual stage Plasmodium falciparum parasites
Experimental Cell Research 2018 Oct 1; 371(1)

II. **Mtakai Ngara** and Rickard Sandberg
Defining limits to subtype discovery in single-cell RNA-sequencing experiments
Manuscript

III. Azadeh Nilchian*, **Mtakai Ngara***, Åsa Segerstolpe*, Vedrana Tabor*, Mikael Karlsson, Rickard Sandberg* and Jonas Fuxe*
Single-cell sequencing reveals autocrine TGF-b- induced EMT as a promoter of late-stage metastasis
Manuscript

IV. Petter S. Woll, Una Kjällquist, Onima Chowdhury, Helen Doolittle, David C. Wedge, Supat Thongjuea, Rikard Erlandsson, **Mtakai Ngara**, Kristina Anderson, Qiaolin Deng, Adam J. Mead, Laura Stenson, Alice Giustacchini, Sara Duarte, Eleni Giannoulatou, Stephen Taylor, Mohsen Karimi, Christian Scharenberg, Teresa Mortera-Blanco, Iain C. Macaulay, Sally-Ann Clark, Ingunn Dybedal, Dag Josefsen, Pierre Fenaux, Peter Hokland, Mette S. Holm, Mario Cazzola, Luca Malcovati, Sudhir Tauro, David Bowen, Jacqueline Boultwood, Andrea Pellagatti, John E. Pimanda, Ashwin Unnikrishnan, Paresh Vyas, Gudrun Go˙hring, Brigitte Schlegelberger, Magnus Tobiasson, Gunnar Kvalheim, Stefan N. Constantinescu, Claus Nerlov, Lars Nilsson, Peter J. Campbell, Rickard Sandberg, Elli Papaemmanuil, Eva Hellström-Lindberg, Sten Linnarsson and Sten Eirik W. Jacobsen
Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells In Vivo
Cancer Cell 2014 June 16; 25(6)

CONTENTS

LIST OF ABBREVIATIONS

ApiAP2	Apicomplexan Apatele-like DNA binding protein
ATP	Adenosine Triphosphate
cDNA	Complimentary DNA
CSC	Cancer Stem Cell
CTC	Circulating Tumor Cell
DNA	Deoxyribonucleotide
ERCC	External RNA Controls Consortium
FACS	Fluorescence-Activated Cell Sorting
FISH	Fluorescent In Situ Hybridization
GFP	Green Fluorescent Protein
GTF	General Transcription Factor
IDC	Intra-erythrocytic development cycle
iRBCs	Plasmodium-infected red blood cells
IVT	In vitro transcription
lincrRNA	Long non-coding RNA
MDS	Myelodysplastic syndromes
mRNA	Messenger RNA
NGS	Next generation sequencing
PCR	Polymerase chain reaction
Pol II	RNA polymerase II
RNA	Ribonucleotide
RT-qPCR	Real Time Quantitative Polymerase Reaction
sci-RNAseq	Single cell combinatorial indexing RNA sequencing
SCT	Single cell transcriptome
Smart	Switch Mechanism at the 5' End of RNA Templates
SP	Subpopulation
STF	Specific transcription factor
STRT-seq	Single-cell transcriptional tagging sequencing

TFIIA	RNA polymerase II associated transcription factor A
TFIIB	RNA polymerase II associated transcription factor B
TFIIC	RNA polymerase II associated transcription factor C
TFIID	RNA polymerase II associated transcription factor D
tSNE	t-distributed Stochastic Neighbour Embedding
UMAP	Uniform Manifold approximation and Projection
UMI	Unique Molecular Identifier

1 BACKGROUND

1.1 EUKARYOTES: FROM UNICELLULAR TO METAZOAN ORGANISMS

The cell is the fundamental unit of life and understanding it at a molecular level is critical in both microbial and complex organisms. Mammalian systems for instance, are made up of functionally specialized tissues and organs that encompass a heterogeneous mix of distinct and dynamic types, subtypes and states. In spite of having identical genomes most mammalian cells adopt specialized form and function that augments the complexity of these organisms (Alberts et al. 2014; Chen & Dent 2014; Levine & Tjian 2003). Also, communities of single celled eukaryotes (e.g. malaria parasites) undergo remarkable transitions in form and function in response to stimuli and developmental progression in order to thrive in different environments (Cowman et al. 2016; Cowman et al. 2017). Hence, intricate regulation of gene expression is a crucial component to eukaryotic life.

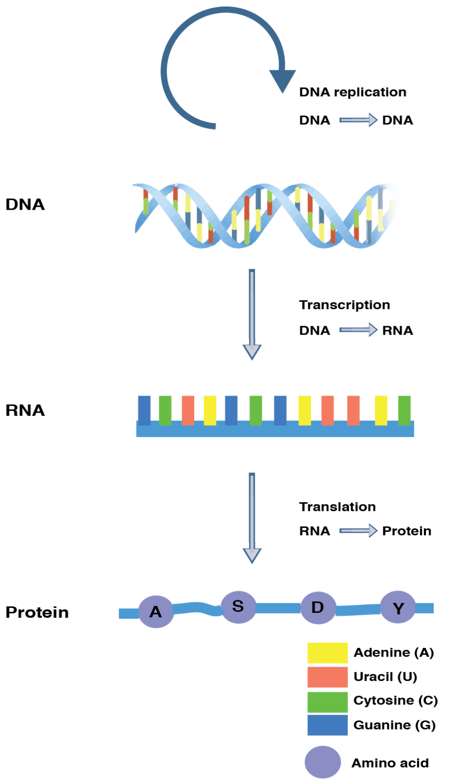
1.2 GENE EXPRESSION AND REGULATION IN EUKARYOTES

The genetic blueprint of a cell is encoded in its genome and this defines its molecular potential. Given that genomes act as blueprint, the encoded information must be transmitted into functional molecular entities in order for it to be useful. In eukaryotes, the flow of information from the genome to the final functional elements is commonly referred to as the central dogma (see figure 1a) and entails RNA transcription, processing, transport, stabilization, translation and protein activation (Alberts et al. 2014). This multistep process is complex and can in principle be regulated at any of the enlisted steps. However, transcriptional control is most critical since it prevents wasteful production of transcripts in a given cell (Alberts et al. 2014).

The set of transcribed genes and their corresponding expression levels are important in determining cells properties. This is controlled at multiple levels such as epigenetic, transcriptional and post-transcriptional regulation (Coulon et al. 2013; Holoch & Moazed 2015; Klemm et al. 2019; Spitz & Furlong 2012). The epigenetic landscape of a cell is an important factor involved in defining the differentiation state. Given the tight packaging of DNA around nucleosomes and higher order chromatin structures, transcription involves remodeling of the closed chromatin region (heterochromatin) into open euchromatin to facilitate accessibility by the transcription machinery (Alberts et al. 2014). Additionally, the overall chromatin accessibility is key in determining the transcriptional activity of a gene. ATP dependent chromatin remodelers facilitate this through post-translational modification of histones, histone variants, DNA modification and non-coding RNAs (Alberts et al., 2014; Chen & Dent, 2014). The epigenetic status of a genic region is dynamic and might be defined in a spatial-temporal fashion depending on the cell type/state, external stimuli, position, and differentiation stage among others (Alberts et al. 2014; Klemm et al. 2019).

For euchromatic genes to be transcribed, general transcriptional factors (GTFs), RNA polymerases and several other regulatory proteins are involved (Alberts et al., 2014). In the nucleus, genes are transcribed into RNA transcripts that can function as intermediates for protein synthesis (messenger RNA), regulatory RNAs (lincrRNA, microRNA), processing factors (small nuclear RNAs), RNA transport, translation (ribosomal RNA and transfer RNA) and degradation (piRNA). RNA only occupies 10% of the cellular volume where the majority are rRNAs (about 80%) and mRNAs 5-10% (Alberts et al., 2014). Unlike prokaryotes, eukaryotes have three RNA polymerases namely: I, II and III. Only RNA polymerase II (Pol II) catalyzes mRNA synthesis. For Pol II to initiate transcription, the general transcription factors (TFIIA, TFIIB, TFIIC, TFIID, etc) are recruited to the proximal promoter thereby forming the basal transcription initiation complex (Alberts et al. 2014; Haberle & Stark 2018). For most genes, transcription is further regulated by multiple *cis* regulatory elements that are bound by specific transcriptional factors (*trans* acting factors) (Alberts et al. 2014; Haberle & Stark 2018; Zabidi & Stark 2016). Additionally, co-activators and co-repressors may play a role (see figure 1b). Pol II recruitment is critical and normally effected by several specific transcriptional factors (STFs) acting proximally or distally to the target gene. DNA looping makes it possible for distally located STFs to regulate genes independent of direction and distance. Additionally, insulators demarcate transcriptional domains to prevent non-specific action. These STFs act in combinatorial and context dependent fashion to enable diverse responses to stimuli. Regulatory factors may influence gene transcription in a variety of ways namely: remodeling chromatin structure, influencing transcriptional machinery assembly, control Pol II release and pausing (Alberts et al., 2014).

Once Pol II driven transcription terminates, the new transcripts are processed into mature mRNA for transport and translation in the cytoplasm. This processing adds another layer of regulation and involves for example 5'- capping, 3'-polyadenylation and splicing (Alberts et al. 2014; Hocine et al. 2010). For instance, alternative splicing allows for a single gene to encode distinct isoforms which would generate variant while capping and polyadenylation are key in the transport and stabilization of RNA transcripts.



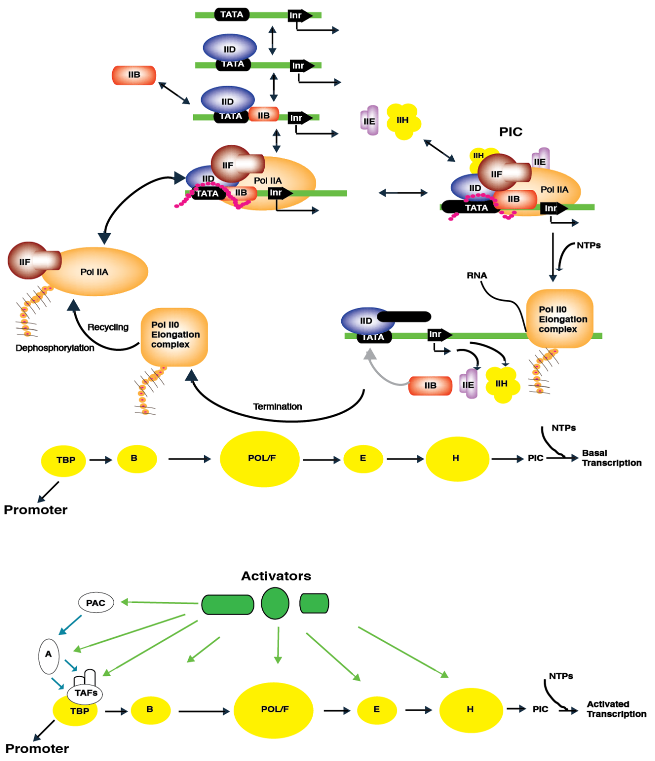


Figure 1a: An illustration of the central dogma of molecular biology

Figure 1b: An illustration of the process and components of transcription initiation machinery in eukaryotes

1.3 DNA SEQUENCING

De novo sequencing of the human genome, several model organisms and important disease causing agents were carried out by the laborious, expensive and low throughput Sanger sequencing technique using a whole genome/chromosome shotgun approach (International Human Genome Sequencing Consortium 2001; International Human Genome Sequencing Consortium 2004; Adams et al. 2000; Arabidopsis Genome Initiative 2000; C. elegans Sequencing Consortium 1998; Mouse Genome Sequencing Consortium et al. 2002; Fleischmann et al. 1995; Gardner et al. 2002). The emergence of massively parallel sequencing technologies in the mid-2000s instigated the genome-wide molecular examination of biological processes. This resulted in a massive increase in throughput, reduction in labour and reductions of up to 50,000-fold in sequencing cost relative to the human genome project (Goodwin et al. 2016). These techniques that have been referred to as next generation sequencing (NGS) technologies and each has its distinguishing chemistry. Their marked leap from the first generation chain termination approaches such as Sanger sequencing was the ability to carry out massively parallelized reactions and detect signals in miniaturized reaction centers without the DNA cloning step (Goodwin et al. 2016; Metzker 2010; Shendure et al. 2017). These advances in NGS technologies have played a critical role in opening up genomics field for intriguing biological pursuits such as genetic variation, epigenomics and

RNA expression (Metzker 2010; Shendure et al. 2017; Shendure et al. 2019). Most of these studies were based on an ensemble of cells that made it unattainable to tackle studies that required single cell resolution. However, the last decade has ushered in genome-wide single cell methods that utilize NGS in their workflow to interrogate gene expression (Sandberg 2013; Shapiro et al. 2013).

Roche/454 pyrosequencing NGS platform was the first to be commercially distributed and followed by multiple technologies namely: Illumina/Solexa sequencing by synthesis, Ion Torrent, Helicos, and BGI's Complete Genomics (Goodwin et al. 2016; Metzker 2010). These have been referred to as the second generation (or short reads) sequencing technologies and vary in their chemistry, read lengths, error rates, throughput and base calling. 454 platform utilizes emulsion PCR to clonally amplify DNA template attached to a bead in nano-sized well then detect emitted light signal from a luciferase and pyrophosphate driven reaction cascade. Ion torrent detects pH changes from the released H+ ions. Illumina sequencing by synthesis uses a cyclic-reversible termination strategy with imaging of the added flour-bound deoxynucleotide.

While the second generation NGS platforms rely on clonally amplified DNA fragments to generate sufficient signal for detection, third generation methods sequence very long reads without the need for amplification. The two prominent technologies are real-time single-molecule sequencing and nanosequencing by Pacific Biosciences (PacBio) and Oxford Nanopore Techonogies (ONT). PacBio relies on optical detection of the flourophore signal released from incorporated nucleotide during DNA polymerase synthesis of the growing DNA strand (Goodwin et al. 2016). Unlike the short reads sequencing-by-synthesis approaches, the target DNA template strand is bound to a fixed DNA polymerase and using zero mode waveguide ensures that only one nucleotide is added and detected per time. ONT exploits the ionic changes that occur when single DNA strand is passed through tiny protein pore and reads template sequence based on the generated electric signal (Goodwin et al. 2016). Unlike other platforms ONT reads the native sequence of the template DNA and does not depend on detecting signals such as pH, color or light. Since ONT relies on electric signal detection there is no need for highly specialized optics and the platform is therefore attractive in remote settings with poor infrastructure. Both PacBio and ONT technologies generate very long sequence reads, typically 10kb or more, and may facilitate the detection of covalent DNA modifications (Flusberg et al. 2010; Wescoe et al. 2014). However, in comparison to the second generation NGS platforms they still suffer from substantial error rates. Hence these platforms could complement other NGS platforms such as Illumina. With continued improvements in terms of throughput and sequencing errors minimization, these technologies might catalyze genome-wide studies such as structural variation, splicing and allelic variation at the single-cell level. The recent acquisition of PacBio by Illumina could potentially lead to improved long reads sequencing platforms (Eisenstein 2019).

2 SINGLE-CELL GENE EXPRESSION ANALYSES

2.1 INTRODUCTION

Over the last decade, scRNA-seq has blossomed and transitioned from a proof-of-concept to addressing intriguing biological questions in development (Petropoulos et al. 2016; Xue et al. 2013), gene regulation (Deng et al. 2014; Reinius et al. 2016) and infectious disease (Poran et al. 2017; Reid et al. 2018) among others. scRNA-seq can elucidate regulatory networks which might not be possible with the ensemble-based methods, for example in instances where two factors regulate a set of genes independently (Shalek et al. 2014; Trapnell et al. 2014; Zeisel et al. 2015). Another emerging and equally exciting application is the inference of transcriptional kinetic parameters at allelic resolution (Larsson et al. 2019). However one of the most established applications of scRNA-seq has been in the characterization and discovery of sub-populations of cells and transitory developmental states (Wagner et al. 2016).

2.2 TRADITIONAL METHODS FOR SINGLE-CELL GENE EXPRESSION ANALYSES

In the early years of single-cell gene expression profiling, low throughput techniques were commonly used to profile a target set of genes. These included the use of fluorescent reporter constructs (Chalfie et al. 1994; Young et al. 2012), quantitative PCR (Bengtsson et al. 2008; Taniguchi et al. 2009; Warren et al. 2006; Wills et al. 2013), and single-molecule RNA FISH (Femino et al. 1998; Raj et al. 2008). While these methods have facilitated some critical findings (Wills et al. 2013), they are limited to the detection of few numbers of genes. Following the emergence of microarray technique, single cell profiling of all annotated transcriptomes after amplification of the cellular transcripts was implemented on this platform (Kamme et al. 2003; Kurimoto et al. 2006; Subkhankulova et al. 2008). However, the appearance of massively parallel sequencing platforms ushered in RNA sequencing initially for bulk then single cell profiling of transcriptomes in unbiased and high throughput fashion. The first single cell RNA sequencing (scRNA-seq) study was published in 2009 (Tang et al. 2009) and followed by several improved techniques (Kolodziejczyk et al. 2015).

2.3 SINGLE-CELL RNA-SEQUENCING

Over the past decade, several protocols have been developed for capturing, converting and amplifying the limited mRNA transcripts from single cells (Lafzi et al. 2018; Picelli 2016; Chen et al. 2018; Svensson et al. 2017). Out of these, some have gained wide prominence namely: CEL-Seq (Hashimshony et al. 2012), CEL-Seq2(Hashimshony et al. 2016), Drop-seq (Macosko et al. 2015), MARS-seq (Jaitin et al. 2014), STRT (Islam et al. 2011; Islam et al. 2014), STRT-2i (Hochgerner et al. 2017), Smart-seq (Ramskold et al. 2012), Smart-seq2 (Picelli et al. 2013), SCRB-seq (Bagnoli et al. 2018; Soumillon et al. 2014), Quartz-seq (Sasagawa et al. 2013), Quartz-seq2 (Sasagawa et al. 2018), inDrop (Klein et al. 2015), sci-RNA-seq (Cao et al. 2017) and SPLiT-seq (Rosenberg et al. 2018). While these protocols differ in their processing steps, the samples are often derived from tissues, organs, cultured cells or microbes. In the case of solid tissues, the first step is to disaggregate the sample into small pieces and then dissociate it into single cells suspension using proteases. Individual cells are then isolated from the suspension using one of the different methods namely: manual picking (using micropipettes), fluorescence-activated cell sorting (FACS) into plates or automated capture in nanolitre chambers in droplets or microwells. Additionally, *in vitro* cell pickers combined with computer vision techniques are becoming useful (Környei et al. 2013; Ungai-Salánki et al. 2016). These approaches differ in terms of their throughput, reagent volumes and labour intensity. The recent advent of single-cell combinatorial indexing techniques in combination with multiple rounds of sample pool-split strategy have made it possible to profile large number of single cells without the need for specialized equipment for automated (Cao et al. 2017; Rosenberg et al. 2018).

Once isolated, individual cells can be lysed using hypotonic solution in the presence of RNAse inhibitors to prevent transcripts degradation. A majority of protocols target polyadenylated mRNA transcripts using polyT oligonucleotides primers and then generate complementary DNA strands (cDNA) through a reverse transcriptase catalyzed synthesis process. There are three main second strand synthesis strategies: poly A tailing using terminal transferase activity (Quartz-seq, Quartz-seq2 and Tang et al), template-switching (Smart-seq, Smart-seq2, Chromium (10x Genomics) (Zheng et al. 2017), STRT-seq, Seq-Well (Gierahn et al. 2017)) and combination of Ribonuclease (RNase H) with DNA pol I from E. coli (CEL-seq, CEL-seq2, MARS-seq, inDrop and sci-RNA-seq). Given the limited amount of material the cDNA is amplified using an *in vitro* transcription (IVT) strategy (CEL-seq, CEL-seq2 and inDrop) with remaining ones using PCR approach. Unlike PCR approaches IVT attains linear amplification hence reducing amplification bias. However, the multiple steps involved IVT approach is time consuming. While Quartz-seq, Smart-seq and Smart-seq2 capture fragments from across the entire transcript length the other protocols only tag 5'- or 3'-ends and referred to as end counting methods. In order to overcome amplification bias, the end counting methods often use Unique Molecular Identifiers (UMIs) that are short nucleotide sequences which act as barcodes for keeping track of the absolute counts of transcripts pre-amplification. While the use of UMIs makes end counting methods attractive for studies that mainly depend on quantifying gene expression levels, the inability to sample fragments from the entire transcript body is a drawback for studies aimed at allelic expression resolution and splicing isoform analysis.

In all protocols, profound caution is required when preparing and processing the libraries in order to impede the loss of already limited material. Some of the measures often taken include pooling of indexed samples, combining multiple steps into single reaction, avoiding unnecessary transfer of reaction mixes and working in ice-cold conditions when processing samples.

In spite of the measures described above, technical noise is inevitable in scRNA-seq and specific experimental & analysis strategies must be considered when designing studies so as to minimize their impact on the biological signal. One of the common strategies aimed at tackling technical noise is to include a set of spike-in control mRNAs, to experimentally distinguish the technical noise level from biological variability (Brennecke et al. 2013; Grün et al. 2014; Grün et al. 2015). Spike-ins are exogenous RNAs added in known amounts per cell lysate before cDNA generation and processed together with the endogenous mRNA transcripts. Since they are added in equal amounts per sample any variability in their expression measurements should represent the magnitude of technical noise. Ideally, the abundances of individual spike-in species should span the dynamic range of the target cells gene expression, reflect the endogenous genes properties (such as GC content, polyA tail length and gene size) and be properly calibrated. One of the commonly used spike-in is ERCC's synthetic set of 92 bacterial derived transcripts (Consortium et al. 2005; Grün et al. 2014). Some experiments have demonstrated the use of transcripts from species unrelated to the cells under investigation (Brennecke et al. 2013). While broadly used in scRNA-seq, spike-ins often have pitfalls and deconvolving technical noise from their measurements can be difficult. Their properties could be non-reflective of the endogenous transcripts (e.g. shorter polyA tails in ERCC spike-ins), cannot track any variations that occur before the reverse transcription step and are usually technically challenging to calibrate especially for small cells with minute amounts of RNA. In some studies, modeling the relationship between mean and the coefficient of variation of endogenous genes and statistically calculating candidates with noise beyond the expected variance can be useful (Buettner et al. 2014).

Several factors contribute towards technical noise in scRNA-seq studies, namely: non-uniform cell lysis, RT efficiency, amplification, sequencing, contaminations and processing batch effects (Grün et al. 2014; Vallejos et al. 2015; Wagner et al. 2016). Due to the small amounts of starting material, transcript dropouts (i.e. genes that are transcribed but fail to be captured due to technical reasons) are prevalent, hence leading to sparse gene expression matrices. This is a hallmark of scRNA-seq data and poses vital analytical challenge that must be taken into consideration (Kharchenko et al. 2014). Dropout rates vary across protocols though it is higher per cell in the less sensitive end counting methods (estimated sensitivity ranges between 5% and 40%). Full-length coverage by protocols such as Smart-seq2 (Picelli et al., 2013) sample fragments from the entire transcript body thus substantially improving their sensitivity per cell. This makes them suitable for studies requiring allele-resolved expression or the detection of lowly expressed genes. However, the recent upscaling of cell numbers by end counting methods on droplet microfluidics (Zheng et al., 2017) and combinatorial indexing implementations (Cao et al. 2017; Rosenberg et al. 2018), presently in the thousands and more, augments their power to detect technical dropouts. The recent adoption of split-pool strategy with combinatorial indexing, substantially increases throughput in terms of cell numbers without the requirement for specialized cell isolation equipment. This may make these approaches attractive to low-resourced research units and for studies targeting enormous cell numbers such as whole tissue or organism profiling (Cao et al. 2019).

Given the substantial number of scRNA-seq protocols it is important to reflect on which method would be most suitable in a study. Some of the guiding factors could be: the research question at hand, financial standing, accessibility to specialized equipment and computational resources. Previous benchmarking efforts may also provide useful insights (Svensson et al. 2017; Ziegenhain et al. 2017). For some research pursuits, two or more protocols might be beneficial. For instance, in subtype discovery end counting protocols on automated platform may be used initially to generate low coverage transcriptomes from large cell numbers at shallow sequencing. Next, applying a high-depth/sensitive protocol to a subset of cells would give an exhaustive molecular understanding of the biological quest.

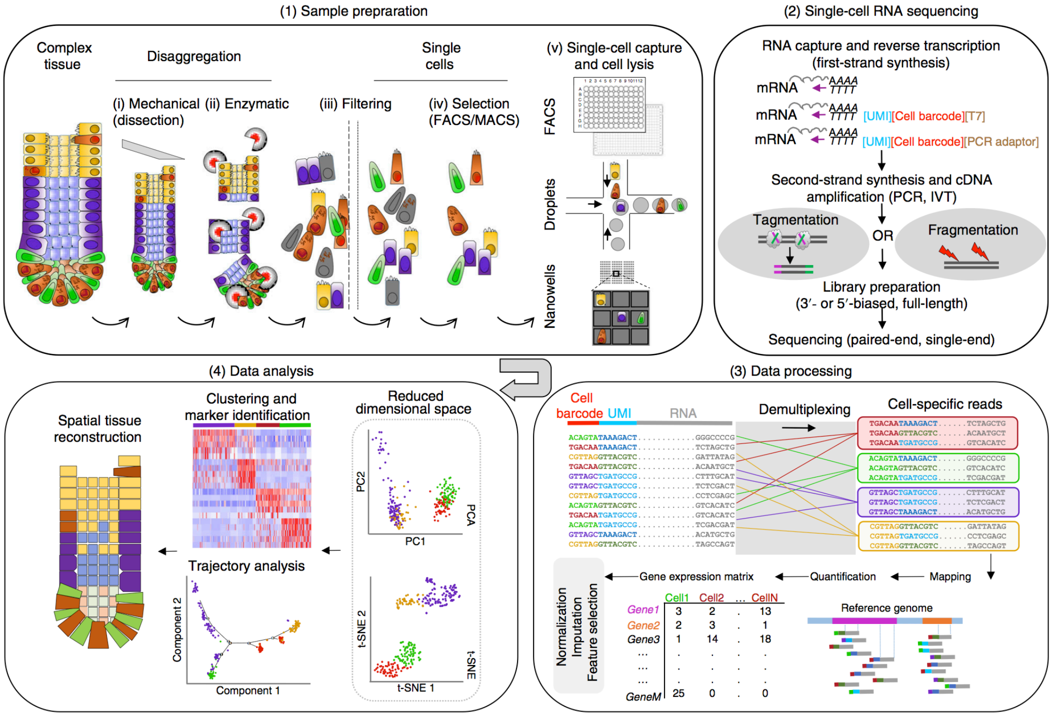


Figure 2: The single-cell RNA sequencing process. The design of single-cell transcriptomics experiments includes four major phases: (1) During sample preparation, cells are physically separated into a single-cell solution from which specific cell types can be enriched or excluded (optional). After they have been captured in wells or droplets, single cells are lysed, and the RNA is released for subsequent processing. (2) To convert RNA into sequencing-ready libraries, poly(A)-tailed RNA molecules are captured on poly(T) oligonucleotides that can contain unique molecular identifier (UMI) sequences and single-cell-specific barcodes (5′- and 3′-biased methods). To allow for subsequent amplification of the RNA by PCR or IVT, adaptors or T7 polymerase promoter sequences, respectively, are included in the oligonucleotides. After RT into cDNA and second-strand synthesis (optional), the transcriptome is amplified (PCR or IVT). For conversion into sequencing libraries, the amplicons are fragmented by enzymatic (e.g., tagmentation) or mechanical (e.g., ultrasound) forces. Sequencing adaptors are attached during a final amplification step. Full-length sequencing can be carried out, or 5′ or 3′ transcript ends can be selected for sequencing using specific amplification primers (optional). For most applications, paired-end sequencing is required. (3) The sequencing reads are demultiplexed on the basis of cell-specific barcodes and mapped to the respective reference genome. UMI sequences are used for the digital counting of RNA molecules and for correction of amplification biases. The resulting gene- expression quantification matrix can subsequently be normalized, and missing values imputed, before informative genes are extracted for the analysis. (4) Dimensional-reduction representations guide the estimation of sample heterogeneity and the data interpretation. Data analysis can then be tailored to the underlying dataset, which allows cells to be clustered into potential cell types and states, or ordered along a predicted trajectory in pseudotime. Eventually, the spatial cellular organization can be reconstructed through the interrogation of marker genes (experimentally) or through marker-guided computational reconstruction (inference). PC, principal

component. Adapted by permission from Springer Nature: Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies, Lafzi, A. et al., 2018. Nature protocols, 13(12)

2.4 COMPUTATIONAL ANALYSES OF SINGLE-CELL RNA-SEQUENCING

With the expansion of scRNA-seq, the demand for suitable computational techniques for processing, analyzing and interpreting the data has intensified. This has culminated into the development of several algorithms and tools addressing the specific analytical and technical challenges in the single cell gene expression domain (Stegle et al. 2015; Wagner et al. 2016; Zappia et al. 2018). Currently, the "scRNA-tools" database has 393 applications enlisted (https://www.scrna-tools.org/) and these are addressing 32 different functional categories namely allele specific analysis, clustering, cell cycle, normalization, imputation and visualization among others (see figure 3). It is noteworthy that a substantial number of the tools were classified in two or more categories. For instance Scanpy, a python-based scalable toolkit (Wolf et al. 2018) made up of APIs for Quality Control, Normalisation, Gene Filtering, Clustering, Ordering, Differential Expression, Variable Genes, Dimensionality Reduction, Visualisation and Simulation. Seurat (Satija et al. 2015; Butler et al. 2018) also provides access to single cell analysis tools with similar functions as Scanpy within the R platform. These suites provide convenience and enhance efficiency in scRNA-seq data analysis by serving as 'one-stop shop' for the relevant tools and also reducing the tedious exercise of formatting data to be compatible with the different applications in the workflow.

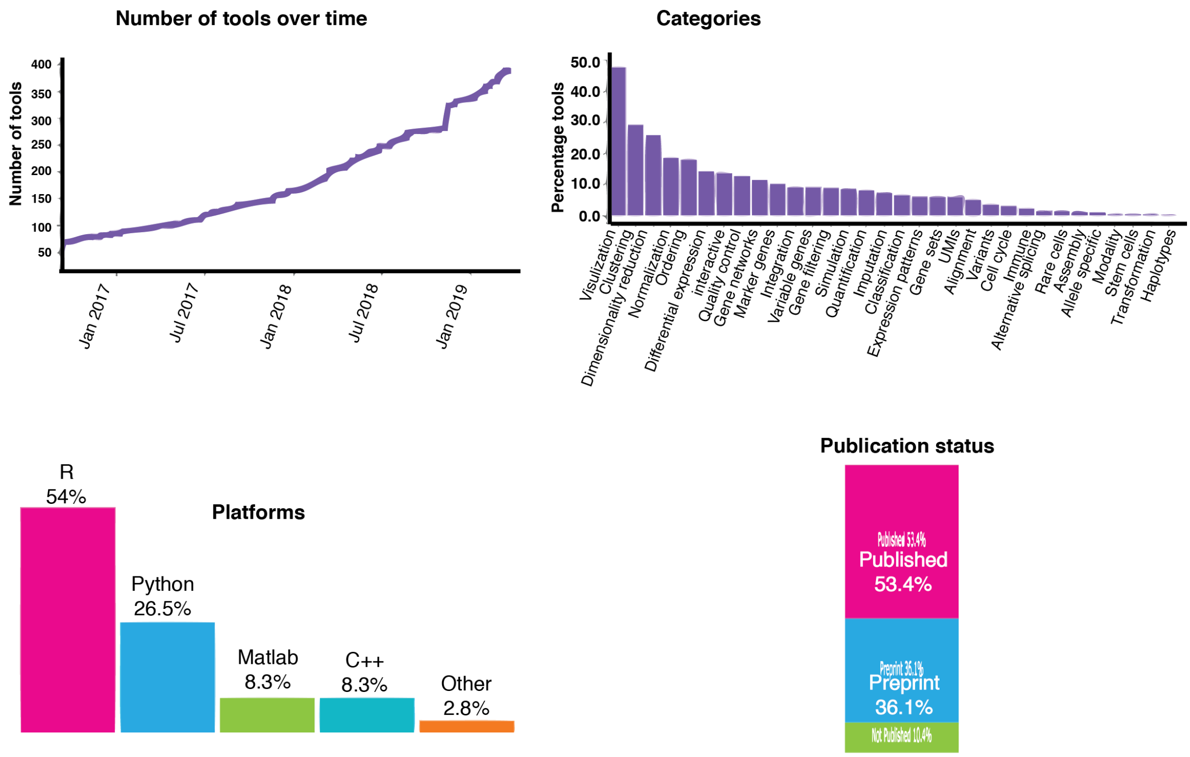


Figure 3: Statistics of tools from scRNA-tools database

Quality control assessment and gene expression quantification

It is common practice to pool several scRNA-seq libraries for sequencing, hence the very first analysis step is to demultiplex the resulting reads into respective samples based on their cell-specific molecular indexes. Illumina's bcl2fastq tool can demultiplex samples using the user provided sample indices in addition to converting the base call BCL files into the fastq format. However, with sufficient computational skills, bioinformaticians/analysts can code scripts to accomplish sample demultiplexing on unix-based platforms or other programming languages. For the UMI-based methods, identical UMIs are further collapsed into their unique tags to gain the absolute counts of the transcripts. Here tools such as: CEL-seq2, Cell Ranger, Drop-seq tools, zUMIs and UMI-tools have been used (Hashimshony et al. 2016; Zheng et al. 2017; Macosko et al. 2015; Parekh et al. 2018; Smith et al. 2017). These tools differ in their methods for determining identical barcodes, the scRNA-seq protocols and mapper compatibility among others (Parekh et al. 2018) and some provide additional preliminary quality evaluation functions. For instance, zUMIs provides read distributions, read downsampling, sample variation, number of genes detected, and is compatible with a large number of the scRNA-seq protocolsb(Parekh et al. 2018). For droplet-based techniques, the challenge of doublets is significant, hence different approaches have been implemented to alleviate this problem. While demuxlet (Kang et al. 2018) uses SNP genotype information to predict doublets, DoubletFinder (McGinnis et al. 2019) uses an average transcriptome created from single cell expression pairs and Scrublet (Wolock et al. 2019) simulates multiplets from the input expression data and uses nearest neighbour classification.

Next, reads are mapped to a reference genome or transcriptome sequence for gene quantification. Aligners have to tackle the challenge of efficiently mapping large number of short reads while addressing the potential errors from sequencing. Here short-read, splice aware aligners have been most popular for RNA-seq (Garber et al. 2011; Baruzzo et al. 2017). Most of these were initially developed for processing bulk-RNA-seq data but have subsequently been adopted in scRNA-seq analysis. The applications adopt an efficient data structure for efficient searches and a divide and conquer strategy to enable efficient memory use. These aligners differ in terms of memory requirements, speed and accuracy among other metrics and benchmarking studies have been done (Baruzzo et al. 2017). Two of the most commonly used aligners are STAR (Dobin et al. 2013) and HISAT (Kim et al. 2015) both have implemented a two-pass strategy to allow for improved alignment of reads with short anchor fragments across splice junctions. The former uses suffix arrays to match reads to the indexed reference and is one of the fastest mappers (Baruzzo et al. 2017) though requires large RAM. For less memory intensive mapping HISAT, which uses a novel approach based on Burrows-Wheeler Index and FM indexing, would be a better option. The emergence of pseudo aligners such as Salmon (Patro et al. 2017) and Kallisto (Bray et al. 2016) capable of gene quantification without explicitly mapping reads may improve efficiency in addressing single cell questions involving thousands and even million of cells that may not need exhaustive expression measurements such as subtype discovery. For the end counting protocols with UMIs it is important to remove the barcodes before aligning the reads.

The long reads from the non-clonal amplification NGS platforms pose unique challenge hence the emergence of various approaches to tackle this (Chaisson & Tesler 2012; Li 2018). Additionally, memory efficient algorithm that can be executed on mobile devices for handling ONT data has been developed (Gamaarachchi et al. 2019). This makes it suitable for the remote settings that lack memory efficient computing platforms. Once mapping is complete gene/transcript quantification is carried out using features counting tools such as HTseq-counts (Pyl et al. 2014) to generate an expression matrix. For the end counting protocols with UMI-barcoded oligos, counts are derived from number of unique UMI barcodes mapped towards a gene but caution must be observed to prevent ghost UMIs resulting from PCR or sequencing error and barcode collisions (Islam et al., 2014).

Quality control (QC) assessment of scRNA-seq data forms the foundation for generating accurate results from experiments hence several metrics must be evaluated to ensure data quality. These include total number of reads, library duplication rate and complexity, reads mapping rates (uniquely mapped, multimaps, ratio of exon to intergenic mapped reads), contaminations from non-target organisms and number of detected genes. Some of the QC analysis tools are FastQC (Andrews 2016) and Kraken (Davis et al. 2013). In experimental designs in which samples have RNA controls the ratio of reads mapped to endogenous genes and spike-ins can be an informative metric for library quality i.e. low quality cells tend to have lower ratio. Additionally spike-ins may be useful for estimating the RNA starting amounts per cell. However, this could mislead in instances of small sized and/or low transcribing cells where the ratio reflects the biological state. In mammalian cells low expression of the core nuclear genes with concomitant high expression of mitochondrial genes has been associated with cells undergoing apoptosis and could be a signature of low quality libraries. After gene counts have been normalized clustering samples based on their correlations distances or using dimensionality reduction methods (e.g. PCA) while overlaying categorical meta information (such as batches, number of genes detected) can help uncover low quality cells which may stick out as outliers.

In practice evidence from multiple metrics is often taken into account before discarding the low quality cells from downstream analysis and it has been demonstrated that certain quality metrics such as number of genes detected and library mapped read have a prominent association with the leading principal components (Wagner et al. 2016; Gaublomme et al. 2015).

Batch effects are systemic factors usually caused by technical variations in the scRNA-seq experimental procedure. However these can also be biological such as the cycling state of proliferating cells or variation by the donor in human derived cells. One of the recommendations is to adopt balanced design whenever possible in performing scRNA-seq experiments. For instance in treatment versus control study, mixed subsets of samples from the two categories should be processed together on independent experiments. Once the normalized single cell expression has been generated, it common for the samples (cells) to be visualized in lower dimensional subspace, using techniques such as PCA, then superimposing meta data so as to detect any systemic patterns that are non-biological. With the aim of overcoming the subjective nature of the visualization approach, novel technique referred to as kBET (Büttner et al. 2019) uses nearest neighbour classification after singular value decomposition to statistically test if a given variable(meta information) has significant batch effect. Additionally different computational methods have implemented techniques such as linear regression in ComBat (Jaffe et al. 2012), nearest mutual neighbour (Haghverdi et al. 2018) and canonical correlation analysis in Seurat (Butler et al. 2018). However, this remains an active area of research and robust benchmarking analysis may help reveal the strengths and weakness of some of these batch correction methods in the context of scRNA-seq.

Sub-population discovery

Unbiased discovery of sub-populations is one of the most dominant applications of scRNA-seq and in well-defined cell lineages, types or states the method is sufficiently robust. However in the context of reduced boundaries between subpopulations separating this groups may pose a challenge and one of the reasons would be the curse of dimensionality(Kiselev et al. 2019). As the number of features (genes) quantified increases the distance between data points (cells) becomes smaller resulting in poorly defined groupings. In order to overcome this different schemes have been used including: dimensionality reduction techniques, feature selection, down sampling strategies and iterative approaches. The common dimensionality reduction methods used are principal component analysis (PCA), t-distributed Stochastic Neighbour Embedding (tSNE) (van der Maaten & Hinton 2008) and more recently UMAP(Becht et al. 2019). There are studies that have combined either of these methods to detect subtypes for example in Petropoulos et al (Petropoulos et al. 2016) the cells were reduced into lower dimensions based on PCA then projected onto tSNE subspace. The methods are normally applied to a subset of genes that may be known drivers of specific biological phenotype (Durruthy-Durruthy et al. 2014), highly variable genes or differentially expressed genes identified from complementary techniques (e.g. bulk-RNA-seq, in situ RNA hybridization). In selecting most variable genes the commonly implemented strategy is to model mean-coefficient of variance (CV) dependency using spike-in controls (or even endogenous genes) in order to capture the expected variability (Brennecke et al. 2013). Each gene may then be tested for statistical significance in variability beyond the technical noise (null). For big cells with sufficient amounts of RNA pool-and-split strategy can be adopted in modeling technical dropouts (Deng et al. 2014).

scRNA-seq data is sparse in nature due to the high dropouts. Several approaches have been implemented to mitigate their impact. Mixed models have been used to capture the distribution resulting from both successfully sequenced transcripts and dropouts (Kharchenko et al. 2014). Data imputation and false-negative curves have also been utilized (Wagner et al. 2016).

Making sense of scRNA-seq data is complicated due to both technical and biological several factors. Transcriptomes complexity, library quality, impact of batch confounders and quantity of cells are some of the factors contributing to the challenge. The degree of challenge also varies with cells types with certain cells being more challenging to profile and analyze. However, with the unrelenting improvement in both experimental and computational methods in scRNA-seq fascinating and important biological pursuits are set to be unraveled.

3 PROJECT SPECIFIC BACKGROUND

In this section I will provide the relevant background to all the different studies that form part of my thesis.

3.1 *P. FALCIPARUM* AND MALARIA

Malaria

Malaria poses heavy health and economic burden globally and is still responsible for close to half a million deaths annually mostly in children under five years (Ashley et al. 2018; Bousema et al. 2014). Encouragingly the past decade has witnessed a steady decline in disease prevalence with up to 60% drop in deaths recorded in some endemic countries (Feachem & Sabot 2008). This maybe attributed to the complex interplay between biotic (such as vector management, use of efficacious drugs and education) and abiotic factors though compelling evidence is lacking (Snow et al. 2017). The reduction in disease prevalence notwithstanding, numerous challenges still exist including the emergence of resistance to anti-malarial drugs (Miller et al. 2013). Hence there is need for innovative technologies to enable in-depth understanding of both the disease and its etiological agent. This may culminate in the detection of potential targets for therapies and/or vaccines.

Malaria is caused by Apicomplexans from the genus *Plasmodium* with five species infecting human namely: *P. malariae, P. ovale, P. knowlesi, P. vivax* and *P. falciparum* (Ashley et al. 2018; Kantele & Jokiranta 2011). Out of these *P. falciparum* is the most virulent accounting for about 90% of malaria-associated mortalities (Murray et al. 2012).

P. falciparum life cycle is complex and constitutes two main development phases: asexual and sexual that occur in human and anopheline mosquito respectively (Bousema et al., 2014). The cycle comprises several stages with distinct morphology and cellular functions that mediate parasite's survival in different host niches (see figure 4). While several of these have been targets for developing drugs and vaccines (Delves et al. 2012; Todryk & Hill 2007), the IDC (intra-erythrocyte development) stage, which is also referred to as blood phase, is responsible for the clinical symptoms linked to the disease. This 48-hour cycle occurring within the host erythrocytes starts when merozoite invades the RBCs (red blood cells) followed by three cytological stages namely: ring (early), trophozoite (mid) and schizont (late). Each of these is further subdivided into early and late phase that accomplish unique cellular function(s) (Bousema et al., 2014).

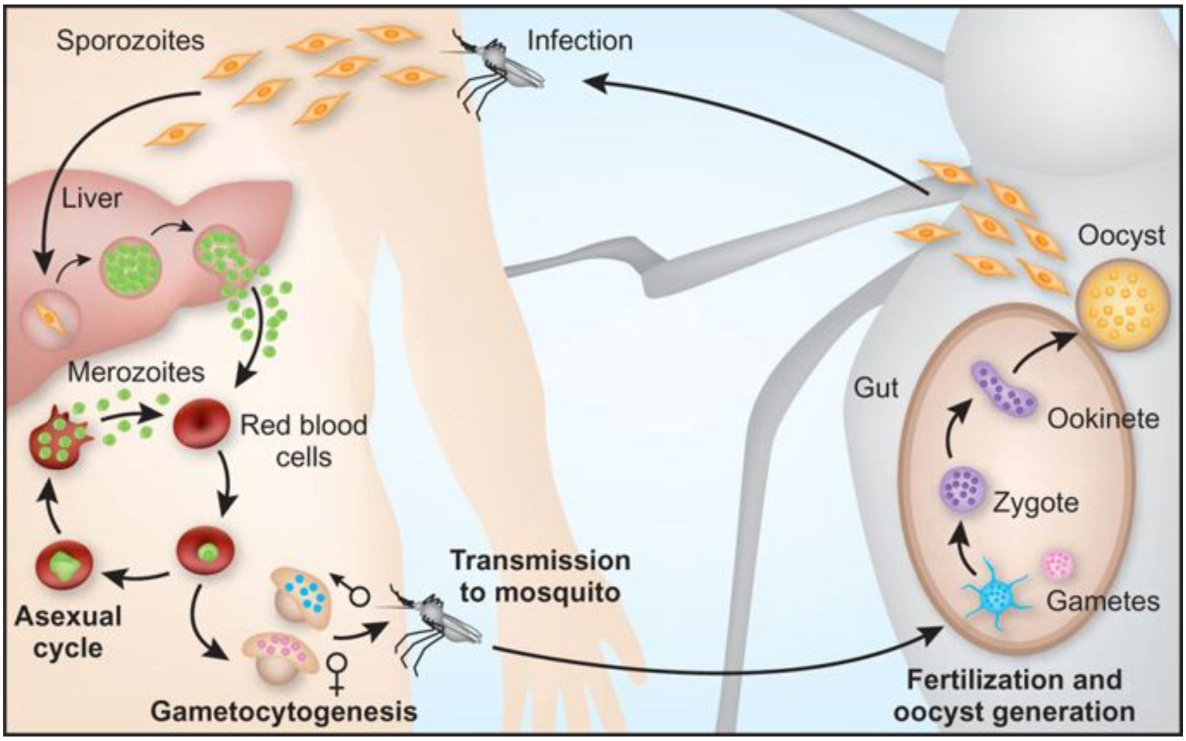


Figure 4: Illustrating *P. falciparum* life cycle. The female Anopheles mosquito transmits sporozoites of the malarial parasite to the human host. After a period of maturation in the liver, merozoites are released into the blood; these invade red cells as part of the asexual (erythrocytic) cycle, and the sexual male and female gametocytes are generated from the merozoites. During a subsequent blood meal, a mosquito takes gametocytes into the midgut, leading to macro (female) and micro (male) gametes, which, after fertilization and zygote formation, produce an ookinete that penetrates the mosquito gut wall and generates oocysts containing sporozoites. Adapted by permission from Springer Nature: Protective hemoglobinopathies and *P. falciparum* transmission, Pasvol, G., 2010. . Nat Genet, 42(4)

Gene expression regulation in *P. falciparum*

To achieve the remarkably complex life cycle *P. falciparum* is endowed with a fine-tuned gene expression control program. This regulatory program occurs at multiple levels namely: epigenetic, transcriptional, post-transcriptional and translation (Painter et al. 2011). Like other eukaryotes a coordinated interplay between these multiple levels define the expression outcome.

The genome encodes the basal transcriptional machinery constituting Pol II, general and specific transcription factors. However, to date only 27 *trans* acting factors have been annotated and these belong to the plant-like Apatela factors referred to as ApiAP2 factors (Painter et al. 2011; Painter et al. 2018). With this limited number of regulatory factors it is possible that post-transcriptional and epigenetic may be involved in controlling the expression of most genes. Epigenetic regulation directs the mutual exclusive expression of the 60-member *var* gene family that are important for host immune evasion and confer parasite virulence (Lopez-Rubio et al. 2009). Heterochromatin marks have been found to correlate with the expression of clonally variant gene families located in the sub-telomeric regions of chromosomes (Anon 2009). Intriguingly, active epigenetic marks and the modifying enzymes have been shown to correlate with the dominant *var* gene and facilitate clonal variant switching (Freitas-Junior et al. 2005; Lopez-Rubio et al. 2007; Petter et al. 2011; Volz et al. 2012). Epigenetics also play a role in regulating the other variant gene families (*rif* and *stevor*) during IDC (Guizetti & Scherf 2013; Kirkman & Deitsch 2012). Though not well-studied splicing is believed to contribute to the regulatory framework (Horrocks et al. 2009).

Non-coding RNAs play significant regulatory role in eukaryotes and *P. falciparum* snRNA, snoRNAs, telomerase RNAs and NATs (natural antisense transcripts) have been described (Vembar et al. 2014). Adjacently located ncRNAs could be involved in the allelic regulation of *var* genes and NATs are also assumed to have a repressive effect on their corresponding genes (Vembar et al. 2014; Siegel et al. 2014). It is worthy to note that several levels of regulation can jointly control the expression of gene(s) to achieve a specific phenotype or function as evidenced by *var* genes control. With recent compelling evidence for chromosomal organization role in controlling virulence genes (Bunnik et al. 2019) it is suffice to infer that we are at the initial stages of understanding how this important parasite regulates its transcriptome in various context.

Transcriptional control of IDC

IDC is an important development stage with significant implications for the intervention and management of the disease. It is responsible for all the clinical symptoms attributed to malaria and the target for all effective drugs (Delves et al. 2012). Additionally it forms the foundation for a number of vaccine design strategies (Todryk & Hill 2007). Most studies reveal that over 90% of the genome is transcribed at this stage (Otto et al. 2010; Painter et al. 2018). Hence unraveling the underlying gene expression mechanisms provide an opportunity to understand parasite's biology at the transcriptional level.

IDC ring stage starts after merozoite invasion of erythrocytes during which the parasite remodels host's intracellular environment for its successful establishment. To achieve these complex transitions between functionally and morphologically divergent stages while being able to mount protective responses against external stress (e.g. host's immunity, temperature changes, drugs, etc) an elaborate molecular regulation program is essential (Bozdech et al. 2003).

P. falciparum genome is pervasively transcribed in IDC. This remarkable feat is accomplished in spite of the limited number of specific transcriptional factors characterized to date. The parasite has evolved an intricate transcription mechanism that has been described as a 'just-in-time' or 'transcripts-to-go' model in which genes are transcribed almost in an instant fashion at the time the proteins they encode are required (Bozdech et al. 2003; Otto et al. 2010; Le Roch et al. 2003). For instance, following erythrocyte invasion, ring stages express general cytoplasmic transcriptional and translational machinery, and then the metabolically active trophozoites express DNA replication and metabolism genes. At schizontal stage the parasite undergoes replication and initiate the expression of invasive genes in preparation for erythrocytes reinvasion.

Gene expression profiling has been key in providing the molecular understanding of IDC. The 1 hr time-scale resolution of *P. falciparum* HB3 strain using microarray technique has been the most comprehensive investigation of IDC (Bozdech et al., 2003). The study established that at least 60% of the genome is transcribed and most of the genes display periodicity in their expression with one minima and maxima peak per cycle. Additionally, it uncovered the outstanding cascades of stage-specific gene expression that mirrored the corresponding cellular and morphological function(s). This malleable transcriptional landscape has been reaffirmed in other studies (Le Roch et al., 2003; Otto et al., 2010) and was found to be conserved even in geographically diverse isolates (Rovira-Graells et al. 2012).

Understanding gene expression regulation during IDC has been of critical interest and currently there is growing evidence for multiple levels of control. *P. falciparum* genome encodes DNA-binding transcriptional factors with a putative AP2 DNA-binding domain which is common in plants transcriptional factors. These factors belong to the Apicomplexan AP2 (ApiAP2) protein family (Balaji et al. 2005) and are utilized in the classical *cis-trans* regulation of the core promoter complex during transcription. Most of them are expressed in a stage-specific fashion and have varying degree of importance in IDC progression as established in knockout experiments (Otto et al., 2010; Painter et al., 2011; 2018). However, how they regulate transcription initiation in IDC is still an open research question.

Post-transcriptional regulation is also crucial during IDC and it encompasses transcript maturation (5'-capping, 3'-polyadenylation, splicing, etc) and stability (rate of de novo synthesis and degradation) (Horrocks et al. 2009). The mRNA half-life has been established to change during IDC with transcript stability increasing from ring to schizont development stages (Shock et al. 2007; Sims et al. 2009). However, recent results from nascent transcription profiling study failed to support this across numerous genes (Painter et al. 2018). In regard to splicing, more than half of *P. falciparum* protein coding genes have introns (Gardner et al., 2002) and all the components of splicing machinery have been characterized. Though the exact mechanism of splicing regulation remains elusive, gene isoforms have been detected during IDC, which could represent an additional layer of regulation. In the study by Sorber et al (Sorber et al. 2011) a total of 405 splicing events were detected with majority being alternative 5'- or 3'- splice sites and 254 alternative isoforms. These events are speculated to generate truncated proteins or could be involved in fine-tuning gene expression profile in IDC. Long non-coding RNAs (lncRNAs) and natural antisense transcripts provide an additionally level of post-transcriptional regulation by silencing the translation of target transcripts through degradation or translational repression and have also been established to regulate IDC gene expression (Vembar et al. 2014).

Single cell transcriptomics in malaria

While scRNA-seq technique was becoming mainstream and impacting several domains in mammals and man it was not until 2017 that initial studies profiling *P. falciparum* at cellular resolution were published (Poran et al., 2017). Since then another single cell expression study looking mainly at sexual differentiation in *P. falciparum* has been carried out (Reid et al., 2018). However the application of scRNA-seq still lags far behind in comparison to mammalian or model organisms. This is attributable to the technical difficulties such as the tiny sized parasites with low amounts of RNA, extreme AT richness yielding low complexity libraries and dearth in scRNA-seq protocols optimized to process this parasite among others (Gardner et al. 2002; Nair et al. 2014; Reid et al. 2018). These challenges notwithstanding, scRNA-seq still holds immense promise in advancing several aspects of malaria biology including sexual development.

IDC development and sexual commitment have been the only studied aspect of the parasite using scRNA-seq methods. So far only lab-adapted isolates sampled at broad range of timepoints have been investigated. It is conceivable that scRNA-seq targeting shorter resolution times with improved number of detected genes may enhance the transcriptional resolution of both IDC and sexual commitment. In the Poran et al study where wild type isolates were profiled the use of less sensitive Drop-seq platform may have restricted the resolved sub-populations (Poran et al., 2017).

scRNA-seq profiling of field isolates from patients without or with minimal lab-adaptation may provide holistic insights into the disease pathogenesis and the impact of host immune system even though the logistics and technical challenges would be immense. Using *in vivo* murine malaria models for such studies may also provide complementary results. Additionally, the regulation of clonally variant gene expression (Reid, 2015), infected RBCs binding phenotypes (Goel et al. 2014; Miller et al. 2013) and drug resistance (Artemova et al. 2015) are some of the aspects of the disease that scRNA-seq may impact. For instance rosetting has been established to occur in a blood type dependent manner (Goel et al. 2014). Using single-cell methods it would be possible to at least delineate the transcriptional mechanisms underlying such a significant pathophenotype.

The application of scRNA-seq to understanding the blood stage and even the other stages is in its early phases. With improved designs mirroring the disease, enhanced scRNA-seq protocols and complimentary technologies the regulation of key genes will be unraveled and these may become potential targets for drug/vaccine development.

3.2 COMPARATIVE ANALYSES OF SINGLE-CELL RNA-SEQUENCING METHODS

In most studies aimed at characterizing sub-population(s) at cellular resolution, a specific scRNA-seq protocol is used for library generation proceeded by multiple computational steps culminating in unsupervised detection of distinct cell clusters. Presently alternatives exist for protocols, normalization, genes (features) selection, dimensionality reduction and clustering hence posing a critical challenge when designing such studies. This has invigorated the need to benchmark their performance.

scRNA-seq protocols benchmarking

Several protocols for generating scRNA-seq libraries for sequencing exist. These differ in their chemistries, implementation platforms and performance. To this end several benchmarking attempts have been carried out (Bagnoli et al. 2018; Sasagawa et al. 2013; Sheng et al. 2017; Svensson et al. 2017; Wu et al. 2014; Ziegenhain et al. 2017).

These comparisons often evaluate the protocols in terms of their number of detected genes per cell (sensitivity), cells throughput, precision (variability in gene expression measurement between replicates), accuracy (based on comparison with other expression profiling methods e.g. single-cell multiplex qPCR), cost, transcript coverage and library complexity. While end counting methods have been attractive for their ability to mitigate amplification bias by using UMIs these may underestimate the expression of endogenous genes (Svensson et al., 2017).

Small variations in the implementation of a given protocol give different performance results e.g. CEL-seq2 on Fluidigm C1 and microwell-plate results in poor sensitivity (Svensson et al. 2017). In numerous protocols newer versions with optimizations from the initial method have been established (Hashimshony et al. 2012; Hashimshony et al. 2016; Soumillon et al. 2014; Bagnoli et al. 2018). Until recently most benchmarking studies utilized cell lines or spike-in controls to compare protocols. However a new study evaluating three main commercial platforms (10X genomics, Drop-seq and inDrop) used human pancreatic islet cells hence making it possible to compare their potential to recapitulate known distinct cell constituents (X. Zhang et al. 2019).

Clustering methods comparison

Unsupervised clustering is fundamental in the characterization and/or discovery of sub-populations of cells belonging to distinct cell type or state. At the time of writing 115 tools were listed in the clustering category in the online database (Zappia et al., 2018). While several specialized algorithms have been developed to tackle the idiosyncrasies of scRNA-seq data, the use of conventional clustering methods such as K-means , hierarchical and graph-based one (Pijuan-Sala et al. 2019) has continued. It is noteworthy that some of the prevailing methods are modification of pre-existing clustering techniques for instance SIMLR (Wang et al. 2017) and RaceID (Grün et al. 2015) have optimized k-means for tackling the problem of rare cell type detection. CIDR (Lin et al. 2017) modifies hierarchical approach for single cell data.

There are several metrics used for estimating the distance between clusters including Euclidean, cosine and correlation. The latter two are scale independent making them robust to the high variation in gene expression measurements (Kiselev et al. 2019). In terms of performance evaluation it has been established that the distance metric and protocol choice influence clusters discovery (Kim et al. 2018).

Bi-clustering technique, BackSPIN, has been useful in determining sub-clusters of cells with distinct signatures of expression (Zeisel et al. 2015).

Different clustering methods have their merits and demerits (Kiselev et al., 2019). Hence methods such as SC3 have adopted a consensus-based strategy to facilitate the identification of robust groupings (Kiselev et al. 2017).

Normalization comparison

Normalization is a critical step in making sense of scRNA-seq data since it seeks to remove unwanted technical and/or biological variation from the digital expression measurements. A number of methods exist each addressing different aspect of scRNA-seq e.g. zero-inflation (L Lun et al. 2016), technical variance (Yip et al. 2017) and sequencing fluctuations between cells (Bacher et al. 2017; Wolf et al. 2018). Methods such as BASiCs (Vallejos et al. 2017) and DESEq2-sc (Brennecke et al., 2013) require exogenous spike-ins to model the technical variation. It is noteworthy that there are scRNA-seq studies that have adopted normalization schemes earlier used for bulk-RNAseq (such as TMM (Robinson & Oshlack 2010), DESeq2 (Love et al. 2014) and RPKM (Mortazavi et al. 2008).

Log-transformation of the normalized gene expression matrix is common practice and usually a pseudo value is added to preclude undefined values from non-detected genes.

Using unsuitable methods may lead to erroneous outcomes in the downstream analysis such as highly variable and differential gene expression detection (Vallejos et al., 2017) resulting into incorrect conclusions. While there are earlier studies that showed scRNAseq-tailored methods (scran and BASiCs) outperforming bulk ones in generating accurate scaling factors (Vallejos et al., 2017) an exhaustive assessment is still lacking.

Dimensionality reduction and feature selection

Processing scRNA-seq data results in high dimensional expression matrix that comprises 10s to 1000s of cells and several thousands of genes. In order to reduce this complexity while retaining the biological signal several strategies are involved. Instead of using all detected genes one may opt for differentially expressed genes between sub-groups, highly variable genes (HVGs) or key markers if the biological process is well characterized. The use of HVGs is widely adopted especially in the context of unsupervised discovery of sub-populations. To determine HVGs the frequently adopted approach entails modeling mean-CV (coefficient of variation) relationship and determining candidates with variation above the null. Varied cut-offs for top HVGs have been utilized: 100 (Posfai et al. 2017), 2000 (Wagner et al. 2018), 3000 (Cadwell et al. 2016), 5000 and even more (Giustacchini et al. 2017; Zeisel et al. 2015). The appropriate cut-off largely depends on the biological question and the variability within the data.

After feature selection the cells may be projected onto lower dimensions using specialized techniques that maybe linear or non-linear (Becht et al. 2019; Woodhouse et al. 2015). Here the common methods are: Correlation (Pearson or Spearman), PCA, tSNE (van der Maaten & Hinton 2008) and UMAP (Becht et al. 2019).

Simulation approaches

Several studies looking into the performance of scRNAseq have adopted various simulation strategies (Zappia et al. 2017; Vieth et al. 2017). It is noteworthy that these studies have been applied in the context of distinct cell states/types or were simulated using parametric models that may not reflect the complete complexity of scRNA-seq data. Furthermore in certain instances spike-in controls have been used to evaluate the impact of scRNA-seq protocols (Svensson et al. 2017) and clustering metrics on subtype discovery (Kim et al. 2018). Hence the need for accurate simulation of subtypes, unbiased comparisons and comprehensive assessment of different parameters in the context of subtle differences between sub-populations remains to be explored.

3.3 EPITHELIAL-TO-MESENCHYMAL TRANSITION

Cancer is a leading cause of death globally and approximately 90% of the mortalities are attributed to the metastatic spread of malignant tissues from the primary site to vital secondary organs (Hanahan & Weinberg 2000; Hanahan & Weinberg 2011). Compelling evidence implicates EMT (epithelial-to-mesenchymal transition) in cancer metastasis (Lawson et al. 2015; Pastushenko et al. 2018). EMT entails the switching of an epithelial cell to mesenchymal state and the reversal process is known as mesenchymal-to-epithelial transition (named MET). First described in the 80s, EMT and MET (simply abbreviated EMT-MET) have been established to be critical in gastrulation and neural crest migration during embryonic development, the formation of septa and valve in heart development and post-embryonic tissue maintenance in human and other mammalian systems (Thiery, Acloque, Huang & Nieto 2009a; Gonzalez & Medici 2014). However in disease, EMT-MET program are reactivated leading to pathogenic progression of conditions such as tissue fibrosis and malignant cancer (Thiery, Acloque, Huang & Nieto 2009b). In light of the significant health implications of EMT-MET, understanding the mechanism(s) is critical goal in cancer biology.

Molecular biology of EMT

EMT-MET are complex and transitory processes that transform the sessile carcinoma cells into invasive and migratory mesenchymal state. During the process concurrent conferment of anti-apoptic, anti-therapeutic and stem-like competences occur (Lawson et al. 2015). It is regulated at different levels namely: epigenetic, transcriptional and post-transcriptional (Lamouille et al. 2014; L. Li & W. Li 2015). These divergent regulatory mechanisms operate in an integrated fashion leading to the activation of genes that are important in inducing and maintaining EMT. A number of these genes and signaling pathways have been described based on *in vivo* and *in vitro* models.

To accomplish these remarkable changes malignant carcinomas, in response to extracellular stimuli from tumor microenvironment, transduce an intracellular signaling cascade that culminates into the activation of core EMT transcriptional factors (core EMT-TFs) (Nieto et al. 2016; De Craene & Berx 2013; Shibue & Weinberg 2017). The core EMT-TFs directly or indirectly repress the transcription of key epithelial phenotype maintenance genes and activate mesenchymal ones. The generated tumorigenic mesenchymal cells can be disseminated to secondary/distal organs via the vascular system and through the reverse MET process result in metastatic malignancy. This has been referred to as the classical view of EMT-MET. It encompasses two major states (epithelial and mesenchymal), its regulation is principally driven at the transcriptional level and its basis relies on *in vitro* experimental models.

Over the years with improved molecular phenotyping technologies and robust *in vivo* cancer models alternative views of EMT-MET are emerging. Growing evidence suggest that the process is highly plastic encompassing several intermediate and reversible states between the epithelial and mesenchymal spectrum (Pastushenko et al. 2018; Lawson et al. 2018). In terms of regulation, non-transcriptional mechanisms may play a significant role in the transitory process and different EMT programs can be utilized by malignant neoplasms depending on tumor subtypes, tissues affected, cell states or types and the microenvironment (Lamouille et al. 2014). These novel discoveries have intensified the need to define and refine the potential intermediate states during EMT, establish all the regulatory mechanisms governing EMT-MET and describe the different EMT-MET programs employed by distinct cancer types during metastasis. To this end single-cell technologies in combination with robust *in vivo* cancer models have been began generating invaluable molecular understanding of EMT-MET (Pastushenko et al. 2018; Lawson et al. 2018). These studies hold immense promise in tackling metastasis with the ultimate goal of inspiring innovative therapies for managing malignant cancer.

During EMT the strong epithelial cell-cell junctions have to be broken down and several genes encoding these junctional proteins are down regulated. This leads to cells with an enhanced migratory phenotype and higher affinity for mesechymal cells (Wheelock et al. 2008; Yilmaz & Christofori 2009a). To accomplish this E-cadherin (*CDH1*) gene, is down regulated and the loosely binding N-cadherin (*CDH2*) gene is up regulated. This switch from *CDH1* to *CDH2* is an established hallmark of EMT (Lamouille et al. 2014). Epithelial cells loose their apical-basal polarity during EMT and this is facilitated by the down-regulation of polarity complex genes and enhanced by the interruption of their interactions with the junctional proteins (Moreno-Bueno et al. 2008). Cytoskeleton modification occurs during EMT and this is facilitated by the down-regulation of cytokeratin encoding genes and up-regulation of vimentin (Huang et al. 2012). To modulate the interactions of epithelial cells with the extracellular matrix the expression of integrin encoding genes also change during EMT (Yilmaz & Christofori 2009b).

Key transcription factors (TFs) regulating EMT are the zinc-finger binding TFs (Snail1 and Snail2) and the basic-helix-loop-helix binding proteins (*Twist*, zinc finger E-box-binding homeobox 1(*ZEB1*) and *ZEB2*) (Gonzalez & Medici 2014). They operate synergistically by binding the promoter regions of genes that may activate mesenchymal state or deactivate epithelial phenotypes in EMT. For instance, Snail1 binds to the promoter region of *CDH1*, the gene encoding E-cadherin, leading to its repression (Batlle et al. 2000; Cano et al. 2000) and its accumulation in the nucleus has been associated with metastasis in breast cancer (Yook et al. 2006). There are additional TFs that work in unison with the master regulators to achieve more specific functions in EMT. A comprehensive account of the transcriptional regulators and gene expression changes in EMT have been reviewed by Lamouille et al (Lamouille et al. 2014).

Post-transcriptional regulation by pre-mRNA splicing and microRNA-mediated degradation is established in EMT. Splicing is mainly driven by the down-regulation of *ESRP1* (epithelial splicing regulatory protein) and *ESRP2* resulting in isoforms that enhance mesechymal phenotype (motility, loose adhesion, signaling pathways) (Warzecha et al. 2010). Additionally, the expression regulation of *RBFOX2* (RNA binding protein FOX1 homologue 2) and *SRSF1* (Ser-Arg-rich splicing factor 1) facilitate splicing alterations in EMT (Braeutigam et al. 2013; Valacca et al. 2010). Example genes that undergo alternative splicing include p120 catenin, CD44 (adhesion protein cluster of differentiation 44), among others.

Non-coding microRNAs (miRNAs) block the translation of specific RNA transcripts by targeting them for degradation or inhibiting their translation. Several miRNAs have been characterized as regulators of EMT with some binding EMT master TFs (Lamouille et al. 2013). For instance, miR-29b (Ru et al. 2012) and miR-30a (J. Zhang et al. 2012) repress EMT by binding Snail1. Other EMT genes that are miRNA targets include E-cadherin, N-cadherin, PAR3, among others (Lamouille et al. 2014).

For successful metastases, an epithelial tumor-initiating cell needs to undergo de-differentiation into a migratory mesechnymal cell and subsequently re-differentiation into an epithelial state. This process requires some reversible reprogramming mechanism and several epigenetic modifications have been established to drive this transient process (Tam & Weinberg 2013).

Cell signaling in EMT

Cell signaling pathways are key in EMT since they activate the expression of EMT-inducing transcription factors via a cascade of intracellular kinases(Gonzalez & Medici 2014; Lamouille et al. 2014). The EMT-inducing signaling pathways include the ones mediated by TGF-β (transforming growth factor - beta), FGF (fibroblast growth factor), EGF (epidermal growth factor), BMP (bone morphogenetic protein), Wnt, Shh (Sonic hedge hog), Notch, PDGF (platelate-derived growth factor) and integrin. These pathways may act predominantly at specific stages of EMT and in a cell type dependent manner (Lamouille et al. 2013).

TGF-β pathway, the most well characterized signaling in EMT, is activated by the TGF-β superfamily ligands with TGF-β1, BMP2 and BMP4 known to induce the process in cancer (Gonzalez & Medici 2014). It evokes EMT in a cascade of intracellular interactions that generate SMAD complexes that combine with DNA-binding TFs at regulatory elements to control transcription (Lamouille et al. 2013; Lamouille et al. 2014). For instance Snail1 expression is induced through TGFβ-mediated activation of SMAD3 and both SMAD3 and SMAD4 regulate the expression of HMGA2 (high mobility group A2) in TGFβ-induced EMT (Lamouille et al. 2013). However, it is worthy to note that there are TGFβ-induced EMT pathways that do not bind EMT-TFs directly but instead activate the expression of mesenchymal genes during EMT through active SMADs. Additionally, TGFβ induces EMT by complimentary non-SMAD signaling cascades such as RHO-like GTPases, ERK MAPK and PI3K (Lamouille et al. 2013).

The EMT-inducing signaling pathways often interact and cooperate with each other to activate or repress EMT progression. For instance TGFβ-induced EMT cause adherens destabilization leading to β-catenin accumulation in the nucleus that feed into Wnt signaling. For a comprehensive description of the key EMT signaling pathways the review by Gonzalez et al (Gonzalez & Medici 2014) is recommended.

EMT and single cell gene expression profiling

In malignancies single cell profiling can be useful in enhancing cancer biology (Tirosh, Izar, et al. 2016; Tirosh, Venteicher, et al. 2016) and addressing multiple aspects including prognosis (Dalerba et al., 2011). Tackling the transcriptional basis of EMT at cellular resolution has burgeoned over recent years. Based on PDX models of breast cancer metastasis 116 genes were profiled using FACS sorting in combination with Fluidigm's multplex qPCR (Lawson et al., 2015). These genes are involved in EMT, stemness, proliferation, cycling and lineage specification. The study established that there was distinct signature between low and high burden metastasis with the former expressing key stem cell maintenance (*Sox2, Oct4/Pouf5*), EMT programme (*Nai2*, *Skp2* and *Twist1*) and dormancy genes (*Cdkn1b, Chek1, Tgfbr3* and *Tgfb2*). Low-burden metastases were more prolific in singly seeding tumors than high-burden derived xenografts. Differentiation status correlates with metastatic burden with the low burden metastasis expressing basal/stem-like signature while high burden having a more luminal expression signature both at transcriptional and protein level (Lawson et al., 2015).

The first *in vivo* model of skin squamus cell carcinoma (SSC) based on genetically engineered mice established the existence of partial-EMT sub-populations using single-cell approach (Pastushenko et al., 2018). The model is mediated by the conditional and targeted expression of the oncogene, *KRas*, and repression of tumor suppressor gene *p53* in hair follicles.

With increasing adoption of scRNA-seq techniques in EMT-MET studies and improved models a refined molecular understanding of this complex process may be achieved in the coming decade.

3.4 MYELODYSPLATIC SYNDROME AND CANCER STEM

Myelodysplastic syndrome (MDS) is a collection of clonal hematopoietic disorders characterized by abnormal blood cells morphology in one or more lineages (dysplasia), defective hematopoiesis leading to low blood counts (cytopenia) and a predisposition to secondary acute myeloid leukemia (Sperling et al. 2017). It is the most prevalent hematopoietic malignancy with 5.3 to 13.1 cases per 100,000 individuals reported in United States. Due to the difficulty in diagnosing MDS and limited records of incidence, it is suspected that the incidence could be higher than current estimates (Özcan, Ilhan, Ozcebe, Nalcaci & Gülbas 2013a; Sperling et al. 2017). Overall, every indication from the inadequate data is that MDS incidences are increasing which has been attributed to the aging population, rise in therapy-related MDS and enhanced awareness. Currently three United States FDA approved agents for treatment (azacitidine, decitabine, and lenalidomide) but only allogeneic bone marrow transplant has curative capacity (DeZern 2018).

Molecular and genetic basis of MDS

Several chromosomal aberrations have been detected in distinct MDS subgroups (Özcan, Ilhan, Ozcebe, Nalcaci & Gülbas 2013b; Sperling et al. 2017). These include interstitial deletion of chromosome 5 long arm (del 5q-MDS), trisomy 8, monosomy 7 and 17p syndrome among others. Joint operation between initiating and cooperative mutations in hematopoietic stem cells or progenitors establishes clonal dominance that progresses to MDS and in some cases culminates into secondary AML (Sperling et al. 2017). This involves the interplay of epigenetic alteration, stepwise acquisition of cooperative mutations and bone marrow microenvironment. The initial foundational mutations generate clonal hematopoiesis which then acquire cooperative mutations that drive the progression to MDS and ultimately sAML. Some of the commonly mutated genes in MDS are categorized as epigenetic (e.g. *DNMT3A, TET2*), splicing genes (e.g. *SF3B1, SRSF2, U2AF1*) and transcription factors (e.g. *RUNX1 CUX1*).

Cancer stem cell theory and myelodysplastic syndrome

Intra-tumoral heterogeneity facilitates distinct cell types/states in malignancies and some of these sub-populations may confer fitness such as anti-therapeutic response in malignancies (Lawson et al. 2018). Cancer stem cells (CSC) concept from the 1960s posits a heterogeneous and hierarchical organization of malignant tumors in a stem cell-like configuration that mirrors the corresponding non-neoplastic tissue. In this organization the small sub-population of distinct CSCs, with self-renewal and multi-lineage differentiation capability, maintain the tumor tissue by repopulating the non-tumorigenic cancer cells through differentiation while retaining the reservoir of CSCs (Kreso & Dick 2014). This phenomenon was initially proven in the hematopoietic malignancy, Acute Myeloid Leukemia (Bonnet & Dick 1997) in mouse-models and later in solid tumors (Al-Hajj et al. 2003). The CSC were shown to have a distinct immunophenotype and long-term tumor initiating potential in *in vivo* xenografts (Jones & Armstrong 2008). Since then CSCs have been identified in various cancers including brain, lung, head and neck, and colon (Kreso & Dick 2014). In some cancers it has been proven that the CSC state is dynamic such that non-tumorigenic sub-populations may gain tumor initiating potential and vice-versa. This is mediated by genetic, non-genetic (such as epigenetic modification) and context-dependent mechanisms (Kreso & Dick 2014).

Given the clinical significance CSCs there is greater need to understand their biology in depth. For instance in del5q MDS patients a distinct sub-population of multi-potent stem cells (CD34+CD38-CD90+Lin-) was shown to confer lenalidomide resistance and had potential to reconstitute myeloid progenitors in *in vitro* long-term colony assays (Tehranchi et al. 2010; Woll et al. 2014). In AML transcriptional signature of tumorigenic sub-population was linked to patients prognosis (Eppert et al. 2011). Therefore with improved models and enhanced techniques (e.g. use of microRNA and reporter assays) the characterization and molecular validation of CSCs is set to enhance our knowledge in cancer biology.

4 AIMS

In this thesis, I used scRNA-seq to tackle key biological problems in infectious disease and cancer as well as performed computational assessment of the general possibilities and limitations of scRNA-seq for enumerating cell types and states de novo.

4.1 PAPER I

- Establish and optimize a workflow for isolating and preparing scRNA-seq libraries from *P. falciparum* intra-erythrocytic development cycle (IDC) stages.
- Analyze the heterogeneity and describe sub-populations of *P. falciparum* during IDC development at cellular resolution based on scRNA-seq.
- Identify novel and known markers involved in IDC and sexual differentiation at single cell level.

4.2 PAPER II

- Establish a computational workflow for simulating and discovering sub-populations of cells from pre-existing homogenous scRNA-seq data.
- Define the limits of subtype discovery and the impact of seven parameters in the computational workflow for sub-population discovery, and across scRNA-seq protocols.

4.3 PAPER III

- Investigate the impact of TGF-ß1 induced EMT (epithelial-to-mesenchymal transition) on lung metastasis colonization and growth.
- Monitor how TGB-ß1 induced EMT affected long-term molecular programs during metastatic cancer growth in the lung.

4.4 PAPER IV

- The generation and computational analysis of gene expression profiles of MDS stem cells and hematopoeitic progenitors.
- Determine the transcriptional signature in relation to specific genetic abberations in MDS stem cells, in comparison to hematopoietic progenitors.

5 RESULTS AND DISCUSSION

5.1 PAPER I

Malaria is an important infectious disease with the most severe form caused by *Plasmodium falciparum*. During infection there are several phenotypes of the disease that occur in a limited subset of the infected erythrocytes such as cytoadhesion, drug resistance and sexual commitment among others. These often have significant implications for the disease virulence and transmission. For instance, sexual differentiation generates a reservoir of metabolically inactive gametocytes that can be taken up by the female anopheline mosquito and further transmitted to human host. While the genetic basis and mechanisms responsible for some of these phenotypes have been unraveled an exhaustive molecular understanding is lacking partly due to shortage of high throughput and sensitive technologies that can resolve these rare phenotypes at a single cell level.

Hence in this study we applied scRNA-seq to the intra-erythrocyte development (IDC) stage of *P. falciparum* both as a proof-of-principle and to assess if it is possible to study sexual differentiation at cellular resolution. It is noteworthy that while scRNA-seq technologies had begun impacting various domains in biology in 2009 and earlier in the decade (Xue et al. 2013; Deng et al. 2014; Jaitin et al. 2014; Fan et al. 2015; Patel et al. 2014), this technology remained unexplored in malaria. In this study, we demonstrated that scRNA-seq could capture the distinct states of the parasite during IDC as evidenced by the correlation between aggregated single cell expression and bulk controls per IDC sub-stage and the expression of IDC regulated genes. The average number of genes we detected per cell in our dataset was comparable to the Reid et al dataset (Reid et al. 2018 data and higher than Poran et al (Poran et al. 2017). The Reid et al study further optimized the application of Smart-seq2 for parasite profiling and they also implemented strategies to filter rRNAs, which may explain their detection of higher numbers of genes. It is noteworthy that these two studies only profiled late trophozoite, schizont and gametocyte stages but we additionally sampled earlier timepoints. The number of detected genes per sample increased with the sampling time. This corroborates the findings that ring stages tend to have lowest transcriptional activity in comparison to the other IDC stages and mRNA half-life increases with the IDC progression (Shock et al. 2007; Sims et al. 2009). However, the drop in number of genes in T3 timepoint cells (likely early trophozoite) was unexpected**.** This pattern may reflect a sub-state in the developmental trajectory since it was observed in all independent experiments targeting this stage. While earlier ensemble-based expression profiling had revealed cascades of transcriptional changes during IDC (Bozdech et al. 2015; Le Roch et al. 2003; Otto et al.

2010) here we observe discrete expression signatures within the eight subpopulations that we established. The cascades at bulk level maybe an indication of the gradual transitions down the IDC development at unsynchronized rate by the parasites.

On the host side we detected a number of remnant human mRNA transcripts including *HBA2, HBB, HBD, SAT1 NKX3-1* and others. These are likely leftovers of initially transcribed genes during erythropoiesis before enucleation and maturation of the terminally differentiated erythrocytes.

Intriguingly, we illuminated a subset of sexually differentiated *P. falciparum* that had a defined gene signature of ten novel genes, five of which were validated to have higher expression in populations of gametocyte-enriched parasites. These 10 genes could separate sexual and asexual parasites in Reid et al data (Reid et al. 2018) and had been detected in independent datasets accessible in PlasmoDB database (Aurrecoechea et al. 2009). These candidate markers of sexual commitment may denote important genes for parasite transmission that could be exploited for the design of transmission-blocking drugs and/or vaccines. However the fact that all the ten are annotated as hypothetical proteins is limiting and is illustrative of the large numbers of P. *falciparum* genes that still lack biological functional annotation. The two candidates Pf3D7_1474000 and Pf3D7_0205100 are most intriguing since further evidence from independent transcriptome and proteome profiling experiments deposited in PlasmoDB database (Aurrecoechea et al. 2009) also shows their gametocyte specificity and critical role in the parasites survival.

Mutually exclusive expression of gene families is a hallmark of *P. falciparum* survival within host. Here we profiled the *var* and *clag* genes that are involved in cytoadhesion and erythrocyte invasion respectively. We confirm dominant expression of *clag3.1* and *clag3.2* in a subset of the SCTs with co-expression of the two genes in a few of the cells. We confirm the dominant expression of var genes: *Upsc1, Upsb1* and *Upse* in subet of cells. The latter is the pregnancy-associated *var2csa* antigen that facilitates iRBCs adhesion to the uterine membrane.

After sequencing and aligning the read fragments, we utilized multiple metrics of quality control including detection of outliers based on correlation distance between samples and PCA and tSNE projections, mapping rates of reads and the qualities of base calls from sequencing. Most informative was to require at least 10,000 uniquely mapped reads per cell and the detection of 200 *P.falciparum* genes detected per cell. Since mature erythrocytes are transcriptionally inactive, I reasoned that higher proportion of the reads must map towards the parasite reference genome in comparison to the host. In line with this an average 66.92% of the total reads mapped to *P.falciparum* and 33.08% to human genome in the scRNA-seq libraries. One of the emerging techniques in scRNA-seq field is integration of datasets from different batches or methods (Stuart & Satija 2019; Butler et al. 2018; Haghverdi et al. 2018). I tried to align our gene expression alongside the two other published single cell data (Poran et al. 2017; Reid et al. 2018) using Canonical Correlation Analysis algorithm as implemented in Seurat (Butler et al. 2018) but the expression data remained clustered by the lab of origin. This is attributable to both biological and technical variations within the data.

While FACS sorting or droplet microfluidics have become the standard approaches for high throughput isolation of single cells here we used an automated CellSorter (Környei et al. 2013) to pick the mitotracker stained infected erythrocytes. The merit with this approach is it afforded us the ability to discriminate non-singly infected RBCs and the minimal pressure was useful for preventing the rapture of the delicate RBCs. Great care was observed to ensure the parasites were isolated within 30 minutes of removal from culturing conditions to ensure RNA integrity.

Overall this study provides an initial demonstration that scRNA-seq can be useful in understanding the molecular basis of the disease at a cellular level. It opens up the possibility of investigating several rare phenotypes of malaria such as sexual commitment, cytoadhesion and clonal variation. Possibilities of studying *Plasmodium* species (such as *P.vivax*) that are uncultivable in the lab could be realized with further optimization.

5.2 PAPER II

Even though scRNA-seq contributes immensely to the enumeration of distinct cell types, subtypes and states in mammalian cells, studies assessing the impact of different parameters in the subpopulation discovery workflow are lacking. In this study, we set out to define the limits of uncovering such subgroups using a simulation strategy that enabled tracking the magnitude of difference between two subpopulations while retaining the intrinsic properties of scRNA-seq data. Here, we used existing published single cell expression data from mouse embryonic stem cells (Ziegenhain et al. 2017). While studies have simulated scRNA-seq using parametric models (L Lun et al. 2016; Zappia et al. 2017; Kharchenko et al. 2014; Vieth et al. 2017) for the goal of appropriate differential expression analysis, testing algorithms performance and modeling technical noise, these studies assumed models that may both not capture the full complexity and dependencies in scRNA-seq data, and they might also be unequally suited for different kinds of scRNA-seq methods. For instance, my attempts at using Smart-seq2 data from mESCs to infer parameters then simulate scRNA-seq data using the Gamma-poisson model in splat (Zappia et al. 2017) or the mean-variance model with local polynomial regression (Vieth et al. 2017) both resulted in marked reduction in the complexity of the simulated data in comparison to the real data. Hence, we reasoned that by partitioning cells into two groups and perturbing a subset of genes in one of them by matching them to the expression of highly or lowly expressed genes, we not only establish the subpopulations from a homogenous group but also retain the complexity and other properties such as dropout rates and variability. By adopting this approach, we did not have to assume any distribution models, and our strategy would be equally suitable to all kinds of scRNA-seq data. We also avoided the comparison of datasets with mixed batches where the effects were evident since this is still an active field of research. While batch correction techniques have been proposed and used in different scRNA-seq analysis, the robustness of these algorithms is not clear. For instance, using regression methods as implemented in Seurat (Butler et al. 2018), SVA-based techniques (Jaffe et al. 2012) and Bayesian approach as implemented in ComBat (Johnson et al. 2006) resulted in some negative values in the normalized gene expression matrix which would be challenging to process downstream. The downside of using single batches in the analysis is the limited number of cells in all the methods except Smart-seq2.

We evaluated the impact of seven different parameters in subpopulation discovery workflow namely: protocols, dimensionality reduction, normalization techniques, top variable genes cut-offs, clustering methods and total number of cells. Our analysis included six common scRNA-seq protocols, three key dimensionality reduction techniques and several normalization strategies. In our approach we adopted the popular approach of using of top variable genes to discover subpopulations of cells. We confirm that sensitive scRNA-seq protocols (such as SCRBseq, CEL-seq2 and Smart-seq2) outperform non-sensitive ones (Drop-seq, MARS-seq and Smart-seq). This was especially true when low- to intermediate-expressed genes were perturbed. Intriguingly, the full transcript coverage techniques (Smart-seq and Smart-seq2) generated lower cluster scores than the other four protocols when highly expressed genes were perturbed. We speculated that this trend might be driven by a change in the gene variability as quantified by coefficient of variance (CV). However, this was not apparent from comparing the protocols data. Each protocol's expression data had ERCC spike-ins but I opted not to use them in identifying the most variable genes. This decision was informed by the fact that synthetic spike-ins have several drawbacks and may not be reflective of the endogenous genes properties (Svensson et al. 2017). Their properties could be non-reflective of the endogenous transcripts (e.g. shorter polyA tails in ERCC's spike-ins), cannot track any variations that occur before the reverse transcription step and are usually technically challenging to calibrate especially for small cells with minute amounts of RNA.

Manifold-based visualization techniques in combination with Principal Component Analysis (PCA) generated the most sensitive separation of sub-populations. Interestingly, applying manifold methods directly to the top variable genes was poor at delineating the simulated subpopulations and this may demonstrate their limitation when only marginal differences exist between the subpopulations.

For normalization, single cell tailored normalization methods (mainly linnorm) outperformed bulk-RNA-seq approaches (such as RPKM, TMM), particularly when the magnitude of difference between sub-populations was small. Reassuringly, when the perturbation magnitude was large all cluster scores were high independent of the normalization approach. This shows that the use of normalization methods might only improve sub-type discovery within a window where subtle but distinct biological variation between types are present.

The number of perturbed genes required to delineate sub-populations varied based on their level of expression, protocol and magnitude of modification. Interestingly, we discovered that for lowly to intermediately expressed genes high level of perturbation was needed (in terms of numbers of genes or the fold expressions alterations) for interpretable sub-clusters to emerge. This may illustrate the challenge in delineating such groups of cells using the current scRNA-seq workflows.

In terms of the number of most variable genes to use for the discovering the subpopulations 500 provided the right balance between sensitivity and specificity in capturing the perturbed genes in the most variable genes at different degrees of perturbations.

Overall this analysis reveals the substantial differences between subpopulations that would escape detection, irrespective of the scRNA-seq protocol and the normalization technique. We also describe how other common parameters in the unsupervised workflow affect the subgroups discovery.

5.3 PAPER III

Metastatic spread from primary to secondary distal organs and subsequent metastatic tumor growth is responsible for more than 90% of cancer mortalities. Compelling evidence implicate epithelial-to-mesenchymal transition (EMT) as a driving mechanism of metastasis (Ye et al. 2015; Dongre & Weinberg 2019). In this study, we collaborated with Jonas Fuxe's lab to establish an *in vivo* EMT model in mice to investigate the molecular programs occurring during long-term metastatic growth and role of EMT. We demonstrate that Transforming Growth Factor beta I (TGF-ßI) is a potent inducer of mesechymal phenotypes in oncogenically primed epithelial cells (EpRas cells) and that autocrinal activity of this cytokine stimulates higher metastatic potential than paracrinal mode of action in epithelial cells. In our experimental setup, EpXT and EpRas-TGF-ßI may mirror the TGF- autocrinal and paracrinal mode of action. Using BALB/c mice with an intact immune system allowed for the immune interaction in cancer unlike majority of studies that use immune deficient murine models.

We observe an increased adhesion and metastatic burden in murine lungs and poor survival curves for TGF-ß1 treated cells with tumor-seeding potential. While the three distinct oncogenic cell lines corresponding to autocrinal TGF-ß1 (EpXT), paracrinal TGF-ß1(EpRas-TGF-ß1) and none TGF-ß1 treated (EpRas) have distinguishing transcriptional profiles *in vitro*, this is lost *in vivo* with the intermixing in PCA and tSNE subspace. Importantly, all cell lines established a mesenchymal characteristic irrespectively of their EMT status before injection into mice, and even after 8 days of metastatic growth they had maintained this phenotype. Although a few cells, mostly of EpRas origin, had more epithelial characters in the metastatic organ, we cannot determine whether such cells were colonizing and growing in the lung together with the cells that had undergone EMT, or whether these cells transited back from a MET (mesenchymal to epithelial transition) in the lung. For such resolution, we would need to simultaneously adopt lineage-tracing strategies.

5.4 PAPER IV

The cancer stem cell hypothesis postulates that malignant cells are hierarchically organized in a tissue-like fashion. At the top, a small subpopulation of cancer stem cells (CSCs) propagates through self-renewal and differentiation into non-tumorigenic cancer cells that form the tumor mass. While the existence of CSCs in hematological malignancies and some solid tumors had been established *in vitro* and in murine models up until this study, their existence in human malignancies had been debated. Our main collaborators at Oxford isolated and then molecularly and functionally characterized CSCs in myelodysplastic syndrome (MDS-SCs) patients with low- and intermediate- risk MDS. Using our in-house computational pipelines and the Smart-seq protocol (Ramskold et al. 2012), we profiled an ensemble of the MDS-SCs and progenitors compartments. PCA analysis established that the MDS-stem and progenitors distinctly separated on the first and second components with over 50% of the variance captured. This provided further evidence that MDS-SC and the progenitors had unique molecular phenotype even at the transcriptome level.

6 CONCLUSIONS AND FUTURE PERSPECTIVES

The maturity of scRNA-seq and the simultaneous analytical computational methods has transformed the field from a proof of concept endeavor to the pursuit of mechanistic basis of intriguing biological questions. In this thesis I have presented our efforts in using the technology to deepen the understanding of parasite biology and EMT-driven cancer metastasis. In spite of this remarkable progress fascinating but equally challenging biological and technical research questions remain to be addressed.

With the current goal of eliminating malaria (Winzeler et al. 2016) advanced molecular understanding of some of the rare but clinically important phenotypes during infection would be significant. These include: iRBCs cytoadhesive processes (such as rosetting, sequestration), sexual differentiation, non-symptomatic infections and drug resistance, among others. The recent studies (Poran et al. 2017; Reid et al. 2018) and our paper I marked the application of scRNA-seq towards IDC and sexual commitment. This has provided significant evidence on the viability of this technology to enhance the genome-wide understanding malaria disease. However there are still key questions. Recent finding that one of the earliest markers of commitment *pfGEXP5* can occur independently of *PfAp2-G* activation, the hitherto presumed master regulator of sexual differentiation, implies the existence of alternative mechanism(s) for gametocyte generation (Henry et al. 2019). This opens up potential areas of exploration e.g. when exactly does commitment occur? How does it vary across different isolates both lab-adapted and from the field? What are determinants of the gametocyte sex? What are the genetic and molecular drivers of drug response differences between male and female gametocytes? Which are the key determinants of gametocytes survival within the mammalian?

Cytoadhesion of iRBCs to the vasculature and rosetting are mediated by the clonally variant surface proteins and confer virulence to the parasite (Wahlgren et al. 2017). However exhaustive understanding of the mechanisms driving these critical phenotypes is unknown. For instance it has been shown that rosetting, varies between parasite strains and human blood types (Wahlgren et al. 2017) but the underlying mechanism is yet to be untangled. In the future it will be important to thoroughly describe how these phenotypes are transcriptionally regulated in both lab-adapted and field isolates of the parasite. It will be important to elucidate the switching mechanism of these surface variants between clones, which is an essential strategy adopted by the parasite to evade the immune system. With the recent founding of combinatorial index techniques such as sci-RNAseq (Cao et al. 2017), which do not need specialized equipment, real time analysis of malaria patients samples in some of the resource limited settings may lead to improved understanding of the disease. However these combinatorial indexing methods are just starting to generate results in mammalian and model organisms and remain unexplored in infectious diseases hence further optimization will be needed. With sequencing technologies becoming cheaper and platforms such as Oxford Nanopore ONT emerging then single cell technologies may further expand to tackle malaria in the field. The prospects of long reads sequencing to enable the generation of splicing isoforms and other allelic variants may facilitate a comprehensive understanding of the important clonally driven phenotypes.

Cells spatial context is important in defining its function and form (Crosetto et al. 2015; Yosef & Regev 2016). For instance in cancer, tumor microenvironment can have an impact on the metastatic potential of tumorigenic cells (Shibue & Weinberg 2017). The invasive cells on the tumor edges may be influenced by a distict microenvironment from the core malignancy cells due to the infiltrating exogenous cells. However, current high throughput scRNA-seq protocols require the dissociation of cells from primary tissue leading to loss of spatial dimension. There have been attempts to integrate spatial cues from a few markers to define the cells contexts using algorithms such as Seurat and Achim et al (Satija et al. 2015). However these approaches cannot be used in novel cases where cells are not clearly demarcated, are without known markers and in stances where there is spatial mix in sub-populations e.g. cancerous tissue. The emergence of elegant techniques such as MERFISH (Chen et al. 2015; Moffitt et al. 2018) will facilitate spatially resolved gene expression profiling and allow for the contextual interpretation of cellular processes. For instance in cancers way begin to understand how the different subpopulations (CSCs, stromal cells, infiltrating immune) respond to the varying microenvironment? Do the leading metastatic cells differ from the core cancer cells? How do the different cancers differ within tissues?

The emergence of multi-omics techniques combining various entities in the gene expression pathway will facilitate a truly integrated examination of cells. Already combining scRNA-seq and epigenomics techniques (sc-ATAC-seq, sc-CHIP-seq, chromosome conformation) are promising to unravel the epigenetic mechanisms controlling key biological processes. The combination of proteomic (mass spectrometry and CITE-seq) and transcriptomics technologies is also emerging. Though in its nascent years integrative technologies promise deliver an augmented and comprehensive view of biological processes.

Co-morbidities are commonplace in populations and these can involve disparate diseases. For instance cancer and malaria even though been studied separately these two diseases have concealed interactions (Nordor et al., 2018). With the rise of cancer incidences in malaria-endemic regions an integrative approach to understanding the involved diseases would be revolutionary. This may become feasible with the emerging sensitive and integrative technologies. In certain instances understanding the potential connections might motivate novel approaches for treatment, diagnosis and management. For instance a proof of concept that *Var2CSA* can be used to isolate/enrich for CTCs has been demonstrated (Agerbæk et al. 2018).

With technological breakthroughs in the offing, tackling some of the most fascinating problems will continue impacting life sciences.

7 ACKNOWLEDGEMENTS

My deepest gratitude go to my main supervisor **Rickard Sandberg** for granting me this magnificent and privileged opportunity to pursue doctoral studies in the lab. You have been an excellent scientific mentor, leader, friend and the most supportive group leader. Thank you for your leadership, always keeping your door open, providing me with great learning opportunities, openness to new ideas and consistent drive for scientific pursuits.

Thank you very much **Björn Andersson** for being my co-supervisor. You have been a great guide, mentor and always available when I had questions on *P. falciparum* biology and doctoral studies.

Big appreciation to my opponent, **Prof. Bart Deplancke**, each one of my examination committee members namely: **Dr. Goncalo Castelo-Branco**, **Dr. Ola Larsson** and **Prof. Staffan Svärd** and my defence chair **Prof. Qiaolin Deng**. Thank you all for making time for the entire evaluation process.

To my department at CMB, it has been an engaging and exciting learning experience with the entire community. I appreciated the unconditional and professional support I received over the years. Special mention to: **Matti Nikkola**, **Lina Pettersson**, **Linda Lindell** and **Margaret Ulander**. I could always count on your counsel and guidance whenever I needed it.

To all my collaborators and colleagues in the single cell *P. falciparum* study. **Prof. Mats Walhgren** for the invaluable inputs, resources, ideas and discussion on possible areas of focus in tackling this significant parasite. To the 'malaria crew': **Mia Palmkvist**, **Sven Sagasser**, **Daisy Hjelmqvist** and **Johan Ankarklev**, I immensely enjoyed working with each of you in the project. Your resilience, passion for malaria research and friendship were energizing learning experiences.

To my main collaborator in the myelodysplastic syndromes (MDS) study, **Petter Woll,** thank you very much for the excellent opportunity to work on this exciting and important health problem. I admired your passion, drive for getting work done and clarity in explaining both the MDS biology and techniques. Wishing you the best as new group leader.

To all my colleagues both former and current in the **Sandberg lab** you have been my scientific family, an inspiration, most passionate and dedicated team I have had the privilege to work with. You made my PhD journey the most stimulating and rewarding experience!

Daniel Ramskold for your brilliance, willingness to share knowledge and the humility. I enjoyed the occasional jogs that we had at Hagaparken and hope you'll keep on as you pursue science. **Ilgar Abdullayev**, you were brilliant, wonderful and encouraging soul during your days in the lab. I still cherish the scientific and non-scientific discussions we had over fika (or at Friskis gym). Hope you are having fun in industry and still keeping fit. **Omid Faridani** it was a wonderful privilege knowing you both for your scientific insights and friendship. Your passion for research, constant emphasis on proper study design and teaching talent were rewarding learning experiences. And of course the enthralling kayaking event in Stockholm is still etched in memory. Best of luck as you start your research in Australia. To **Helena Storvall** for sharing your in-depth knowledge in python programming, willingness to assist when I had coding issues and friendship. I enjoyed the scientific and social discussions we had with the rest of the Ludwig crew on the rooftop. **Daniel Edsgard** thank so much for the guidance and teaching on statistical modeling, quantitative computational biology concepts and friendship. I learned a great deal on apt coding, code documentation and that with adept understanding of R language one can achieve so much. **Hussein Talukdar** thank you so much for your friendship and the great interactions we've had since you joined the lab. I learned so much from your inquisitive scientific nature and encouragement on how to formulate my cell type study. Keep up the passion for systems biology and I am certain you will achieve your future career goals. Best wishes. **Emma Inns** since joining the lab you have provided us with the unconditional and professional support. I have enjoyed your wonderful sense of humor and engaging conversations on topics ranging from science to politics. Thank you very much for the assistance you've offered while I was preparing for the defence. To **Ping Chen** you have been a wonderful colleague and friend since you joined the lab. Thank you for sharing Python coding tips, bioinformatics knowledge and providing useful scientific critique during our discussions. I wish you the best in your post-doc and the rest of your research career. **Leonard Hartmanis** it has been great pleasure having you in the group and I have enjoyed the engaging and humorous interactions we have had in the lab. Your drive, openness to learning and passion for science, sports and life in general are admirable. I wish you the best in the exciting and challenging human transcriptional regulation study. Lots of success in your travels down the PhD road. **Anton Larsson** it has been fantastic having you in the lab and thank you for the invaluable scientific inputs and friendship. Your brilliance and adeptness in quantitative, molecular biology and computing skills will take you places in science. Best of luck in your PhD studies and scientific career. **Gosta Winberg** it has been rewarding to have you in the lab for your friendship, humor and admirable dedication to science. Thank you for teaching me some of the cool lab-bench skills and always coming by to check on us just to find out how we were doing. I wish you and Liudmila the best as you transition from science. **Gert-Jan** it has been a wonderful experience having you in the lab for the stimulating scientific discussions, remarkable input, invaluable critiquing of ideas and wonderful friendship. I appreciated your openness to sharing your knowledge. Keep up your drive and passion for basic research and I wish you the best in your career. **Michael Hagemann** you've been a great and interactive colleague and thank you so much for your willingness to collaborate and optimize the single cell malaria study. I have learned and enjoyed working with you and hopeful we'll get some initial datasets to move to the next level. Keep up the passion and drive for excellence in your experiments and all the best as you hone your computational biology skills. Best wishes in the post-doc and future career. **Per Johnsson** you have been a sociable and insightful colleague in the lab. Thank you for your beneficial scientific input during group meetings and encouragement to publish as soon as I could. Keep up the hard work and best wishes in your scientific career. **Christoph Ziegenhain** you have been a great addition to the lab and I have enjoyed the engaging scientific interactions we had. Your in-depth knowledge in molecular biology, computational biology and lab-bench experience coupled with your collaborative talent were inspiring. I wish you and Leo the best as you establish high throughput Smart-seq2 and best of luck in your career. **Oscar Forsman** for the brief period we interacted I enjoyed your passion for learning and mathematical acumen. Best wishes in your Medical degree. **Lisa Anna** welcome to the lab and it has been wonderful interacting with you for the brief duration you have been around. I wish you best of luck in your post-doc. **Asa Bjorkland** it was great pleasure working and getting inputs from you in the single cell malaria project. You helped me familiarize with the in-house computational methods and perl programming at the beginning of my PhD. Thank you and best of luck. **Asa Segerstolpe** it was a wonderful pleasure interacting and working with you especially in the EMT project. I appreciated your collaborative and consultative approach to research and the encouragement. Best wishes in your career and hope you are having fun in Harvard. **Sophie Petropoulos** I really enjoyed your camaraderie, motivation and the engaging interactions on science and social aspects. Thank you and the University of Toronto/KI crew for the fun-learning experience at the Developmental biology course in Toronto. It has been inspiring to see you transition from Post-doc to a group leader. I wish you the best as you establish your research group. **Qiaolin Deng** it was always wonderful having you as colleague during your Post-doc days and to see you successfully transition to a group leader has been inspiring. Your tenacity for achieving scientific goals and proactive nature to solving both scientific and non-scientific tasks is admirable. I wish you continued success in your research and academic pursuits. **Björn**

Reinius it was a great pleasure having you in the lab during your Post-doc years and congratulations for becoming group leader. Your drive and insightful ideas in genetics were always stimulating and I look forward to reading more exciting findings from your group. Best wishes in your career.

To the **Ludwig Institute colleagues** from all the six different groups I deeply enjoyed the camaraderie, the broad and stimulating scientific discussions over Friday fikas, lucia and the many memorable paper celebrations. **Thomas Perlman** for the excellent leadership and establishing a tight knit family at the institute. **Charlotta Linderholm** for your unrelenting support and clear guidance that helped me settle smoothly at the institute. To the PIs **Schlisio**, **Muhr** and **Johan** for the occasional inspiration and scientific inputs. **Eliza Joodmardi, Jorge Villarroel** and **Soheilla Rezaian** for your unconditional and efficient assistance in ordering lab reagents and being available to answer my questions with a lot enthusiasm. I miss the loud lunch time laughs!! **Yu Pei, Han-Pin Pui, Shangli Cheng, Shuijie, Geng Chen, Nigel Lee, Danny Topcic, Daniel Hagey, Maria Bergsland, Cécile Zaouter, Susanne Klum, Stuart, Andre Nobre, Linda Gillberg, Bhumica Singla, Hilda, Shuijie, Nick Volakakis, Konstantinos, Idha, Katarina and Isabelle Westerlund.**

To my family and friends you have been the fuel over the entire duration of my studies. **Cecillia Oman, Johanna** and **Nelson, Pekka Kohonen, Stephen Ochaya, Sheila, Linet** and **Axel, Vivian, Wawuda and Pat, Roba** and **Maureen, Novel and Joanne, Bryo, Ivy, Mr** and **Mrs Ojuki, dad** and **mum in-law, entire Ngara, Maguke** and **Mtakai family.**

To my beloved siblings **Eddy, Sabina, Ouma** and **Awendo** you have always been the best! To my **loving mum** always a firm foundation, an inspiration and our life coach. To **dad** miss you more especially at these moments!

To my loving wife and best friend **Purity** thank you for your unconditional support, counsel and motivation through the PhD journey and in life. Nakupenda sana!

8 REFERENCES

Adams, M.D. et al., 2000. The Genome Sequence of &lt;em&gt;Drosophila melanogaster&lt;/em&gt. *Science*, 287(5461), p.2185.

Agerbæk, M.Ø. et al., 2018. The VAR2CSA malaria protein efficiently retrieves circulating tumor cells in an EpCAM-independent manner. *Nature communications*, 9(1), p.3279.

Al-Hajj, M. et al., 2003. Prospective identification of tumorigenic breast cancer cells. *Proceedings of the National Academy of Sciences*, 100(7), pp.3983–3988.

Alberts, B. et al., 2014. *Molecular Biology of the Cell* 6 ed., Garland Scientific, Taylor & Francis, New York.

Andrews, S., 2016. FastQC A Quality Control tool for High Throughput Sequence Data. *www.bioinformatics.babraham.ac.uk*. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed August 3, 2016].

Anon, 2009. Genome-wide Analysis of Heterochromatin Associates Clonally Variant Gene Regulation with Perinuclear Repressive Centers in Malaria Parasites. *Cell Host & Microbe*, 5(2), pp.179–190.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814), pp.796–815.

Artemova, T. et al., 2015. *Isolated cell behavior drives the evolution of antibiotic resistance*,

Ashley, E.A., Pyae Phyo, A. & Woodrow, C.J., 2018. Malaria. *The Lancet*, 391(10130), pp.1608–1621.

Aurrecoechea, C. et al., 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research*, 37(suppl 1), pp.D539–D543.

Bacher, R. et al., 2017. SCnorm: robust normalization of single-cell RNA-seq data. *Nat Meth*, 14(6), pp.584–586.

Bagnoli, J.W. et al., 2018. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nature communications*, 9(1), p.2937.

Balaji, S. et al., 2005. *Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains*,

Baruzzo, G. et al., 2017. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature methods*, 14(2), pp.135–139.

Batlle, E. et al., 2000. The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. 2(2), pp.84–89.

Becht, E. et al., 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 37(1), pp.38–44.

Bengtsson, M. et al., 2008. Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. *BMC Molecular Biology*, 9(1), p.63.

Bonnet, D. & Dick, J.E., 1997. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nature medicine*, 3(7), pp.730–737.

Bousema, T. et al., 2014. Asymptomatic malaria infections: detectability, transmissibility and public health relevance. *Nature reviews. Microbiology*, 12(12), pp.833–840.

Bozdech, Z. et al., 2003. The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum. *PLoS biology*, 1(1), p.E5.

Bozdech, Z., Ferreira, P.E. & Mok, S., 2015. A crucial piece in the puzzle of the artemisinin resistance mechanism in Plasmodium falciparum. *Trends in parasitology*.

Braeutigam, C. et al., 2013. The RNA-binding protein Rbfox2: an essential regulator of EMT-driven alternative splicing and a mediator of cellular invasion. 33(9), pp.1082–1092.

Bray, N.L. et al., 2016. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), pp.525–527.

Brennecke, P. et al., 2013. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, 10(11), pp.1093–1095.

Buettner, F. et al., 2014. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotech*, 33(2), pp.155–160.

Bunnik, E.M. et al., 2019. Comparative 3D genome organization in apicomplexan parasites. *Proc Natl Acad Sci U S A*, 116(8), pp.3183–3192.

Butler, A. et al., 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5), pp.411–420.

Büttner, M. et al., 2019. A test metric for assessing single-cell RNA-seq batch correction. *Nature methods*, 16(1), pp.43–49.

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

Cadwell, C.R. et al., 2016. Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature biotechnology*, 34(2), pp.199–203.

Cano, A. et al., 2000. The transcription factor Snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. 2(2), pp.76–83.

Cao, J. et al., 2017. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352), pp.661–667.

Cao, J. et al., 2019. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 34, p.1.

Chaisson, M.J. & Tesler, G., 2012. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, 13(1), p.238.

Chalfie, M. et al., 1994. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148), p.802.

Chen, K.H. et al., 2015. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), pp.aaa6090–aaa6090.

Chen, T. & Dent, S.Y.R., 2014. Chromatin modifiers and remodellers: regulators of cellular differentiation. 15(2), pp.93–106.

Chen, X., Teichmann, S.A. & Meyer, K.B., 2018. From Tissues to Cell Types and Back: Single-Cell Gene Expression Analysis of Tissue Architecture. *Annu. Rev. Biomed. Data Sci.*, 1(1), pp.29–51.

Consortium, T.E.R.C. et al., 2005. The External RNA Controls Consortium: a progress report. *Nature methods*, 2, pp.731 EP –.

Coulon, A. et al., 2013. Eukaryotic transcriptional dynamics: from single molecules to cell populations. 14(8), pp.572–584.

Cowman, A.F. et al., 2016. Malaria: Biology and Disease. *Cell*, 167(3), pp.610–624.

Cowman, A.F. et al., 2017. The Molecular Basis of Erythrocyte Invasion by Malaria Parasites. *Cell Host & Microbe*, 22(2), pp.232–245.

Crosetto, N., Bienko, M. & van Oudenaarden, A., 2015. Spatially resolved transcriptomics and beyond. 16(1), pp.57–66.

Davis, M.P.A. et al., 2013. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods (San Diego, Calif.)*, 63(1), pp.41–49.

De Craene, B. & Berx, G., 2013. Regulatory networks defining EMT during cancer initiation and progression. *Nature Reviews Cancer*, 13(2), pp.97–110.

Delves, M. et al., 2012. The activities of current antimalarial drugs on the life cycle stages of Plasmodium: a comparative study with human and rodent parasites. J. G. Beeson, ed. *PLoS medicine*, 9(2), p.e1001169.

Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R., 2014a. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167), pp.193–196.

Deng, Q., Ramskold, D., Reinius, B. & Sandberg, R., 2014b. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, 343(6167), pp.193–196.

DeZern, A.E., 2018. Treatments targeting MDS genetics: a fool's errand? *Hematology. American Society of Hematology. Education Program*, 2018(1), pp.277–285.

Dobin, A. et al., 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), pp.15–21.

Dongre, A. & Weinberg, R.A., 2019. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature Reviews Molecular Cell Biology*, 20(2), pp.69–84.

Durruthy-Durruthy, R. et al., 2014. Reconstruction of the Mouse Otocyst and Early

Neuroblast Lineage at Single-Cell Resolution. *Cell*, 157(4), pp.964–978.

Eisenstein, M., 2019. Illumina swallows PacBio in long shot for market domination. *Nat Biotech*, 37(1), pp.3–4.

Eppert, K. et al., 2011. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nature medicine*, 17(9), pp.1086–1093.

Fan, H.C., Fu, G.K. & Fodor, S.P.A., 2015. Combinatorial labeling of single cells for gene expression cytometry. *Science*, 347(6222), pp.1258367–1258367.

Feachem, R. & Sabot, O., 2008. A new global malaria eradication strategy. *The Lancet*, 371(9624), pp.1633–1635.

Femino, A.M. et al., 1998. Visualization of Single RNA Transcripts in Situ. *Science*, 280(5363), p.585.

Fleischmann, R.D. et al., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223), p.496.

Flusberg, B.A. et al., 2010. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature methods*, 7(6), pp.461–465.

Freitas-Junior, L.H. et al., 2005. *Telomeric Heterochromatin Propagation and Histone Acetylation Control Mutually Exclusive Expression of Antigenic Variation Genes in Malaria Parasites*,

Gamaarachchi, H., Parameswaran, S. & Smith, M.A., 2019. Featherweight long read alignment using partitioned reference indexes. 9(1), p.4318.

Garber, M. et al., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth*, 8(6), pp.469–477.

Gardner, M.J. et al., 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature*, 419(6906), pp.498–511.

Gaublomme, J.T. et al., 2015. Single-Cell Genomics Unveils Critical Regulators of Th17 Cell Pathogenicity. *Cell*, 163(6), pp.1400–1412.

Gierahn, T.M. et al., 2017. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nature methods*, 14(4), pp.395–398.

Giustacchini, A. et al., 2017. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nature medicine*, 23(6), pp.692–702.

Goel, S. et al., 2014. RIFINs are adhesins implicated in severe Plasmodium falciparum malaria. *Nature medicine*, 21(4), pp.314–317.

Gonzalez, D.M. & Medici, D., 2014. Signaling mechanisms of the epithelial-mesenchymal transition. *Science Signaling*, 7(344), pp.re8–re8.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17(6), pp.333–351.

Grün, D., Kester, L. & van Oudenaarden, A., 2014. Validation of noise models for single-cell transcriptomics. *Nature methods*, 11(6), pp.637–640.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A., 2015a. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568), pp.251–255.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H. & van Oudenaarden, A., 2015b. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*.

Guizetti, J. & Scherf, A., 2013. *Silence, activate, poise and switch! Mechanisms of antigenic variation in Plasmodium falciparum*,

Haberle, V. & Stark, A., 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, 13, p.1.

Haghverdi, L. et al., 2018. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5), pp.421–427.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of Cancer: The Next Generation. *Cell*, 144(5), pp.646–674.

Hanahan, D. & Weinberg, R.A., 2000. The Hallmarks of Cancer. *Cell*, 100(1), pp.57–70.

Hashimshony, T. et al., 2016. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, 17(1), p.892.

Hashimshony, T. et al., 2012. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3), pp.666–673.

Henry, N.B. et al., 2019. Biology of Plasmodium falciparum gametocyte sex ratio and implications in malaria parasite transmission. *Malar J*, 18(1), p.70.

Hochgerner, H. et al., 2017. STRT-seq-2i: dual-index 5' single cell and nucleus RNA-seq on an addressable microwell array., 7(1), p.16327.

Hocine, S., Singer, R.H. & Grünwald, D., 2010. RNA Processing and Export. *Cold Spring Harbor Perspectives in Biology*, 2(12), pp.a000752–a000752.

Holoch, D. & Moazed, D., 2015. RNA-mediated epigenetic regulation of gene expression. 16(2), pp.71–84.

Horrocks, P. et al., 2009. Control of gene expression in Plasmodium falciparum – Ten years on. *Molecular and Biochemical Parasitology*, 164(1), pp.9–25.

Huang, R.Y.-J., Guilford, P. & Thiery, J.P., 2012. Early events in cell adhesion and polarity during epithelial-mesenchymal transition. *Journal of Cell Science*, 125(19), pp.4417–4422.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), pp.931–945.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Islam, S. et al., 2011. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7), pp.1160–1167.

Islam, S. et al., 2014. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(2), pp.163–166.

Jaffe, A.E. et al., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, 28(6), pp.882–883.

Jaitin, D.A. et al., 2014. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172), pp.776–779.

Johnson, W.E., Li, C. & Rabinovic, A., 2006. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1), pp.118–127.

Jones, R.J. & Armstrong, S.A., 2008. Cancer Stem Cells in Hematopoietic Malignancies. *Hematopoietic Stem Cell Transplantation 2008 Education Supplement*, 14(1), pp.12–16.

Kamme, F. et al., 2003. Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J. Neurosci.*, 23(9), pp.3607–3615.

Kang, H.M. et al., 2018. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology*, 36(1), pp.89–94.

Kantele, A. & Jokiranta, T.S., 2011. *Review of Cases With the Emerging Fifth Human Malaria Parasite, Plasmodium knowlesi*,

Kharchenko, P.V., Silberstein, L. & Scadden, D.T., 2014a. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7), pp.740–742.

Kharchenko, P.V., Silberstein, L. & Scadden, D.T., 2014b. Bayesian approach to single-cell differential expression analysis. *Nature methods*, 11(7), pp.740–742.

Kim, D., Ben Langmead & Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4), pp.357–360.

Kim, T. et al., 2018. Impact of similarity metrics on single-cell RNA-seq data clustering. *Briefings in Bioinformatics*, 343, pp.776–bby076 VL – IS –.

Kirkman, L.A. & Deitsch, K.W., 2012. *Antigenic variation and the generation of diversity in malaria parasites*,

Kiselev, V.Y. et al., 2017. SC3: consensus clustering of single-cell RNA-seq data. *Nature methods*, 14(5), pp.483–486.

Kiselev, V.Y., Andrews, T.S. & Hemberg, M., 2019. Challenges in unsupervised clustering of single-cell RNA-seq data. 6, p.1.

Klein, A.M. et al., 2015. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5), pp.1187–1201.

Klemm, S.L., Shipony, Z. & Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. 5, p.1.

Kolodziejczyk, A.A. et al., 2015. The Technology and Biology of Single-Cell RNA Sequencing. *Molecular cell*, 58(4), pp.610–620.

Környei, Z. et al., 2013. Cell sorting in a Petri dish controlled by computer vision. 3, pp.1088 EP –.

Kreso, A. & Dick, J.E., 2014. Evolution of the Cancer Stem Cell Model. *Cell Stem Cell*, 14(3), pp.275–291.

Kurimoto, K. et al., 2006. An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic acids research*, 34(5), pp.e42–e42.

L Lun, A.T., Bach, K. & Marioni, J.C., 2016. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, 17(1), p.133.

Lafzi, A. et al., 2018. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature protocols*, 13(12), pp.2742–2757.

Lamouille, S. et al., 2013. Regulation of epithelial–mesenchymal and mesenchymal–epithelial transitions by microRNAs. *Cell regulation*, 25(2), pp.200–207.

Lamouille, S., Xu, J. & Derynck, R., 2014. Molecular mechanisms of epithelial–mesenchymal transition. 15(3), pp.178–196.

Larsson, A.J.M. et al., 2019. Genomic encoding of transcriptional burst kinetics. *Nature*, 424, p.1.

Lawson, D.A. et al., 2015. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526(7571), pp.131–135.

Lawson, D.A. et al., 2018. Tumour heterogeneity and metastasis at single-cell resolution. *Nature Cell Biology*, 20(12), pp.1349–1360.

Le Roch, K.G. et al., 2003. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639), pp.1503–1508.

Levine, M. & Tjian, R., 2003. Transcription regulation and animal diversity. *Nature*, 424(6945), pp.147–151.

Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. I. Birol, ed. *Bioinformatics*, 34(18), pp.3094–3100.

Li, L. & Li, W., 2015. Epithelial–mesenchymal transition in human cancer: Comprehensive reprogramming of metabolism, epigenetics, and differentiation. *Pharmacology & Therapeutics*, 150, pp.33–46.

Lin, P., Troup, M. & Ho, J.W.K., 2017. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18(1), p.59.

Lopez-Rubio, J.-J. et al., 2007. 5′ flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Molecular microbiology*, 66(6), pp.1296–1305.

Lopez-Rubio, J.-J., Mancio-Silva, L. & Scherf, A., 2009. Genome-wide Analysis of

Heterochromatin Associates Clonally Variant Gene Regulation with Perinuclear Repressive Centers in Malaria Parasites. *Cell Host & Microbe*, 5(2), pp.179–190.

Love, M.I., Huber, W. & Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), p.550.

Macosko, E.Z. et al., 2015. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), pp.1202–1214.

McGinnis, C.S., Murrow, L.M. & Gartner, Z.J., 2019. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*.

Metzker, M.L., 2010. Sequencing technologies — the next generation. 11(1), pp.31–46.

Miller, L.H., Ackerman, H.C., Su, X.-Z. & Wellems, T.E., 2013a. Malaria biology and disease pathogenesis: insights for new treatments. *Nature medicine*, 19(2), pp.156–167.

Miller, L.H., Ackerman, H.C., Su, X.-Z. & Wellems, T.E., 2013b. Malaria biology and disease pathogenesis: insights for new treatments. *Nature medicine*, 19(2), pp.156–167.

Moffitt, J.R. et al., 2018. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, 362(6416), p.eaau5324.

Moreno-Bueno, G., Portillo, F. & Cano, A., 2008. Transcriptional regulation of cell polarity in EMT and cancer. 27(55), pp.6958–6969.

Mortazavi, A. et al., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7), pp.621–628.

Mouse Genome Sequencing Consortium et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Murray, C.J. et al., 2012. Global malaria mortality between 1980 and 2010: a systematic analysis. *The Lancet*, 379(9814), pp.413–431.

Nair, S. et al., 2014. Single-cell genomics for dissection of complex malaria infections. *Genome Research*, 24(6), pp.1028–1038.

Nieto, M.A. et al., 2016. EMT: 2016. *Cell*, 166(1), pp.21–45.

Otto, T.D. et al., 2010. New insights into the blood-stage transcriptome of Plasmodium falciparum using RNA-Seq. *Molecular microbiology*, 76(1), pp.12–24.

Özcan, M.A., Ilhan, O., Ozcebe, O.I., Nalcaci, M. & Gülbas, Z., 2013a. Review of therapeutic options and the management of patients with myelodysplastic syndromes. *Expert Review of Hematology*, 6(2), pp.165–189.

Özcan, M.A., Ilhan, O., Ozcebe, O.I., Nalcaci, M. & Gülbas, Z., 2013b. Review of therapeutic options and the management of patients with myelodysplastic syndromes. *Expert Review of Hematology*, 6(2), pp.165–189.

Painter, H.J. et al., 2018. Genome-wide real-time in vivo transcriptional dynamics during Plasmodium falciparum blood-stage development. *Nature communications*, 9(1), p.2656.

Painter, H.J., Campbell, T.L. & Llinás, M., 2011. The Apicomplexan AP2 family: Integral factors regulating Plasmodium development. *Molecular and Biochemical Parasitology*, 176(1), pp.1–7.

Parekh, S. et al., 2018. zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. *GigaScience*, 7(6), p.22.

Pastushenko, I. et al., 2018. Identification of the tumour transition states occurring during EMT. *Nature*, 556(7702), pp.463–468.

Patel, A.P. et al., 2014. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190), pp.1396–1401.

Patro, R. et al., 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4), pp.417–419.

Petropoulos, S. et al., 2016. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell*, 165(4), pp.1012–1026.

Petter, M. et al., 2011. Expression of P. falciparum var Genes Involves Exchange of the Histone Variant H2A.Z at the Promoter. *PLoS pathogens*, 7(2), p.e1001292.

Picelli, S., 2016. Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biology*, 14(5), pp.637–650.

Picelli, S. et al., 2013. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Meth*, 10(11), pp.1096–1098.

Pijuan-Sala, B. et al., 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 68, p.1.

Poran, A. et al., 2017. Single-cell RNA sequencing reveals a signature of sexual commitment in malaria parasites JA . 551, pp.95 EP –.

Posfai, E. et al., 2017. Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *eLife*, 6, p.2813.

Pyl, P.T., Anders, S. & Huber, W., 2014. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2), pp.166–169.

Raj, A. et al., 2008. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods*, 5, pp.877 EP –.

Ramskold, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P. & Sandberg, R., 2012a. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, 30(8), pp.777–782.

Ramskold, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O.R., Daniels, G.A., Khrebtukova, I., Loring, J.F., Laurent, L.C., Schroth, G.P. & Sandberg, R., 2012b. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, 30(8), pp.777–782.

Reid, A.J. et al., 2018. Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *eLife*, 7, p.e33105.

Reinius, B. et al., 2016. Analysis of allelic expression patterns in clonal somatic cells by single-cell RNA–seq. *Nat Genet*, advance online publication SP - EP .

Robinson, M.D. & Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), pp.R25–R25.

Rosenberg, A.B. et al., 2018. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385), pp.176–182.

Rovira-Graells, N. et al., 2012. Transcriptional variation in the malaria parasite Plasmodium falciparum. *Genome Research*, 22(5), pp.925–938.

Ru, P. et al., 2012. miRNA-29b Suppresses Prostate Cancer Metastasis by Regulating Epithelial–Mesenchymal Transition Signaling. *Molecular Cancer Therapeutics*, 11(5), pp.1166–1173.

Sandberg, R., 2013. Entering the era of single-cell transcriptomics in biology and medicine. *Nat Meth*, 11(1), pp.22–24.

Sasagawa, Y. et al., 2018. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biology*, 19(1), p.29.

Sasagawa, Y. et al., 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology*, 14(4), p.3097.

Satija, R. et al., 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotech*, 33(5), pp.495–502.

Shalek, A.K. et al., 2014. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505), pp.363–369.

Shapiro, E., Biezuner, T. & Linnarsson, S., 2013. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet*, 14(9), pp.618–630.

Shendure, J. et al., 2017. DNA sequencing at 40: past, present and future JA . 550, pp.345 EP –.

Shendure, J., Findlay, G.M. & Snyder, M.W., 2019. Genomic Medicine–Progress, Pitfalls, and Promise. *Cell*, 177(1), pp.45–57.

Sheng, K. et al., 2017. Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nature methods*, 14(3), pp.267–270.

Shibue, T. & Weinberg, R.A., 2017. EMT, CSCs, and drug resistance: the mechanistic link and clinical implications. *Nature Reviews Clinical Oncology*, 14(10), pp.611–629.

Shock, J.L., Fischer, K.F. & DeRisi, J.L., 2007. Whole-genome analysis of mRNA decay in Plasmodium falciparum reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle. *Genome Biol.*, 8(7), pp.R134–R134.

Siegel, T.N. et al., 2014. Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in Plasmodium falciparum. *BMC genomics*, 15, p.150.

Sims, J.S. et al., 2009. Patterns of gene-specific and total transcriptional activity during the Plasmodium falciparum intraerythrocytic developmental cycle. *Eukaryotic cell*, 8(3), pp.327–338.

Smith, T., Heger, A. & Sudbery, I., 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3), pp.491–499.

Snow, R.W. et al., 2017. The prevalence of <i>Plasmodium falciparum</i> in sub-Saharan Africa since 1900. *Nature*, 550(7677), pp.515–518.

Sorber, K., Dimon, M.T. & DeRisi, J.L., 2011. *RNA-Seq analysis of splicing in Plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts*,

Soumillon, M. et al., 2014. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*, p.003236.

Sperling, A.S., Gibson, C.J. & Ebert, B.L., 2017. The genetics of myelodysplastic syndrome: from clonal haematopoiesis to secondary leukaemia. *Nature Reviews Cancer*, 17(1), pp.5–19.

Spitz, F. & Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. 13(9), pp.613–626.

Stegle, O., Teichmann, S.A. & Marioni, J.C., 2015. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet*, 16(3), pp.133–145.

Stuart, T. & Satija, R., 2019. Integrative single-cell analysis. 2, p.1.

Subkhankulova, T., Gilchrist, M.J. & Livesey, F.J., 2008. Modelling and measuring single cell RNA expression levels find considerable transcriptional differences among phenotypically identical cells. *BMC genomics*, 9(1), p.268.

Svensson, V. et al., 2017. Power analysis of single-cell RNA-sequencing experiments. *Nat Meth*, 14(4), pp.381–387.

Tam, W.L. & Weinberg, R.A., 2013. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature medicine*, 19(11), pp.1438–1449.

Tang, F. et al., 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5), pp.377–382.

Taniguchi, K., Kajiyama, T. & Kambara, H., 2009. Quantitative analysis of gene expression in a single cell by qPCR. *Nature methods*, 6, pp.503 EP –.

Tehranchi, R. et al., 2010. Persistent Malignant Stem Cells in del(5q) Myelodysplasia in Remission. *N Engl J Med*, 363(11), pp.1025–1037.

Thiery, J.P., Acloque, H., Huang, R.Y.J. & Nieto, M.A., 2009a. Epithelial-mesenchymal transitions in development and disease. *Cell*, 139(5), pp.871–890.

Thiery, J.P., Acloque, H., Huang, R.Y.J. & Nieto, M.A., 2009b. Epithelial-mesenchymal transitions in development and disease. *Cell*, 139(5), pp.871–890.

Tirosh, I., Izar, B., et al., 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282), pp.189–196.

Tirosh, I., Venteicher, A.S., et al., 2016. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, 539(7628), pp.309–313.

Todryk, S.M. & Hill, A.V.S., 2007a. Malaria vaccines: the stage we are at. 5(7), pp.487–489.

Todryk, S.M. & Hill, A.V.S., 2007b. Malaria vaccines: the stage we are at. 5(7), pp.487–489.

Trapnell, C. et al., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4), pp.381–386.

Ungai-Salánki, R. et al., 2016. Automated single cell isolation from suspension with computer vision. 6, pp.20375 EP –.

Valacca, C. et al., 2010. Sam68 regulates EMT through alternative splicing–activated nonsense-mediated mRNA decay of the SF2/ASF proto-oncogene. *The Journal of Cell Biology*, 191(1), pp.87–99.

Vallejos, C.A. et al., 2017. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature methods*, 14(6), pp.565–571.

Vallejos, C.A., Marioni, J.C. & Richardson, S., 2015. BASiCS: Bayesian Analysis of Single-Cell Sequencing Data Q. Morris, ed. *PLoS Comput Biol*, 11(6), p.e1004333.

van der Maaten, L. & Hinton, G., 2008a. Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9(2579-2605), p.85.

van der Maaten, L. & Hinton, G., 2008b. **Visualizing Data using t-SNE**. *Journal of Machine Learning Research*. Available at: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

Vembar, S.S., Scherf, A. & Siegel, T.N., 2014. Noncoding RNAs as emerging regulators of Plasmodium falciparum virulence gene expression. *Host–microbe interactions: fungi/parasites/viruses*, 20 IS -, pp.153–161.

Vieth, B. et al., 2017. powsimR: power analysis for bulk and single cell RNA-seq experiments. I. Hofacker, ed. *Bioinformatics*, 33(21), pp.3486–3488.

Volz, J.C. et al., 2012. *PfSET10, a Plasmodium falciparum Methyltransferase, Maintains the Active var Gene in a Poised State during Parasite Division*,

Wagner, A., Regev, A. & Yosef, N., 2016. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotech*, 34(11), pp.1145–1160.

Wagner, D.E. et al., 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392), pp.981–987.

Wahlgren, M., Goel, S. & Akhouri, R.R., 2017. Variant surface antigens of Plasmodium falciparum and their roles in severe malaria. 15(8), pp.479–491.

Wang, B. et al., 2017. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods*, advance online publication SP - EP .

Warren, L. et al., 2006. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences*, 103(47), pp.17807–17812.

Warzecha, C.C. et al., 2010. An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *The EMBO Journal*, 29(19), pp.3286–3300.

Wescoe, Z.L., Schreiber, J. & Akeson, M., 2014. Nanopores discriminate among five C5-cytosine variants in DNA. *Journal of the American Chemical Society*, 136(47), pp.16582–16587.

Wheelock, M.J. et al., 2008. Cadherin switching. *Journal of Cell Science*, 121(6), pp.727–735.

Wills, Q.F. et al., 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8), pp.748–752.

Winzeler, E.A. et al., 2016. Eradicating Malaria: Discoveries, Challenges, and Questions. *Cell*, 167(3), pp.595–597.

Wolf, F.A., Angerer, P. & Theis, F.J., 2018. SCANPY : large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), p.15.

Woll, P.S. et al., 2014. Myelodysplastic Syndromes Are Propagated by Rare and Distinct Human Cancer Stem Cells In Vivo. *Cancer Cell*, 25(6), pp.794–808.

Wolock, S.L., Lopez, R. & Klein, A.M., 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*.

Woodhouse, S. et al., 2015. Processing, visualising and reconstructing network models from single-cell data. 94(3), pp.256–265.

Wu, A.R. et al., 2014. Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods*, 11(1), pp.41–46.

Xue, Z. et al., 2013. Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, 500(7464), pp.593–597.

Ye, X. et al., 2015. Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature*, 525, pp.256 EP –.

Yilmaz, M. & Christofori, G., 2009a. EMT, the cytoskeleton, and cancer cell invasion. *Cancer and Metastasis Reviews*, 28(1-2), pp.15–33.

Yilmaz, M. & Christofori, G., 2009b. EMT, the cytoskeleton, and cancer cell invasion. *Cancer and Metastasis Reviews*, 28(1-2), pp.15–33.

Yip, S.H. et al., 2017. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic acids research*, 45(22), pp.e179–e179.

Yook, J.I. et al., 2006. A Wnt–Axin2–GSK3β cascade regulates Snail1 activity in breast cancer cells. 8(12), pp.1398–1406.

Yosef, N. & Regev, A., 2016. Writ large: Genomic dissection of the effect of cellular environment on immune response. *Science*, 354(6308), pp.64–68.

Young, J.W. et al., 2012. Measuring single-cell gene expression dynamics in bacteria using fluorescence time-lapse microscopy. *Nature protocols*, 7(1), pp.80–88.

Zabidi, M.A. & Stark, A., 2016. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in Genetics*, 32(12), pp.801–814.

Zappia, L., Phipson, B. & Oshlack, A., 2018. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database D. Schneidman, ed. *PLoS Comput Biol*, 14(6), p.e1006245.

Zappia, L., Phipson, B. & Oshlack, A., 2017. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1), p.333.

Zeisel, A. et al., 2015. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226), pp.1138–1142.

Zhang, J. et al., 2012. miR-30 inhibits TGF-β1-induced epithelial-to-mesenchymal transition in hepatocyte by targeting Snail1. *Biochemical and Biophysical Research Communications*, 417(3), pp.1100–1105.

Zhang, X. et al., 2019. Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular cell*, 73(1), pp.130–142.e5.

Zheng, G.X.Y. et al., 2017. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8, p.14049.

Ziegenhain, C. et al., 2017. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular cell*, 65(4), pp.631–643.e4.