Università degli studi di Padova

COMPUTATIONAL DETECTION OF DOUBLETS IN SINGLE-CELL RNA SEQUENCING

Laureanda: **Francesca Coden**

Matricola 1171373

Relatore: **Barbara Di Camillo**

Correlatori: **Zlatko Trajanoski**
Francesca Finotello

Corso di Laurea Magistrale in Bioingegneria

Anno Accademico 2018/2019
Padova, 09 Dicembre 2019

Sommario

Lo studio del RNA permette l'analisi delle componenti funzionali delle cellule e delle loro espressioni geniche. È stata sviluppata una tecnica, chiamata sequenziamento dell'RNA (RNA-seq), che permette l'analisi e l'identificazione di questi componenti. L'analisi standard, definita come bulk RNA-seq, studia l'espressione dell'RNA di grandi popolazioni di cellule e fornisce, come valore di espressione, la media dei valori di espressione dell'RNA nella popolazione. Poiché questa misurazione potrebbe nascondere quelle che sono le differenze tra le singole cellule in una popolazione mista, negli ultimi anni è stata modificata e migliorata con lo sviluppo del sequenziamento dell'RNA a singola cellula (scRNA-seq). Questa nuova tecnica permette lo studio di singole cellule rendendo visibili nuove cellule o sottopopolazioni di cellule.

Nonostante lo sviluppo del sequenziamento a singola cellula abbia portato miglioramenti nello studio dell'espressione genica, esso appare anche caratterizzato da diversi problemi. Uno dei più comuni è il problema delle doublets, che si ha con l'aggregazione o la cattura di due cellule insieme durante l'analisi. Queste cellule sono fattori di confusione che possono portare a errati risultati nell'analisi, poiché hanno espressione genica ibrida e possono essere scambiate con nuovi tipi di cellule e dovrebbero essere eliminate dai dati.

Diversi metodi sperimentali e computazionali sono nati per combattere questo problema attraverso la classificazione delle cellule considerate come doublets. Questo lavoro descriverà questi strumenti e analizzerà due dei metodi computazionali, DoubletFinder e DoubletDecon, creati per l'identificazione di tali cellule. Si focalizzerà sull'analisi del loro comportamento e delle loro prestazioni. Set di dati simulati e reali sono utilizzati per confrontare i due metodi e i loro risultati. Viene sviluppato un terzo metodo in cui le cellule classificate come doublets sono le cellule considerate come doublets da entrambi i metodi precedentemente considerati. Viene utilizzato per analizzare la concordanza tra DubletDecon e DoubletFinder e per valutare se questo approccio può essere utile per future analisi.

Abstract

The study of the RNA allows the analysis of the functional components of the cells and their gene expressions. A technique, called RNA sequencing (RNA-seq), that allows the analysis and the identification of these components has been developed. The standard analysis, defined as bulk RNA-seq, studies the expression of RNAs from large populations of cells and gives, as expression value, the average of its expression levels across the population. Since this measurement could hide differences between individual cells in mixed population, in the recent years it was modified and improved with development of the single-cell RNA sequencing (scRNA-seq). This new technique allows the study of single cells making new cells or subpopulations of cells visible.

Despite the development of single-cell RNA sequencing led to improvements in gene expression study, it is also characterized by different problems. One of the most common is the doublets problem that arise with the aggregation or capture of two cells together during the analysis. These cells are confounding factors that can lead to false analysis results, since they have hybrid gene expression and can be exchanged for new cell types and should be eliminated from the data.

Different experimental and computational methods are born to “fight” this problem with the classification of the cells considered as doublets. This work will describe these tools and analyse two of the computational methods, DoubletFinder and DoubletDecon, created for the analysis of such cells. It will focus on the analysis of their behaviour and performances. Simulated and real datasets are used to compare the two methods and their results. A third methods is developed in which the cells classified as doublets are only the cells considered as doublets in the other two methods. It is used to analyse the concordance between DoubletDecon and DoubletFinder and to assess if this approach could be useful for future analysis.

Table of Contents

Introduction

The genome is the set of all the instructions useful to each cell of the organism to encode the proteins necessary for survival, interaction with the environment and reproduction, i.e. life itself.

DNA contains all this information that is useful to the cells in determining their function and properties. The cells have access to this information thanks to gene expression that allows, by selecting specific sets of genes, to pass this “data” first to the RNA and then, in most cases, to the proteins. RNA is a molecule used to transport the information that reflect the state of the cell in defined conditions. It can also be used to identify and analyse the different mechanisms that exist, for example in tissues or organs, that lead to diseases or characterise different gene expressions.

RNA-sequencing was devised to measure the different information in the various states. It is a method for RNA profiling that permits the identification of gene expression patterns and maps the transcribed parts of the genome.

To better understand cell behaviour and cell composition, the study of gene expression or protein expression has improved over the last few years (~10 years), with the invention of methods for single-cell RNA sequencing (scRNA-seq). These methods are important for the identification of cell populations and they help to study complex biological systems measuring the expression of the genes that are expressed in thousands of single cells. scRNA-seq allows, with the analysis of different cell genes in different population in tissues or organs, to detect the existence of new cell types that were not visible with earlier technologies. The study of similarities and differences among cells in a population, thanks to the comparison between their transcriptomes, also permits the discovery of new kinds of subpopulations. The transcriptome analysis of single cells helps also to analyse the cell characteristics before and after cellular states transitions, like differentiation or development. As such it is possible to note the nature of some cell mechanisms that regulate these cell states. It could be useful to understand which genes are important for cellular processes.

All scRNA-seq methods usually follow similar steps that are: RNA molecule capture, reverse transcription and its amplification, preparation of sequencing libraries, and sequencing.

While these methods are important to improve the understanding of biological systems, they have some issues, in fact they tend to produce technical artefacts. One of these is especially problematic, the doublets creation that appear when two cells are captured together. This can be a problem for single-cell analysis since cell expression profiles are mixed and hybrid transcriptomes are generated. The mixture of the profiles of the two captured cells can be interpreted as cell types or states that do not exist leading to false results in the analysis. To improve the study of the RNA expression the doublets should be removed from the data.

To resolve this problem, different experimental and computational methods have been concocted to estimate the probability of each cells, sequenced with scRNA-seq, of being a doublet or to classify the cells into doublets or singlets. They take on different approaches and are yet to provide a uniquely efficient workflow to underline the problem.

This thesis will outline the doublet detection problem and various tools for doublet classification. It will focus on two computational methods, DoubletFinder and DoubletDecon, by attempting to apply them to both simulated and real datasets. The simulated datasets are created in such a way that the classification of the cells into doublets or singlets is already known. This allows the comparison between the true classification and the classification carried out by the methods. A third method is implemented, for the sake of this research, to measure the differences or equality in the results of the other two methods. The analysis will then show their classification efficiency and try to characterise their differences.

In the first chapter the themes of gene expression and RNA-sequencing are briefly described. Next single-cell RNA sequencing is explained describing data generation and the most commonly used technologies for scRNA-seq are exposed and described. The computational analysis is explained with focus on the possible challenges it faces.

In the second chapter the problem of doublets detection is explained in further detail. Experimental and computational tools for doublets identification are presented and described.

The third chapter focuses upon the material and methods used for the analyses behind this thesis. Two datasets, for which ground truth data are provided, are used to validate two computational methods for doublet detection, DoubletFinder and DoubletDecon. Another method is implemented for doublet identification that merges the results from the other two methods. A case study is also put forward to test the performances of the three methods on a real dataset. A pre-processing pipeline is used and described to prepare the data for the analysis.

The results of the analyses on the three datasets are laid out in the last chapter.

1. Single-cell RNA sequencing

1.1 Gene expression

Gene expression represents the process by which the information encoded in genes in DNA is used to synthesise protein or RNA structures present and operating in the cell. The expressed genes include genes that are transcribed into messenger RNA (mRNA) and then translated into protein, as well as genes that are transcribed into RNA, such as transfer and ribosomal RNAs, but not translated into protein.

The process of gene expression follows several steps, expressed in Figure 1, that are: transcription, RNA splicing, RNA export and translation.

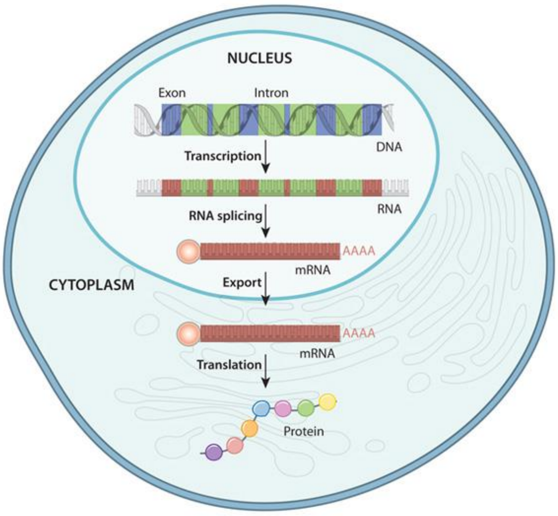


Figure 1. Description of the four steps used to on gene expression to transcribe DNA into RNA and then to translate the RNA into protein: transcription, RNA splicing, RNA export, and translation. Each step is represented by a labelled arrow. Transcription of a DNA template to a pre-mRNA and the splicing of the pre-mRNA into a mature mRNA are shown inside the cell nucleus. The RNA export brings the mature mRNA to the cytoplasm, where the mature mRNA message is translated into a protein.

The transcription is the process in which an RNA molecule is synthesised from a DNA sequence. In the first phase, called initialisation, the enzymes that permit the transcription recognise a DNA area that corresponds to the gene of interest for RNA synthesis. In fact, nearby the gene to be transcribed, there is a site promoter featured by a sequence of 10-300 nucleotides that are treated as consensus signal by the enzymes. The segment of DNA is read thanks to the enzyme RNA polymerase that links to the DNA and, with the help of some proteins, separates the two chains of DNA. The second phase of the transcription is the elongation in which the RNA chain is synthesised. The RNA polymerase moves along the DNA from 3’ end to 5’ end. One nucleotide at a time is added at the end 3’ of the RNA according to the pairing of the bases. As the transcription proceeds, the newly formed RNA chain detaches itself from the DNA. During the process the enzymes meet a sequence that signals the end of the transcription. This phase is called termination. At this time a chain of pre-RNA is formed.

The second step is the RNA splicing where the introns, non-coding regions for the proteins in eukaryotes, are removed to leave only the regions carrying information, the exons. Introns are removed from primary transcripts by the identification of specific sequences called splice sites. The RNA resulting from this step is called messenger RNA (mRNA) since it contains the message for protein synthesis. Another modification that could occur to the RNA in this phase is the creation of the poly(A) tail. It represents the addition of 100-200 nucleotides to the 3’ end of the pre-RNA, useful in the recognition of the mRNA for protein synthesis.

In the RNA export the mRNA is transported from the nucleus to the cytoplasm. Although some RNAs function in the nucleus, most are carried through pores in the nucleus into the cytosol, including all RNAs involved in protein synthesis.

The last step, the translation, is a process during which a polypeptide is synthesised from the genetic information that exists in the sequence of mRNA. Needed for this step are the mRNA synthesised, ribosomes and transfer RNA (tRNA).

Ribosomes are molecular complexes in which translation happens. They are formed by two subunits, one smaller and one bigger. Each subunit is formed by proteins and non-messenger RNA called ribosomal RNA (rRNA).

tRNA is an RNA molecule that allows to convert the language gene (nitrogenous bases) to the language of the proteins (amino acids). It consists of a site for the binding of an amino acid and a site that contains three complementary nucleotides to the triplet (codon) that codifies for the amino acid linked to the tRNA (anticodon).

The translation begins when mRNA meets the smaller subunit of the Ribosome and the tRNA that transports the amino acid methionine and links itself to the codon of START. This fixes the beginning of the initialisation phase. In a second moment the bigger subunit of the Ribosome binds itself to the mRNA. This subunit is formed by three enzymatic sites:

- E (Exit)
- P (Peptide)
- A (Access)

In the initial phase the tRNA with methionine amino acid is placed in the P site.

Next, the second phase of the translation, called elongation, is implemented by three steps. In the first step, the codon recognition, the tRNA anticodon recognises the codon in mRNA content and sticks to the A site of the Ribosome. tRNA brings amino acid codon specific. Next the formation of a peptide bond between the two amino acids in the sites A and P occurs. The amino acid in the P site leaves the tRNA. In the last step, translocation, the two tRNA shift their location occupying the sites E and P and freeing the A site. The tRNA in E position leaves the Ribosome.

At the end of the translation the site A of the Ribosome reaches the STOP codon and the polypeptide chains, and the Ribosome is released.

1.2 RNA sequencing

RNA sequencing (RNA-seq) is a method for RNA profiling where mRNA molecules in a sample are recognised and analysed quantitatively. It is essential to identify the functional elements of the genome and to understand when and how they are expressed and regulated. This allows a more complete mapping of the transcribed regions of the genome. It uses next-generation sequencing (NGS) to measure expression across the transcriptome and it is helpful in the study of cellular responses such as discovery of undetected changes in states of the diseases or response to therapies. RNA-seq allows description of transcription without knowing in advance the genomic sites of transcription origin and has a great ability to distinguish RNA isoforms, determine allelic expression, and reveal sequence variants.

Typically, the steps of an RNA-seq experiment consist in, as shown in the example of Figure 2, the isolation of RNA/mRNA, the fragmentation and conversion of the fragments into complementary DNAs (cDNAs), the amplification of cDNAs later subject to NGS for the sequencing and the mapping of the reads generated on the reference genome.

In the RNA isolation step, it is important to ensure a sufficient quality of RNA since a low-quality RNA, characterised by a genomic sequence degradation, can significantly influence sequencing results and lead to misled biological conclusions.

In the second step, the isolated RNA is chemically fragmented into pieces which are then converted into cDNAs for the subsequent analysis with NGS technologies. The cDNA is the DNA synthesised in vitro from a mould of RNA by the action of the reverse transcriptase. This process is called reverse transcription (RT). The reverse transcriptase is an enzyme that works on a single filament of mRNA generating its complementary DNA through the pairing of the nitrogenous bases. After the conversion some adapters are added at the end of each fragment to permit the sequencing. The oligo(dT) primers are used the most and are composed of the union of different nucleotides (around 12-18) and may have different lengths.

They are created to bind to the complementary poly(A) tails of mRNA and for this reason they are suited to the analysis of mRNA expression.

During the following step the synthesised cDNAs are amplified, i.e. several identical copies (clones) of the considered fragments are created. The most commonly known and most used techniques are Polymerase Chain Reaction (PCR) and in vitro transcription (IVT).

The PRC is developed in different steps. First, comes the DNA denaturation, i.e. the division of its chains and next, the primer annealing that is the creation of a link between the primers and the segments to be reproduced. The primers are appropriate triggers consisting of short DNA sequences (oligonucleotides) complementary to the end 5' and 3'of the two filaments. At the end of this process a copy of the filaments is made using DNA polymerase and DNA ligase.

IVT, on the other hand, consists in a first step of reverse transcription and cDNA synthesis with an Oligo(dT) primer containing a promoter for the T7 polymerase, that is a specific RNA polymerase for the formation of RNA from DNA from 5 end to 3 end. In the IVT it repeatedly binds to the promoter and amplifies RNA, which eventually undergoes a final step of RT.

After the amplification, the data are sequenced with NGS and the results of the sequencing, called reads, represented as character strings, are saved in public databases to be used for the analysis of genome expression. During the sequencing the nucleotides characterising a fragment are identified, thus allowing the identification of the components of the sequence.

The most commonly used NGS platform is Illumina. It is characterised by a bridge-PCR in which DNA molecules are amplified, while immobilised on the surface of a glass flow cell, by “arching” over anchor oligonucleotide fixed on the flow cell surface by hybridisation. Different amplification cycles transform the single-molecule DNA form into clonally amplified cluster with about 1000 clonal molecules. For each flow cell, millions of clusters can be generated. Its sequencing is characterised by the use of fluorescent labelled termination nucleotides that are embedded according to the complementary sequence in each strand of a cluster. Illumina, like other NGS technologies, can apply a “paired-end” sequencing in which both ends of the molecules are sequenced. In such way it provides positional information that make the alignment and the assembly easier, especially for short reads.

Next the reads are aligned on the reference genome allowing the identification of gene regions that agree with the strings from the sequencing. There are different tools for the mapping but all of them have in common the fact that the process begins with the creation of an index of the reference genome or the reads that is used to identify the positions, in the sequence of reference, where the reads are more likely to align themselves.

After the identification of these possible regions of alignment, the mapping is executed permitting the reconstruction of the transcript’s sequences and the quantification of their abundance. The number of reads that is possible to align to each gene is used as a measurement of the gene expression levels in the sample of interest and are called counts.

The standard analysis, defined as bulk RNA-seq, analyse the expression of RNAs from large populations of cells and the expression value discovered for each gene is the average of its expression levels across the population. This represents a problem for the analysis, in fact considering an example in which a study is carried out on mixed cell populations, such measurement could “cover” differences between individual cells within the populations. Another question of bulk RNA-seq is the absence of an insight on the stochastic nature of gene expression. If genetically identical cells are considered, exposed to the same environmental condition, this analysis is unable to see the significant changes in molecular content and the differences in phenotypic characteristics. This sequencing is also unable to examine the kinetics of the transcripts, not allowing the tracking of the expression of individual genes over time.

All these questions can be solved thanks to a new approach, single cell RNA-seq in which expression profiles of a single cell is provided.

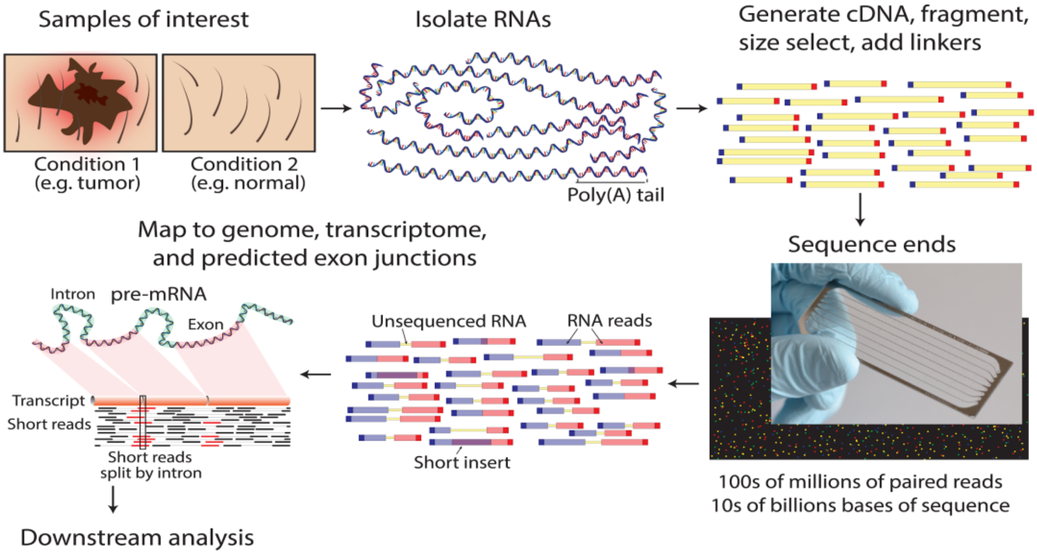

Samples of interest
Condition 1
(e.g. tumor)
Condition 2
(e.g. normal)
Isolate RNAs
Poly(A) tail
Generate cDNA, fragment, size select, add linkers
Sequence ends
100s of millions of paired reads
10s of billions bases of sequence
Unsequenced RNA
RNA reads
Short insert
Map to genome, transcriptome, and predicted exon junctions
Intron
pre-mRNA
Exon
Transcript
Short reads
Short reads split by intron
Downstream analysis

Figure 2. *Description of RNA-seq steps: RNAs isolation from the sample; fragmentation and conversion of the fragments into complementary DNA (cDNA), addition of primers the cDNA end; sequencing and mapping of the reads created with the sequencing. (Griffith, Walker, Spies, Ainscough, & Griffith, 2015)*

1.3 Single-cell RNA sequencing

1.3.1 Technologies and data generation

Single-cell RNA sequencing (scRNA-seq) is the gene expression profiling of singlet cells. It can show a large variety of cell types and subpopulation that were unseen with traditional experimental techniques, and it also led to the discovery of new information in regard to the cell composition. scRNA-seq describes RNA molecule with high resolution and on the genomic level allowing the comparison of single cell transcriptome.

scRNA-seq is not a single method, there are different suitable protocols that can be chosen to study RNA depending on the type of study required. For example, if the interest of the research is to obtain an high number of details per cell, the protocols to use are ones with higher sensitivity while, if the interest is on the quantification of the transcriptome of large number of cells, other protocols may be of more use for the analysis.

Even if there are many different protocols, they all follow some common steps of analysis which are: single cell capture, reverse transcription (RT) and cDNA synthesis, cDNA amplification, libraries preparation and sequencing.

First, it must be considered the fact that the condition of the cells is fundamental for cell capture efficiency and for an optimal performance of the scRNA-seq protocol. For this reason, it is important to ensure that high-quality individual cell suspensions are prepared, whether fresh or frozen cells are used. To be sure to work with high quality cells, it is necessary to minimalize cell aggregation, dead cells, non-cellular nucleic acid and RT inhibitors.

The steps of RT and cDNA synthesis are similar to those for RNA-seq. The difference is that, after the reverse transcription, single cell specific barcodes are added to the poly(T) oligonucleotide to help in the correction of amplification bias and in technical noise reduction. Some methods use also unique molecular identifiers (UMIs) that are short sequences added to the transcripts of interest before amplification to detect and quantify unique RNA molecules.

The amplification of cDNA can be achieved by PCR or with IVT. In choosing the protocol to use for the analysis, it must be kept in mind that IVT leads to linear amplification and for this reason the protocols that use it have less amplification biases, but additional downstream is required to convert the amplified RNA into cDNA and sequencing libraries. PCR instead is a nonlinear amplification process and that leads to biases in the composition of RNA in the final libraries.

Another difference between the protocols is that some give the full-length transcript of the data while others count only the 3’ “tagged” end of each mRNA. Full-length transcriptome sequencing permits the discovery of splice variants and alternative transcripts, but it does not allow the inserting of UMIs and cellular barcodes, which implies higher costs in the preparation of the libraries.

scRNA-seq methodologies can be divided into three groups that differ according to their protocol approach (Lafzi, Moutinho, Picelli, & Heyn, 2018):

- Microtiter-plate-based approach;
- Microfluidic systems-based approach;
- Split-pool barcoding-based approach.

The first approach is based on the idea that the single cells are isolated into microtiter plates by fluorescence activated cell sorting (FACS) and on the application of full-length transcript or 3'-end protocol. A commonly used method based on this approach is Smart-seq2 (Picelli et al., 2013, 2014). It sorts the single cells into well PCR plates containing lysis buffer using FACS. The mRNA is then primed with an oligo(dT) primer and the reverse transcription is performed until the enzyme has reached the 5′-end of each mRNA transcript (Figure 3). After RT, cytosines are added to the cDNA by an enzyme allowing the template-switching reaction that represents the ability of the reverse transcriptase to introduce a few un-templated cytosines when it reaches the 5′-end of the RNA template, that corresponds to the 3′-end of the synthesised cDNA. These cytosines help the Template Switching Oligo, an oligonucleotide that allows the reverse transcriptase to "switch" the template and then to synthesise the cDNA. The introduction of this sequence at the transcript end permits an efficient cDNA amplification in the following PCR step. Subsequently, the tagmentation, that at the same time fragments and indexes the cells, is used to prepare the sequencing libraries. The characteristic of Smart-seq2 is that it generates full-length libraries, choice that could be disadvantageous since it does not allow early barcoding of the cells and the adding of UMIs.

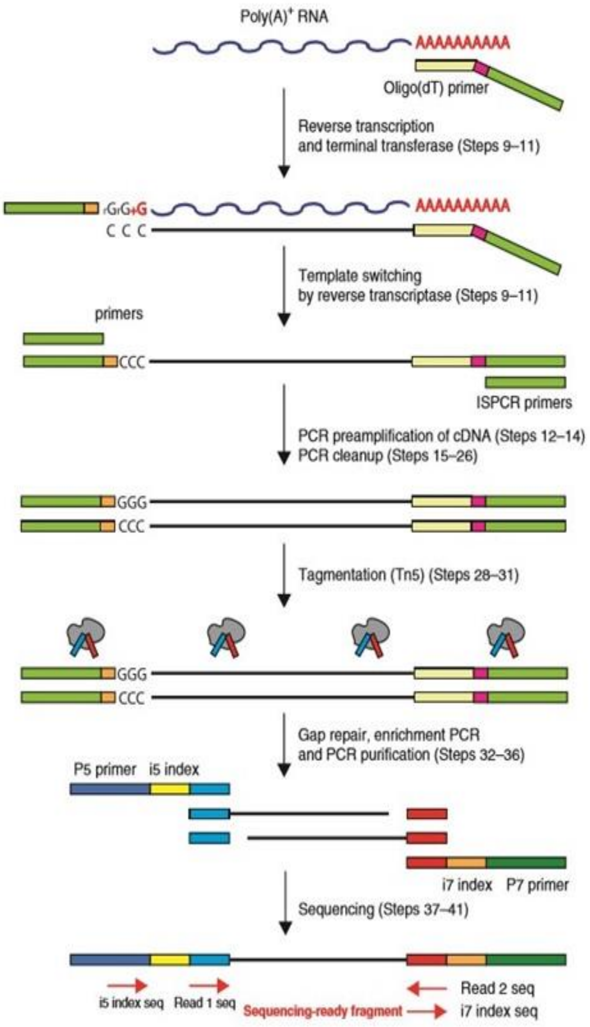

Poly(A)$^+$ RNA
AAAAAAAAAA
Oligo(dT) primer
Reverse transcription
and terminal transferase (Steps 9–11)
rGrGrG
CCC
Template switching
by reverse transcriptase (Steps 9–11)
primers
ISPCR primers
PCR preamplification of cDNA (Steps 12–14)
PCR cleanup (Steps 15–26)
GGG
Tagmentation (Tn5) (Steps 28–31)
Gap repair, enrichment PCR
and PCR purification (Steps 32–36)
P5 primer
i5 index
i7 index
P7 primer
Sequencing (Steps 37–41)
i5 index seq
Read 1 seq
Sequencing-ready fragment
Read 2 seq
i7 index seq

Figure 3.Steps of Smart-seq2 library preparation: oligo(dT) primer binding end reverse transcription; template-switching reaction and cDNA synthesis; cDNA amplification; tagmentation; sequencing. (Picelli et al., 2013, 2014)

Another protocol for the microtiter-plate-based approach is SCRB-seq (Macosko et al., 2015; Soumillon, Cacchiarelli, Semrau, van Oudenaarden, & Mikkelsen, 2014). With this method, after the cells sorting by FACS, cDNA is generated using oligo-dT priming, template switching and PCR amplification of cDNA. In this case the oligo-dT primers contain cell-specific barcodes and UMIs, thus allowing the 3' end amplification of the transcripts. The cDNA from each plate can be pooled and then transformed into sequencing libraries thanks to a modified tagmentation in which 3' end enrichment takes place.

Of interest is also MARS-seq protocol (Jaitin et al., 2014) that implements high-throughput transcriptional profiling of single cells. In this method, after the cell isolation and the RT, the cDNA is amplified through IVT and barcodes and UMIs are added to the transcript.

A fourth method considered following the same approach is CEL-seq2 (Hashimshony et al., 2016). This protocol uses UMIs and linear amplification by in vitro transcription. After IVT the cDNA is synthesised by random priming, that use short oligonucleotides as primers in a random sequence.

The second scRNA-seq approach, microfluidic systems-based, allows higher-throughput scRNA-seq workflow and the downscaling of the reaction volumes from microliters to nanolitres. There exists three ways for capturing the cells: (1) integrated microfluidic chips (IFCs), (2) nanowells and (3) droplets.

All these strategies are characterised by a higher number of cell capture sites than the previous approach.

The first microfluidics system created was Fluidigm C1, that automatically captures single cells in fluidic circuits where they are immobilised on hydrodynamic traps. The captured cells are lysed and subject to reverse transcription and pre-amplification in nanolitre reaction chambers. This is an automated array solution. A limitation of this protocol is the fact that it works only with cells that of a similar size since the formats of the array, i.e. of the capture sites, have only three dimensions specific for the cells (small, medium and large). Another limitation is represented by the capture efficacy which can be low when dealing with sticky or non-spherical cells.

To increase the cell numbers, microfluidic-based technologies shift to the implementation of nanowell systems, which reduce reagent costs and environmental contamination of RNA. An example is STRT-seq in which the cells are loaded in nanowell platforms where cell barcode and UMI are incorporated on the 5’ end of the transcripts. On this platform the second-strand synthesis in the RT reaction is usually carried out using a template-switching mechanism while the amplification of cDNA is achieved by PCR.

IFC and nanowell strategies, although capable of working with many cells and have higher yields, are limited by the number of reaction sites. To overcome this restriction, droplet-based systems, based on the encapsulation of the cells in nanolitre-size emulsion droplets, came about. Two methods, based on this strategy, originated in parallel: inDrops (Klein et al., 2015) and Drop-seq (Macosko et al., 2015).

For the cell encapsulation inDrops uses hydrogel bead containing lysis buffer, RT reagents and oligonucleotides primers with known barcodes. The primers are photo-releasable, and their detachment improve molecule-capture efficiency and initiate in-drop RT reactions. The greatest challenge is to ensure that each droplet contains primers that can encode a different barcode. The barcoded cDNAs are amplified through in vitro transcription and sequenced. A variation of this protocol is represented by 10x Genomics Chromium platform in which each gel bead, called GEM (gel bead in emulsion), is created to contain oligonucleotides with RT primers, UMIs and cell barcodes. It is defined as GemCode technology. Cells are loaded at a slow dilution to minimise the possibility of multiple cells encapsulated in the same GEM. After encapsulation, cell lysis takes place, the gel beads are broken, and the oligonucleotides contained in them are released to carry out reverse transcription. Each cDNA molecule resulting from this process, encloses a UMI and shared barcode per GEM, and ends with a template switching oligo at the 3' end. Next, the barcoded cDNAs are amplified using PCR amplification, with primers complementary switch oligos and sequencing adapters.

The other protocol, Drop-seq, consists of the union of a flow of beads suspended in lysis buffer and a flow of cells. Each bead is composed of oligonucleotides primers containing: a constant sequence (identical on all primers and beads) used as a priming site for downstream PCR and sequencing; a "cell barcode"; a UMI and an oligo(dT) sequence.

The split-pool barcoding-based approach is based on the isolation of the cells in pools, which are divided and mixed with the addition, at each turn, of pool-specific barcodes. The barcodes resulting from this process are unique to each cell through their random assignment during consecutive pooling. A method that uses this approach is SPLiT-seq. It uses four rounds to index the cells, next two rounds of index ligation and a final PCR indexing step to create cell-specific barcoded 3'-transcript libraries. In the indexing UMIs are also added to correct amplification bias.

1.3.2 Computational analysis

scRNA-seq brought about many improvements in the study of cells such as the ability to discover new cell types, an insight on the significant variation in the molecular content and on the differences in the phenotypic characteristics. It has also provided insight into the kinetics of the transcript. It should be noted, however, that despite the improvements it has brought to the study of the gene expression, it has also brought with it many challenges in the analysis of the data.

First, scRNA-seq protocols, during the capture of the cells, tend to stress the cells and that sometimes leads to the rupture and the killing of some. These kinds of cells, called low quality cells, should be discarded from the analysis since they may lead to false results. Additionally, during this process one may come across some empty capture or some sites containing more than one cell (multiplets). These are confounding factors that should be considered and eliminated from the study.

Another challenge is represented by the sparsity of the data. This implies that, since the distance between the cells is big, there is a lot of empty space between the data points. In this case it is possible to notice a high proportion of zero read counts. This is a problem, because, by comparing cells in large spaces, the distances between cells become more homogeneous and therefore makes it more difficult to distinguish between differences in the same population or between different populations. To resolve this problem a dimension reduction technique is usually used such as PCA (Principal Component Analysis). It is a deterministic algorithm that uses a linear transformation to redistribute data in a smaller number of independent dimensions than the number of the original space. Another dimension reduction technique is t-distributed stochastic neighbour embedding (t-SNE) in which a probability distribution that permits the grouping together of data with similar probability is computed. Linked to this problem are also dropouts' events that represent the events in which genes expressed in the cell are not detected through sequencing and therefore are not present in the final dataset. These are treated as zeros (false zeros) in the analysis leading to the loss of information. These dropout events are mainly due to RT inefficiency and shallow sequencing depths and are still open challenges in the study of scRNA-seq.

scRNA-seq permits the analysis of a high number of cells but it can become a problem in the analysis of the data since they can appear heavy on a computational level. To resolve this issue sparse matrix, where the zeros are translated in a different way in order to save memory, are used.

During scRNA-seq studies it is also possible to come across batch effects. They are systematic differences that can be seen in gene expression levels between independent cells from the same population, caused by the variability between experimental batches. They can be considered as a source of variation in the data which is confused with the biological signal of interest and are commonly identified with PCA and diagnostic plots.

Following the development of scRNA-seq data analysis, this field is characterised by the presence of numerous computational tools to analyse the datasets, as can be seen in the database present on www.scRNA-tools.org (Zappia, Phipson, & Oshlack, 2018). The platforms for such analysis mainly use R or Python language. New tools are constantly being produced so being able to follow the development of new platforms is always difficult. In addition, the pipelines for data analysis are not standardised, therefore the tasks they perform vary from tool to tool and to choose which to use can be challenging.

Although the tasks carried out by these tools are different, they may be grouped into four analysis stages (Figure 4):

1. Data acquisition;
2. Data cleaning;
3. Cell assignment;
4. Gene identification.

The first stage considers the raw reads resulting from the sequencing and creates a matrix that represents the expression of each gene in each cell. During this phase, the reads are aligned to the reference genome or to the transcriptome, their assembly is carried out and their expression is quantified. It has to be taken into account the fact that more and more often the tools, in this point of the analysis, must be created to manage Unique Molecular Identifiers.

The second phase, the data cleaning, is focused on cell quality control, i.e. the removal of low-quality cells, and on the filtering of genes that are not of interest for the analysis or are lowly expressed. Some tools may also carry out data normalisation or missing value attribution. Normalisation is commonly used to remove unwelcome variations that might affect the results by levelling the data from the different cells.

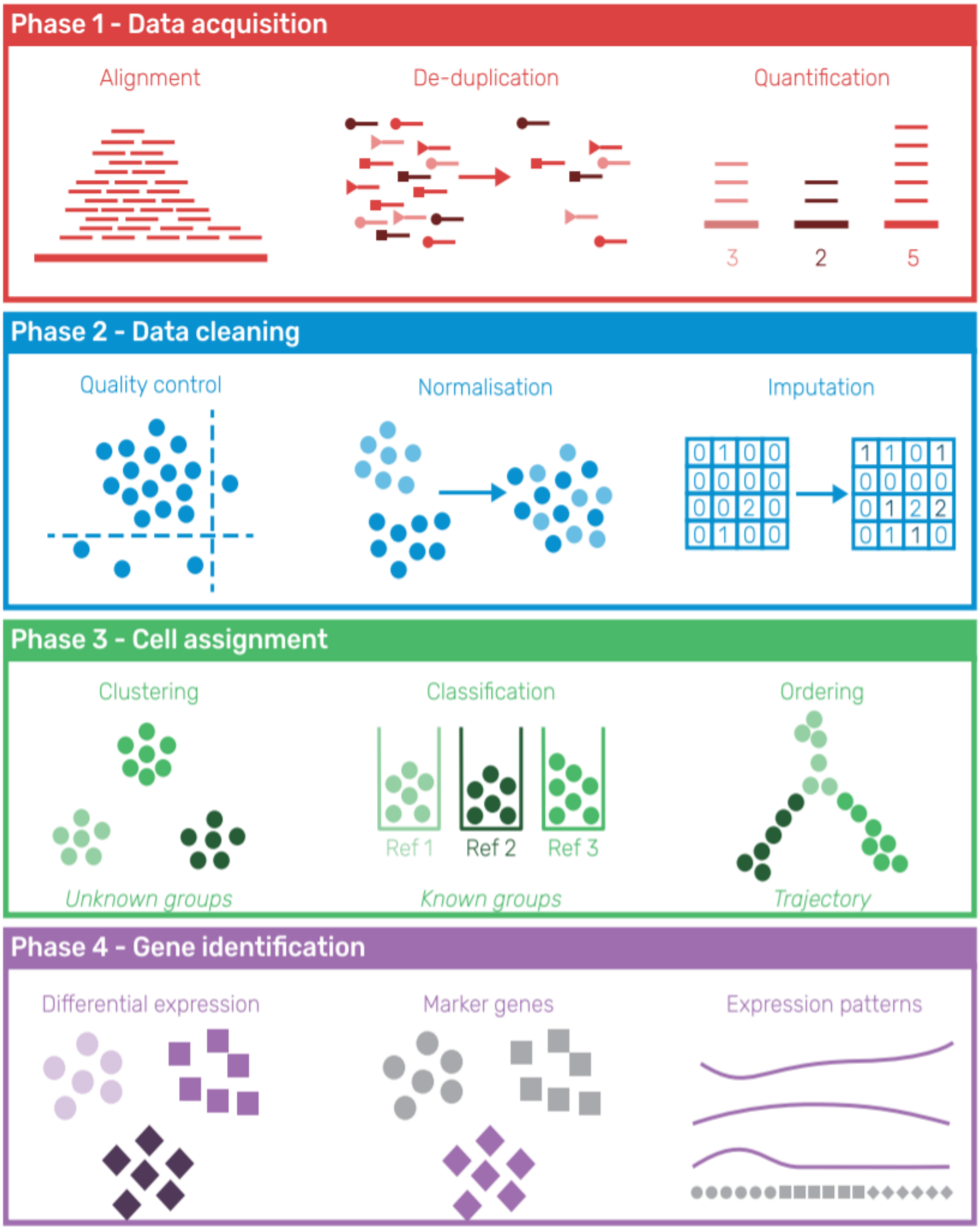
Next the cells are divided and assigned to groups, which may previously have been unknown, by clustering. The grouping of cells is usually based on their expression profile. With the development of the studies on scRNA-seq the assignment of the cells to a cell type becomes possible, i.e. they begin to be classified. It is also possible to order the cells according to their development, thus showing the kinetics of the transcript. At this stage, the identification of new cell population or of cells with stem-like characteristics takes place.

Once cells have been assigned, the focus of analysis turns to interpreting what those assignments mean. The fourth phase, gene identification, is used, as the name suggests, to identify the gene of interest for the study. At this point of the analysis it becomes possible to define which are the genes expressed differently along the different groups, the genes that change their expression over time or even the marker genes of a defined group.

If it is necessary to evaluate a certain hypothesis where cell populations have to be ordered or if the differences between cell types are not of interest, but those of the conditions of development of the experiments are, this path of analysis may be different. In fact, the phase of allocation of the cells may not be necessary and different approaches or tools are needed.

Another task that is usually included on the tools is the visualisation in which scRNA-seq data and results are explored and displayed. Dimensionality reduction is also a commonly used for visualisation or quality control usually carried out with t-SNE. Simulation, i.e. creation of synthetic scRNA-seq datasets, is another valuable technique which has been recently developed to test and validate scRNA-seq tools. More tools are now including packages with simulation functions and some are even developed for the purpose of generating realistic synthetic scRNA-seq datasets.

Considering the tools used for computational data analysis, it can be noticed that they can be constructed to carry out a single task or, as for example Seurat, they can be seen as analysis toolboxes that follow almost all the phases of the analysis, from the creation of the expression matrix up to the identification of the genes. The creation of tools that perform multiple functions arose from the need to simplify the analysis. In fact, using a single tool can avoid data loss resulting from the conversion of data between different formats.



0	1	0	0
0	0	0	0
0	0	2	0
0	1	0	0

1	1	0	1
0	0	0	0
0	1	2	2
0	1	1	0

Figure 4. Description of the possible phases of scRNA-seq analysis. In data acquisition (Phase 1) an expression matrix (gene x cells) is created from the sequencing reads. The reads are aligned to the reference genome or to the transcriptome and their expression is quantified. Some protocols require assignment and de-duplication of Unique Molecular Identifiers (UMIs). In data cleaning (Phase 2), quality control is carried out with removal of low-quality cells and uninformative genes. At this stage several protocols also perform normalization end estimation of expression where zeros have been observed. In class assignment (Phase 3) the cells are assigned, by classification or clustering, respectively in known or unknown groups or in a position along a developmental trajectory. In Phase 4 the gene are identified to explain the groups or trajectories. (Zappia et al., 2018)

2. Detection of doublets in scRNA-seq

Single-cell genomics has brought great advantage in the study of cell population heterogeneity and in the analysis of gene expressions. However, this technology is still limited by several issues, one of which is the presence of doublets. Doublets occur when two cells are captured together at the reaction sites and labelled with the same barcode. Sequenced as the remaining single cells, doublets result in expression profiles that are the mixture of the profiles of the two captured cells. Even though doublets generate more complex read sequences than the individual cells, as they may have a higher number of detected transcripts, the signal strength is not enough to identify them in an unambiguous manner. Doublets can bring serious consequences in the interpretation of the data resulting from scRNA-seq experiments, since their mixed transcriptome can be interpreted as cell types or states that do not exist. For this reason, doublets should be identified and discarded from the analysis.

Experiments performed with a mixture of different species can be used to estimate a lower bound on doublet rate (Bloom, 2018), i.e. the ratio between doublets and total number of sequenced cells. For the 10x Genomics Chromium technology, the doublet rate has been shown to follow a distribution of Poisson (Bloom, 2018), resulting in a linear relationship between total cells and doublet rate (Figure 5). Considering a dataset in which are present two cell types with equal proportion, the multiplets frequency M, that is the probability that a non-empty droplet contains more than a cell, can be defined as:

$$M = \frac{Pr\ (c \geq 2)}{Pr\ (c \geq 1)} = 1 - \frac{(\mu_1 + \mu_2) e^{-\mu_1 + \mu_2}}{1 - e^{-\mu_1 - \mu_2}} \qquad (1)$$

where μ_1 is the average number of type 1 cells per droplet and μ_2 is the average number of type 2 cells per droplet, while $\Pr\ (c \geq 1)$ and $\Pr\ (c \geq 2)$ are the probabilities that the number of cells in a droplet is larger than one or two, respectively.

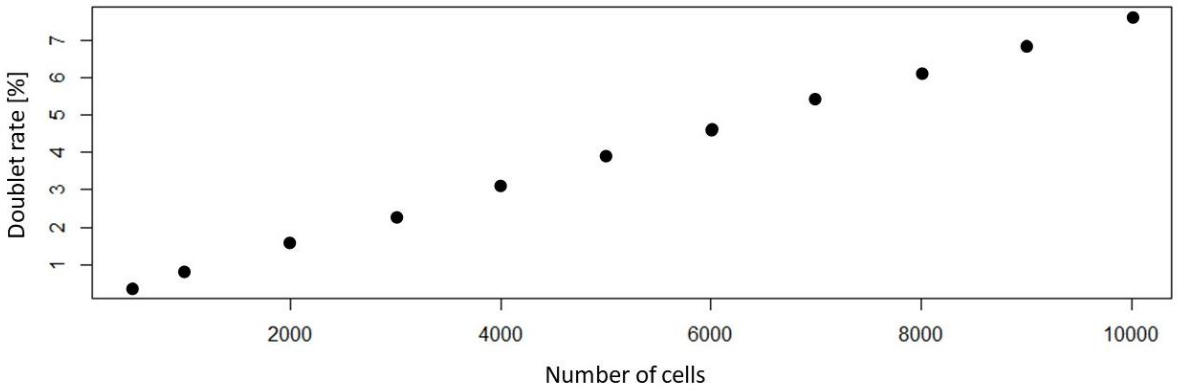


Figure 4. Scatterplot of doublet rate for the number of cells recovered as described on Chromium™ Single Cell 3' Reagent Kits v2 User Guide.

However, the mere definition of a doublet rate does not allow to identify the doublets that, leading to possible errors of analysis, should be excluded from the studies. Different tools are emerging with the purpose of identify doublets, either with experimental and/or computational means (Table 1).

Demuxlet (Kang et al., 2018) and Cell Hashing (Stoeckius et al., 2018) are two promising approaches to generate scRNA-seq data and effectively remove doublets in the downstream analysis by considering additional information from individual genetic variation or cellular barcoding, respectively. One important limitation of these methods is that they cannot be applied to different experimental systems and technologies, as well as to publicly available scRNA-seq data, but they require the generation of novel data using their specific library preparations and setup. Demuxlet requires genetically distinct samples and Cell Hashing requires unique antibody-oligonucleotide conjugate panels for the cell types of interest.

Table 1. Overview of experimental and computational tools for doublet detection.

References	Tool	Method	Performs doublet/singlet classification	Analysis of standard scRNA-seq data	Input data	Programming language	Code Availability
(Kang et al., 2018)	Demuxlet	Maximum likelihood	yes	no	---	ANSI C	https://github.com/statgen/demuxlet
(Stoeckius et al., 2018)	Cell Hashing	Barcoded antibody signal	yes	no	---	---	---
(DePasquale et al., 2018)	DoubletDecon	Deconvolution based strategy	yes	yes	counts	R	https://github.com/EDePasquale/DoubletDecon
(McGinnis, Murrow, & Gartner, 2019)	DoubletFinder	Nearest neighbours search	yes	yes	UMI counts	R	https://github.com/chris-mcginnings-ucsf/DoubletFinder
(Wolock, Lopez, & Klein, 2019)	Scrublet	Target-decoy search	yes	yes	UMI counts	Python	https://github.com/AllonKleinLab/Scrublet
---	doubletCells	Simulation doublet	no	yes	counts	R	Scran
---	doubletCluster	Cluster identification	no	yes	counts	R	Scran
---	DoubletDetection	cell augmented dataset clustering	yes	yes	counts	Python	https://github.com/JonathanShor/DoubletDetection

When scRNA-seq data are generated with the standard technologies instead, doublets must be identified and removed during the bioinformatic analysis by considering only the generated expression profiles. One simple approach consists in the identification of cells characterised by several detected genes (namely, genes with non-null expression) higher than a certain threshold. This heuristic procedure might produce false positives in the presence of cells with higher overall RNA abundance, which are also characterised by a higher number of detected genes and can be biased by the technical variability in mRNA capture efficiency.

More computational methods have been recently proposed for the identification of doublets using more complex, integrative analysis of the data, among which: DoubletDecon (DePasquale et al., 2018), DoubletFinder (McGinnis et al., 2019), Scrublet (Wolock et al., 2019), while the experimental ones are Demuxlet (Kang et al., 2018) and Cell Hashing (Stoeckius et al., 2018) and DoubletDetection. In parallel, methods such doubletCluster and doubletCells, despite not performing a full classification of singlets and doublets, permits the identification of possible clusters of doublet cells or to compute a cell-specific doublet score representing the likelihood that a cell is a doublet.

2.1 Demuxlet

Demuxlet (Kang et al., 2018) is a method for multiplex droplet scRNA-seq (dscRNA-seq) that sequences thousands of cells from different individuals to better identify inter-individual variability. For this reason, it uses an experimental protocol and a computational algorithm that exploits genetic variation between individuals to identify droplet containing one cell (singlet) or two cells (doublet) from different people and their genetic identity. It considers genotype probabilities to assess the likelihood of having RNA-seq reads overlapping a group of single nucleotide polymorphisms (SNPs) from a single cell. This represents a statistical model that helps to find to which individual each cell belongs. Even if only a small number of reads overlap common SNPs it is enough to identify each cell.

It is possible to note that multiplexing, i.e. capturing multiple cells and mRNAs simultaneously, even only a small number of individuals, allows a very high probability
(1 – 1/N) that a doublet is composed of cells from different individuals.

Demuxlet carries out doublets identification by simulating droplets and by analysing real data. The simulated doublets are created by randomly sampling and merging pairs of barcodes within a dataset. The real data are taken from the expression profiles resulting from the 10x Chromium instrument and what are considered for the study are the UMI counts (for the study of inter-individual variability they are usually normalised).

In the implementation of this method, the first step is the identification of the sample identity of each single cell considering that doublets are absent. The maximum likelihood is used to determine best-matching sample. The second step focuses on the identification of the doublets and implements a mixture model that permits the calculation of the likelihood that the sequence reads are from two individuals. The calculated likelihoods are then compared to decide whether a droplet contains cells from one or two samples. The next step is the study of theoretical expectation of deconvoluting singlets. The multiplet rate can be calculated as:

$$[(n-(n-1)\beta)/n]d(x)$$

where d(x) is the multiplet rate expected when x cells are loaded, and α and β represent the fraction of true singlets classified as multiplets (false positives) and the fraction of multiplets correctly classified (true positives), respectively.

In conclusion, Demuxlet can detect multiplets originating from two individuals reducing the number of multiplets that are non-identifiable in a rate that is directly proportional to the number of multiplexed samples.

2.2 Cell Hashing

Cell Hashing (Stoeckius et al., 2018) is another method able to reduce the number of doublets in single-cell data, which is based on an ad hoc experimental procedure, followed by computational analysis to remove the identified doublets. It uses oligo-tagged antibodies against proteins, expressed on the cell surface, to mark cells from distinct samples. Therefore, is possible to pool these cells together and to use the barcoded antibody signal as a fingerprint for demultiplexing. In fact, by sequencing them together with the cellular transcriptome, it is possible to find the sample of origin of each cell and identify doublets forming from multiple samples.

The strategy is a modification of the CITE-seq (Stoeckius et al., 2017) method, where oligonucleotide-tagged antibodies are used on cell-surface proteins to create a sequenceable read-out together with scRNA-seq. In Cell Hashing approach a set of monoclonal antibodies is used on ubiquitously and highly expressed immune surface markers. The antibodies are combined into eight identical pools and each pool is then labelled with distinct hashtag oligonucleotide (HTO). The cells are then sequenced on the 10x Genomics Chromium instrument to generate counts and UMI counts.

Subsequently, to examine two by two the expression of HTO counts, a statistical model is used to classify the barcode for each HTO as ‘positive’ or ‘negative’. Background cells are predicted from the k-medoids clustering of all HTO reads and the background signal for each HTO is modelled independently as a negative binomial distribution. Barcodes that have HTO signals greater than 99% quantile for this distribution are marked as 'positive', and barcodes that are 'positive' for more than one HTO are considered as multiplets.

Cell Hashing can assign each barcode to its original sample and is able to detect cross-sample multiplets. However, it cannot identify doublet from a single sample.

2.3 DoubletDecon

DoubletDecon is a computational method for doublet identification that can be used to analyse scRNA-seq data from different platforms and can account for confounding effects like cell-cycle bias.

DoubletDecon imports single-cell expression profiles, summarised as counts, from the outputs of the ICGS (AltAnalyze - https://github.com/nsalomonis/altanalyze/wiki/cellHarmony) or Seurat (https://github.com/satijalab/seurat). After data pre-processing, counts can be corrected for cell-cycle bias. Then, gene expression medoids or centroids are computed for each cluster (to be specified by the user) and used to determine intra-cluster similarity. Cluster medoids are preferred because they are not affected by doublets contained in a cluster, but centroids can be advantageous in case of very sparse data.

A binary correlation matrix, called ‘blacklist’, is created from medoid correlations (R) that are used to determine cluster similarity. Next, a threshold for cluster similarity is defined as

$$\rho = mean(R) + \rho' \cdot sd(R) \tag{2}$$

where ρ’ is the blacklist correlation threshold, i.e. the level of similarity that a medoid should have to be considered correlated. The blacklisted clusters are the ones for which the medoid correlation surpass the threshold ρ. They represent clusters with the higher level of dissimilarity, so ideally not containing doublets. Then higher values of ρ’ result in fewer blacklisted clusters. Markov clustering is then used to identify new clusters while minimising inter-cluster similarity and determine which clusters are not useful for doublets detection. In this process, new medoids are created based on the blacklisted clusters and the cluster identification is updated.

To determine whether a cell profile is more similar to an individual cell or to a doublet, thirty synthetic doublets are generated through the combination of pairs of cells sampled from all dissimilar clusters. The number of simulated doublets can be set by the user.

The core algorithm of DoubletDecon consists in three main steps that follow and are called: Remove, Re-cluster and Rescue. The first step removes putative doublets, whereas the other two aim at identifying and rescue possible false positives due to transitional cell states. A detailed description of the three steps here follows.

During the “Remove” step, a deconvolution approach based on quadratic programming and implemented the R package ‘DeconRNASeq’ (Gong & Szustakowski, 2013) is applied to each expression profile, including the synthetic doublets. The blacklisted cluster medoids are used as reference profiles for deconvolution. For each cell, the deconvoluted cell-profiles (DCP) are then compared with Pearson correlation to the centroid DCP for cells in each blacklisted cluster and the centroid synthetic doublet DCP. The cells that present DCP most similar to DCP from a simulated doublet are then considered as doublets.

In the “Re-cluster” step, the cells that were removed in the previous step, because identified as doublets, are re-clustered. Then, in the “Rescue” step, ANOVA analysis is used to evaluate unique gene expression in the new doublet clusters. Initial clustered doublets that present at least one unique gene expressed relative to the original blacklisted clusters are re-assigned as singlets and reincorporated into the non-doublet expression matrix.

2.4 DoubletFinder

DoubletFinder is a computational method for the identification of doublets based solely on gene expression profiles. DoubletFinder generates artificial doublets to identify real doublets through neighbourhood search, assuming that artificial and real doublets are likely to cluster together in the gene expression space.

The method consists in a multi-step strategy (Figure 6):

1. Generation of artificial doublets by averaging raw UMI count profiles from pairs of cells selected through random sampling without replacement.
2. Merging of artificial and real data.
3. Dimension reduction using principal component analysis (PCA)
4. Computation of the proportion of artificial nearest neighbours (*pANN*) for every cell in the gene expression space.
5. Ranking and thresholding of *pANN* values according to the number of expected doublets *nExp*.

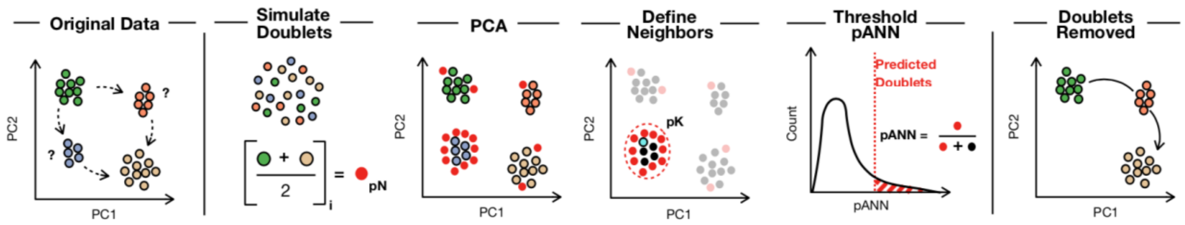


Figure 5. DoubletFinder analysis steps. The first step is the creation of artificial doublets (red) from the original data and their integration at a defined proportion (pN). Next DoubletFinder uses principal component analysis (PCA) for dimension reduction and defines each cell's neighbourhood in gene expression space. The proportion of artificial nearest neighbours (pANN) is then calculated and the cells with the top pANN values are predicted as doublets. Figure from (McGinnis et al., 2019).

The number of expected doublets can be estimated from cell loading densities into the 10X/Drop-Seq device using Poisson statistics. However, DoubletFinder can identify heterotypic doublets (i.e., doublets formed from transcriptionally distinct cell states) but not homotypic doublets (i.e., doublets formed from transcriptionally similar cell states). Therefore, a Poisson-derived estimate *nExpP* can result in false-positive calls. If cell-type annotations are available, this information can be used to estimate the proportion of homotypic doublets *HoD* and correct the number of expected doublets as:

$$nExp \;= nExpP \cdot (1 - HoD) \tag{3}$$

However, it must be pointed out that the estimation of *nExp* is not trivial, especially when analysing public datasets, for which the following information are rarely reported: (1) cell loading densities, important to derive the total number of expected doublets; (2) pools of cells sequenced in the same lane, which should be analysed in batches, as cells from different lanes do not generate doubles; and (3) cell-type annotations.

2.5 Scrublet

Scrublet is another computational method that identifies doublets by comparing cell expression profiles to simulated doublets.

It develops over two principal steps. First, doublets are simulated through the random combination of pairs of observed transcriptomes. Then, the observed transcriptomes are scored using the relative densities of simulated doublets and observed transcriptomes in their vicinity. Thus, also this method, focuses on the identification of heterotypic rather than homotypic doublets.

Scrublet is based on three assumptions: (1) the gene expression space is high-dimensional and sparsely populated by cells, implying that heterotypic doublets will likely fall into an unoccupied region of gene expression space; (2) among all observed transcriptomes, multiplets are relatively rare events, justifying the study of doublets rather than multiplets; (3) all cells used to generate simulated doublets are also present as single cells somewhere in the data.

Under these assumptions, Scrublet identifies heterotypic doublets with the procedure described in the following. First, it generates simulated doublets through the linear combination of pairs of observed cell transcriptomes randomly sampled. Then, it merges the real transcriptomes and simulated doublet profiles. For each real transcriptome i or simulated doublet i' , the scores $\boldsymbol{f_i}$, $\boldsymbol{f}_{i'}$ are defined as the ratio of simulated doublets to real transcriptomes in the neighbourhood of i or i'. A doublet score threshold θ is fixed based on the bimodal distribution of $\boldsymbol{f}_{i'}$. A bimodal distribution of $\boldsymbol{f}_{i'}$ is used because heterotypic doublets tend to have a higher fraction of simulated doublets neighbours than single cells or homotypic doublets. The simulated doublets that have $\boldsymbol{f}_{i'} < \theta$ correspond to homotypic doublets, while those with $\boldsymbol{f}_{i'} > \theta$ to heterotypic doublets. Then, Scrublet calculates the "detectable doublet fraction", φ_D, computed as the fraction of simulated doublets with $\boldsymbol{f}_{i'} > \theta$.

φ_D is used to estimate the fraction of observed doublets that are heterotypic doublets. Finally, the real transcriptomes with $f_i > \theta$ are classified as putative heterotypic doublets.

2.6 DoubletCells

DoubletCells is an R function from the "scran" R package (Huber et al., 2015). It is based on the simulation of doublets from single cells to estimate each cell probability of being a doublet.

DoubletCells first simulates doublets by randomly merging pairs of single-cell profiles. For each real transcriptome, then, it computes the cell the density of simulated doublets and real cells in the neighbourhood is calculated. Finally, for each cell, returns a doublet score computed as the ratio between the two densities. However, it does not perform a full classification of singlets and doublets.

2.7 DoubletCluster

DoubletCluster is another R function from the "scran" package and permits the identification of clusters containing putative doublets.

It is based on the identification of clusters that have expression profiles in between two other clusters using an approach similar to that used in (Bach et al., 2017).

It analyses all possible triplet of clusters considering two clusters as 'sources' and the last one as the 'query' cluster. The basic hypothesis is that the query is formed by doublets born from the union of two cells from the sources. As such, gene expression in the query cluster should be intermediate between the two sources.

DoubletCluster identifies the possible doublet clusters with three characteristics. The first one is the number N of genes in the query cluster that are unique. If N indicates the number of genes that are not present in the cluster ‘sources’, lower number of N represent that the query cluster is likely to contain doublets. The second characteristic is the size of the library. Doublets should have library size higher than single cells. This way it is possible to assume that the query cluster contains doublets if its library size is greater than that of the other two clusters. This leads to the assumption that the doublet cluster contains more RNA and has more counts than either of the two source clusters. The last characteristic to consider is the fact that the proportion of cells in the query cluster should also be reasonable. It is reasonable to have few cells in the query cluster since if the scRNA-seq experiment is properly performed, is desirable to have few doublets.

3. Material and Methods

3.1 Datasets

3.1.1 Cell Hashing datasets

Cell Hashing (Stoeckius et al., 2018) dataset was obtained by pooling eight samples of human peripheral blood mononuclear cells (PBMC) from healthy donors and combining these cells with a set of oligonucleotide-tagged antibodies. The antibodies were created by the union of monoclonal antibodies, directed against immune surface markers, and hashtag oligonucleotides (called HTOs). The HTOs are formed by unique barcodes that can be read and linked to the cellular transcriptome. These data were loaded in a droplet-based instrument, 10x Chromium, and analysed. The dataset is available on Gene Expression Omnibus (accession: GSE108313) and can be used to benchmark computational methods for doublets detection as it contains both the cell expression profiles and single/doublet classification, which can be used as a gold standard measure.

We considered a subset of the sequenced cells, for which a truth table (singlet/doublet) was available from (McGinnis et al., 2019). The single-cell expression data considered consisted in a count matrix composed of 15,646 cells and 40,899 genes. To limit the computational burden of the analysis, to be run on a personal computer, and to test the dependence of method performance on sample size, we derived from this data three sets of subsampled datasets. Each set was composed by 100 datasets with total number of cells reported in Table 2.

Table 2. Simulations created with Cell Hashing data with relative number of datasets and number of cells considered for each dataset. An ID has been assigned to each simulation.

ID Simulation	**Datasets**	**Number of cells**
CellHash1	100	1000
CellHash2	100	2000
CelHash3	100	3000

3.1.2 Simulated datasets

To further evaluate the methods on benchmark dataset, we constructed four sets of simulated datasets (with 100 scRNA-seq datasets each) starting from the real data generated in (Zheng et al., 2017). This dataset was derived from 68k PBMCs sequenced with a droplet-based system, the 10x Genomics Chromium, and they are available at https://support.10xgenomics.com/single-cell-gene-expression/datasets.

From this study, we considered six cells: B cells, CD4$^+$ T cells, monocytes, regulatory T (T_{reg}) cells, CD8$^+$ T cells, and natural killer (NK) cells. The data were pre-processed to select only cells with expression of cell-specific gene markers greater than 0, number of detected genes lower than the mean plus one standard deviation computed across all cells, to filter out possible doublets, and higher or equal to 300, di discard low-quality cell profiles (Table 3).

Table 3. Cell Types considered for the simulations with number of cells before and after the filtering and marker genes.

Cell Types	**Marker genes**	**Num. initial cells**	**Num. cells after filtering**
CD4$^+$ T cells	CD4	11213	681
CD8$^+$ T cells	CD8B	10209	5458
T_{reg} cells	FOXP3	10263	445
B cells	CD19	10085	1553
Monocytes	CD14	2612	1224
NK cells	NCAM1	8385	256

After this filtering all the data left were merged in a single matrix where the columns correspond to the filtered cells of the different datasets, while the rows are the common genes considered in the different datasets. A vector was also created to memorise the cell types for each cells (column) in the matrix. The simulated datasets were then created in an iterative way so that for each number nSim sets, the 100 datasets were created aggregating nDoubl doublet and nSing singlet cell profiles (Table 4). The nSing cells were selected randomly and without replacement and the data from these cells were saved in the matrix X. 2*nDoubl cells were selected randomly and without replacement and the expression sum from each pair of cells was calculated. The data were saved in the Matrix Y. The final matrix for each dataset was then created by combining the X and Y matrices. In Sim1, Sim3, and Sim4, the percentage of doublets was set to increase with decreasing of the single cells , whereas in Sim2 nDoubl was fixed as described in (Wang et al., 2019).

Table 4. Simulations created with (Zheng et al., 2017) *data with relative number of datasets (nSim), number of singlets wanted for datasets (nSing), number of doublets considered for each dataset (nDoubl) and doublet rates (nDoubl over total cells).*

ID simulation	**nSim**	**nDoubl**	**nSing**	**Doublet rate (%)**
Sim1	100	100	3000	3.23
Sim2	100	630	2370	21
Sim3	100	100	2000	4.76
Sim4	100	150	1950	7.14

3.1.3 Case study: scRNA-seq data from the tumour microenvironment

As a case study, we considered real scRNA-seq dataset described in (Jerby-Arnon et al., 2018). The data were obtained from of 33 human melanoma tumours from 31 patients with scRNA-seq analysis. 7,186 cell profiles have been obtained of which 4,199 are cells collected from patient tumours of a previous study (Tirosh et al., 2016) and 2,987 from new patient tumours. From both the newly collected and (Tirosh et al., 2016) melanoma, individual cells were dissociated from fresh tumour tissues, immune and non-immune cells were isolated by FACS and profiled with a modified full length plate-based SMART-Seq2 protocol. In the original publication, cell profiles were assigned to specific cell types considering the expression of marker genes. From the analysis several subsets of cells were derived based on their expression profiles and the following cell types were identified: tumour cells, CD8$^+$ T cells, CD4$^+$ T cells, B cells, NK cells, macrophages, cancer associated fibroblasts (CAFs), endothelial cells, and “unknown” cells. The dataset is available on Gene Expression Omnibus (accession: GSE115978), and it includes the count matrix with the expression profiles and the cell type annotations.

3.2 Data preprocessing: Seurat Pipeline

Seurat is an R package that aims to improve scRNA-seq data analysis and to help the identification of highly expressed transcriptome. We used this package for the preparation of the data for the subsequent analysis.

From the data provided by the analysis of the single cell measurements, the function CreateSeuratObject is used to create a Seurat object. After the creation of the Seurat object, this tool permits the use of different functions to pre-process the scRNA-seq data.

The first pre-processing process applied on the data (in the Seurat object) was the normalisation with the function NormalizeData so that is possible to limit the redundancy of the data and to level the data from the different cells.

After the normalisation since potential confounding factors, such as the cell cycle, can be found in the data we have run the Seurat function named ScaleData that permits to scale and centre the genes in the data and remove unwanted sources of variation

Another necessary step in our pre-processing was the detection of highly variable genes across the single cells since these are the genes of interest for the analysis. The function FindVariableGenes was then used on the scaled data to identify genes that are outliers while controlling for the strong relationship between variability and average expression. Its first step is to calculate the average expression and the dispersion for each gene and then, secondly, to divide genes into bins based on their average expression, and to calculate the z-scores for dispersion within each bin.

Next the PCA was performed on the scaled data using the Seurat function RunPCA, to improve the efficiency of the analysis. The Principal components analysis is used to reduce the dimensionality of the data and to project the cells into the first two principal components (PCs), to represent each cell as a point in a two-dimension space. It was applied on a Seurat object and was run with a parameter that contains the genes to use as input for the PCA.

Another Seurat function used is JackStraw that permits to determine statistically significant principal components permuting, in a random way, a subset of data and assessing projected PCA scores for these 'random' genes. Next it compares the founded scores with the observed PCA scores to determine statistical significance. In the end it gives a p-value for each gene's association with each principal component. This help to identify as ‘significant’ PCs those who are strongly enriched by low p-value features. In the running of this function it is important to specify the number of replicate samplings to perform.

We had also used the non-linear dimensional reduction t-distributed stochastic neighbour embedding (tSNE) as a tool to visualise and explore the datasets. For this we have run the function RunTSNE that is implemented using the tSNE idea of “transporting” cells that have similar local neighbourhoods in a space of high dimension, all together to a space that is dimensionally reduced.

After that analysis we included a graph-based clustering approach based on the expression data, founded with the function FindVariableGenes, which identified clusters of cells using a shared nearest neighbour (SNN) modularity optimisation-based clustering algorithm. The function FindClusters calculates this process by computing first the k-nearest neighbours and building the SNN graph and then optimising the modularity function to determine clusters

3.3 Doublet detection

The methods considered in this study are two computational methods from Table 1: DoubletDecon and DoubletFinder. In addition, we create a new method called Consensus that is created on the agreement in doublets classification of the other two methods.

The choice fell on computational methods and not experimental, because the formers are applicable to any dataset, which make real classification, and can process a large number of data in less time. The choice of DoubletDecon and DoubletFinder was made because they are both based on the R language, thus allowing a unified analysis in the same environment.

3.3.1 DoubletDecon R package

DoubletDecon has a principal prediction function for doublet identification, based on deconvolution-based strategy, called “Main_Doublet_Decon” and takes different parameters as input. We considered for our analysis only some: *rawFile, groupsFile, filename, location, species, rhop, PMF, centroids, num_doubs* and *min_uniq.*

rawDataFile represents the name of the file in ICGS format containing the counts gene by cell and *groupsFile* indicates the name of the file also in ICGS format where are saved the cluster assignments for each cell. This file contains a matrix in which the number of rows is equal to the number of cells and in the columns are saved the cell names and the cluster of each cell.

Filename is a character string that is pasted to the names of the output files from the function, while the parameter *location* represents the path to the directory where output files should be saved.

Species parameter indicates the scientific species name ("mmu", "hsa") to be considered when removing cell-cycle genes. We have use the label "hsa" since we have studies only human datasets.

rhop, instead, is the blacklist correlation threshold ρ' of Equation (2). We used its default value 1.

PMF is a logical parameter that claims whether "unique gene expression" should be used as doublet-determination criterium (default: TRUE). Another logical parameter used is *centroids* which indicates whether centroids should be used as reference profiles for deconvolution instead of medoids. For our analysis centroids were not needed so we set this parameter to FALSE.

The last two parameters that were considered are *num_doubs*, the number of artificial doublets to be generated for each pair of clusters, usually 30, and *min_uniq*, the minimum number of unique genes required for a cluster to be rescued, set to 4.

rawDataFile and *groupsFile* files for "Main_Doublet_Decon" can be prepared with the "Seurat_Pre_Process" function within DoubletDecon, which uses t-tests to identify marker genes in each cluster. This function takes as input the normalised expression file from Seurat (*expressionFile*), the gene list from Seurat (*genesFile*) and the cluster list from Seurat (*clustersFile)*.

3.3.2 DoubletFinder R package

DoubletFinder R package is characterised by a core doublet prediction function for doublet identification, based on neighbourhood search, called "doubletFinder". In the analysis we considered as input for the function: a Seurat object fully processed with the Seurat's functions; *pK* which represents the PC neighbourhood size used to compute pANN, expressed as a proportion of the merged real-artificial data, and *nExp* that is the number expected doublets. No default is set for the pK, as it should be adjusted for each scRNA-seq dataset.

The optimal *pK* value can be estimated in four steps. The first step is the computation of pN-pK parameter sweeps on a 10,000-cell subset of a pre-processed Seurat object using the “doubletFinder_ParamSweep” function. *pN* represents the number of artificial doublets to be generated, expressed as a proportion of the merged real-artificial data. The output is a list containing the pANN vectors for every *pN* and *pK* combination, computed for both real and artificial doublets. The second step is the processing of the list generated at step one with the “summarizeSweep” function to compute the bimodality coefficient in the pN-pK parameter space. The third step is the processing of the data generated at the second step with the “find.pK” function to compute (and visualize) the mean-variance normalised bimodality coefficient (BCmvn) score for each *pK* value tested in the parameter sweep. The last step is the identification of the optimal *pK* that maximizes BCmvn.

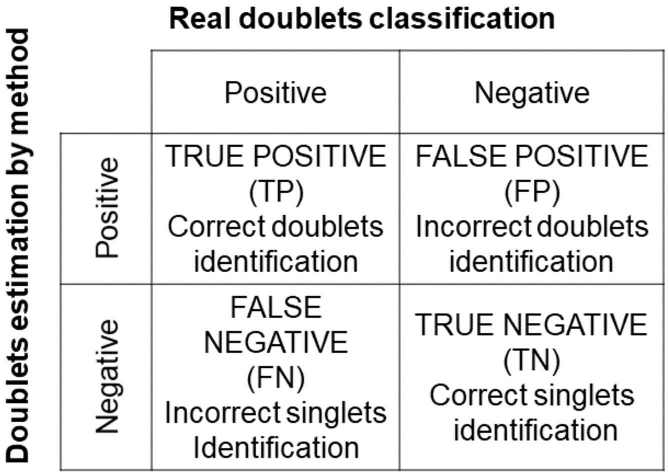
The identification of the optimal *nExp* parameter is not trivial, as discussed in Chapter 2.4. When cell annotations are available, the proportion of homotypic doublets (*HoP)* to be used for correcting Poisson-derived estimates of the expected number of doublets (Equation 3), can be estimated with the “modelHomotypic” function. For the purpose of this thesis, we chose to diverge from this procedure and to use as *nExp* the number of doublets detected by DoubletDecon method. This choice was made to compare the classification of the two methods by starting from the same conditions.

3.3.3 Consensus method

Beside testing DoubletFinder and DoubletDecon R packages, we explore the possibility to derive a consensus classification with improved accuracy. We implemented this approach, that we called Consensus, so that the final doublets were only the ones classified as doublets by both DoubletDecon and DoubleFinder. All the other cells were classified as singlets.

3.4 Performance metrics

To assess the classification performance of the three doublet detection methods, we considered four performance metrics based on the computation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) defined as shown in Figure 7Figure. In particular, TP were calculated as the number of doublets correctly identified by the method considered in the analysis. FP were calculated as the number of cells that were considered as singlets by the methods but are doublets for the real classification of the dataset. FN were calculated as the number of cells identified as doublets while being single cells. TN were computed as the number of singlets correctly estimated by the methods.



Real doublets classification

		Positive	Negative
Doublets estimation by method	Positive	TRUE POSITIVE (TP) Correct doublets identification	FALSE POSITIVE (FP) Incorrect doublets identification
	Negative	FALSE NEGATIVE (FN) Incorrect singlets Identification	TRUE NEGATIVE (TN) Correct singlets identification

Figure 7. Table of confusion that describes the performance of a methods for doublets detection on a dataset for which the real values are known.

These four parameters were then used to calculate the measure Accuracy, Precision, Sensitivity (also called Recall) and F1 Score.

The Accuracy is the ratio of doublets and singlets correctly identified in the total number of cells considered in the datasets. It can be calculated as:

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \tag{4}$$

Obtaining high accuracy usually means that the model gives good results but when the datasets are asymmetric, i.e. the values of FP and FN are very different, it is better to look at other measures to evaluate the performance of the methods. For this reason, Precision, Sensitivity and F1 Score are also used.

Precision is a parameter that indicates the ratio of doublets correctly identified compared to the total number of doublets estimated by the method and it is referred to as:

$$Precision = (TP)/(TP + FP) \tag{5}$$

Sensitivity represents the ratio of doublets correctly identified in the total number of doublets present in the datasets. It is calculated as:

$$Sensitivity = (TP)/(TP + FN) \tag{6}$$

The F1 score is a parameter that finds a balance between Precision and Sensitivity. It is the weighted average of the other the measurements. It is represented as:

$$F1\ Score = 2 \cdot (Sensitivity \cdot Precision)/(Sensitivity + Precision) \tag{7}$$

This measurement works better than Accuracy when dealing with datasets with uneven classification, as in the case of doublet/singlet identification.

Finally, for the simulated datasets, we also considered the percentage of homotypic doublets estimated by the methods compared to their real number in the dataset. Homotypic doublets in the simulated data were identified as simulated doublets generated from two profiles coming from the same cell type.

4. Results

4.1 Cell hashing datasets

Errore. L'origine riferimento non è stata trovata.-Figure 11 and Table 5 report the performance metrics obtained on the subsampled datasets derived from the Cell Hashing data (CellHash1, CellHash2, and CellHash3, respectively) and obtained by the three tested methods: DoubletDecon (DD), DoubletFinder (DF), and Consensus (Con).

Figure 8 is describes the performance of DoubletDecon in estimating the number of doublets. DoubletDecon performs a good estimation of the doublets number, except for ChellHash3 for which we have an overestimation. It is also shown that the relative error for each dataset could be large.

DoubletDecon has an average accuracy ranging between 76.18 and 76.49 (Table 5), which is slightly lower than that of the other methods. This method obtained the lowest precision on all datasets (ranging between 22.92 and 25.01), which is due to both the high number of FP and low number of TP obtained. This method also scored a higher number of FN with respect to DoubletFinder. DoubletDecon obtained the highest sensitivity in all datasets, due to the high performance in terms of TP (3.36-4.19) and FN (11.28-12.25) (Table 5). The Consensus approach scored the best performance in terms of accuracy, precision, and F1 score, but showed the lowest sensitivity due to the small number of TP and high number of FN obtained. Better TP, FN, and F1 scores could be obtained by all methods on datasets with higher total-cell sizes (CellHash1 < CellHash2 < CellHash3).

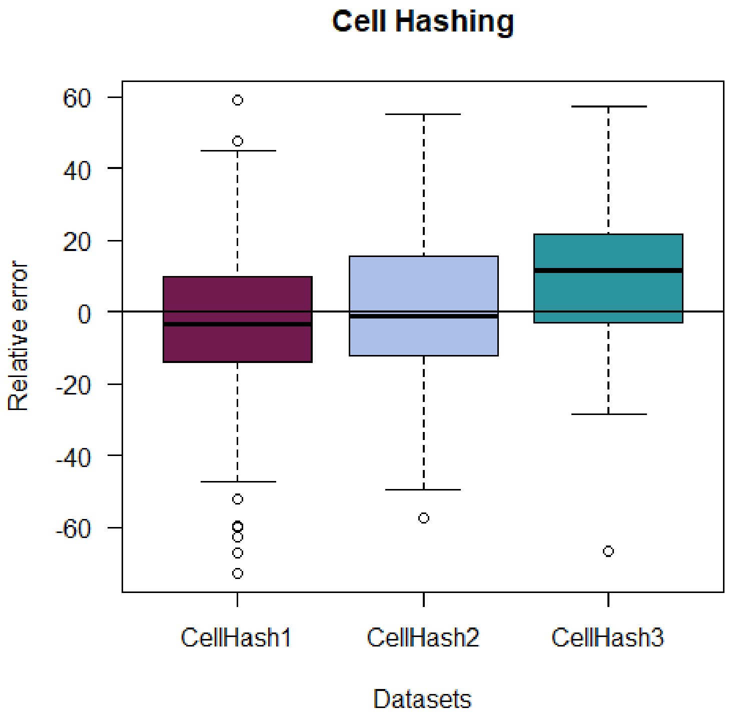


Figure 8. Boxplots of the doublet relative error of DoubetDecon method in the CellHash1, CellHash2 and Cellhash3 datasets.

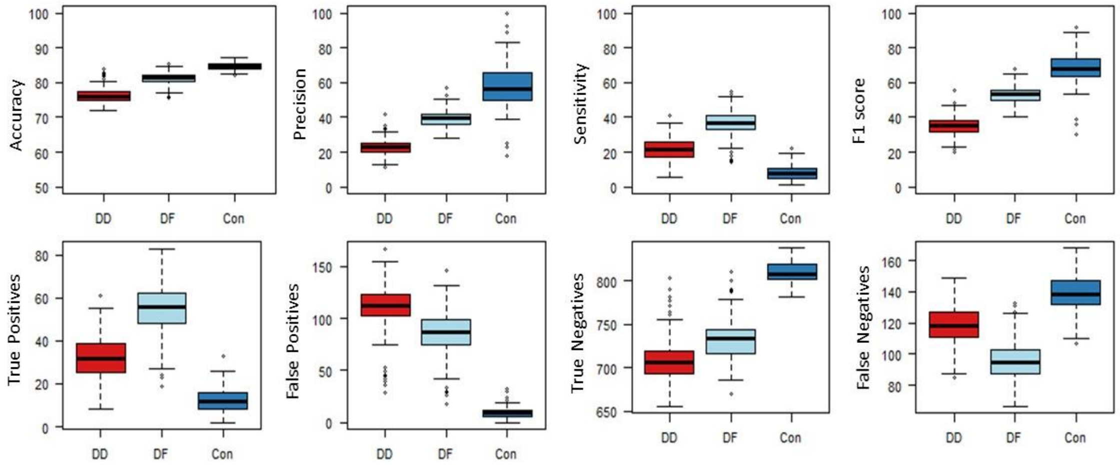


Figure 9. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the CellHash1 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives and false negatives

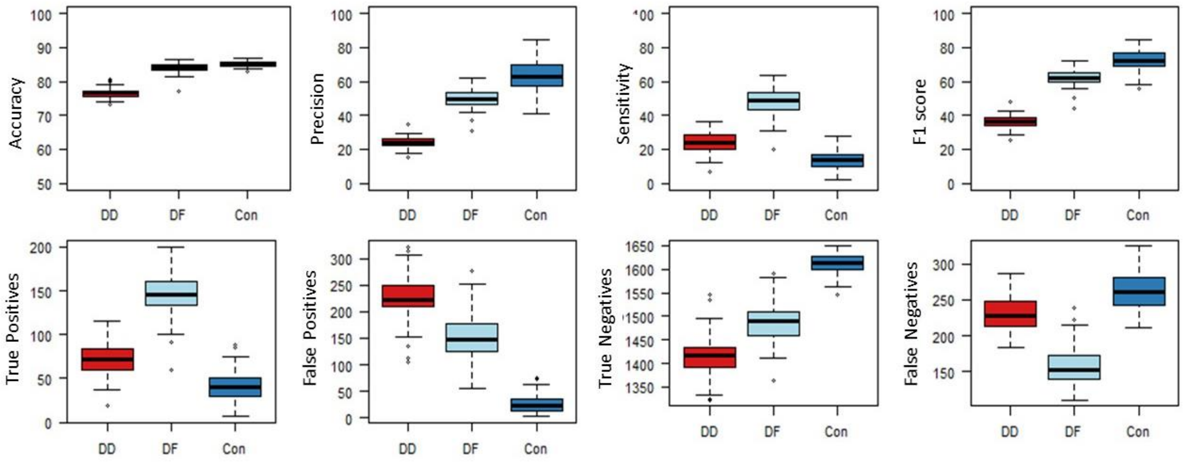


Figure 10. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the CellHash2 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives and false negatives

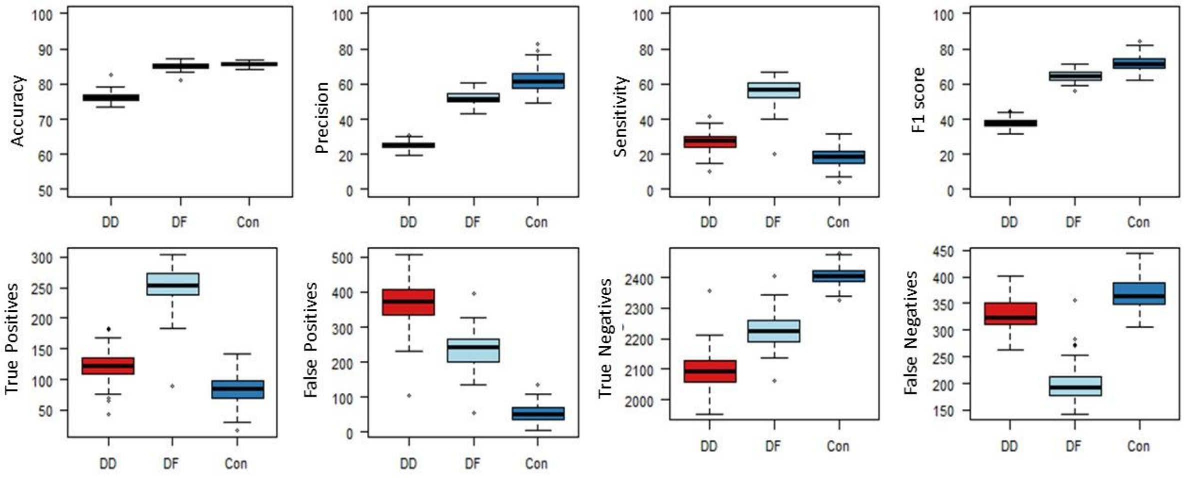


Figure 11. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the CellHash3 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives and false negatives.

Table 5. Average values of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) in the CellHash1, CellHAsh2 and CellHash3 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives and false negatives. The shaded cells identify the method that has the best performance for each metric.

Statistic	Simulation	DD	DF	Con
Accuracy	CellHash1	76.39	81.11	84.73
Accuracy	CellHash2	76.49	84.09	85.17
Accuracy	CellHash3	76.18	85.11	85.52
Precision	CellHash1	22.92	39.24	58.07
Precision	CellHash2	24.02	49.46	63.38
Precision	CellHash3	25.01	52.19	61.97
Sensitivity	CellHash1	21.56	36.26	8.34
Sensitivity	CellHash2	23.86	48.28	13.61
Sensitivity	CellHash3	27.08	56.04	18.46
F1 Score	CellHash1	35.03	52.70	68.05
F1 Score	CellHash2	36.45	62.18	72.31
F1 Score	CellHash3	37.60	64.65	71.69
TP	CellHash1	3.36	5.66	1.30
TP	CellHash2	3.71	7.50	2.11
TP	CellHash3	4.19	8.66	2.85
FP	CellHash1	11.36	8.94	0.97
FP	CellHash2	11.24	7.82	1.36
FP	CellHash3	12.54	8.08	1.87
FN	CellHash1	12.25	9.95	14.31
FN	CellHash2	11.87	8.09	13.48
FN	CellHash3	11.28	6.81	12.61
TN	CellHash1	73.03	75.45	83.43
TN	CellHash2	72.78	76.60	83.06
TN	CellHash3	72.00	76.46	82.66

4.2 Simulated datasets

Figure 14-Figure 17 and Table 6 report the performance metrics obtained on the simulated datasets (Sim1-Sim4) and obtained by the three tested methods: DoubletDecon, DoubletFinder and Consensus.

In Figure 12 is described the performance of DoubletDecon in estimating the number of doublets. As it is shown, DD tends to underestimate the number of doublets, except for Sim1, and the relative error for each dataset is large. By comparing Figure 12 and Figure 13Figure it is possible to notice that even if we calculate the relative error considering only the heterotypic doublets, DoubletDecon tends to underestimate the number of doublets. So, the problem on doublets estimation cannot solely be linked to the difficulty on homotypic doublets identification.

On this dataset, DoubletDecon accuracy, precision, and F1 score values were always lower than those obtained with the competitor methods, but generally high (>74 for F1 score), with the only exception of dataset Sim1. The fraction of homotypic doublets identified ranged in 3.81 and 11.25%. DoubletFinder obtained the highest accuracy on all datasets (except Sim1) and sensitivity. Moreover, it identified the highest number of homotypic doublets in all datasets. This fraction, however, did not exceed 33%. The Consensus method obtained the best results in term of precision (83.66-98.77) and F1 score (88.98-94.85), at the expenses of sensitivity (always lower than 20%) and fraction of homotypic doublets identified (1.18-5.56%). The stringency of this method is also evident from Figure 18, which shows the percentage of doublets identified by Consensus in the four simulated datasets, referred to total doublets (which are the same for DoubletDecon and DoubletFinder). This percentage represents the fraction of doublets in common between DoubletDecon and DoubletFinder, and ranges between 0 and 9%

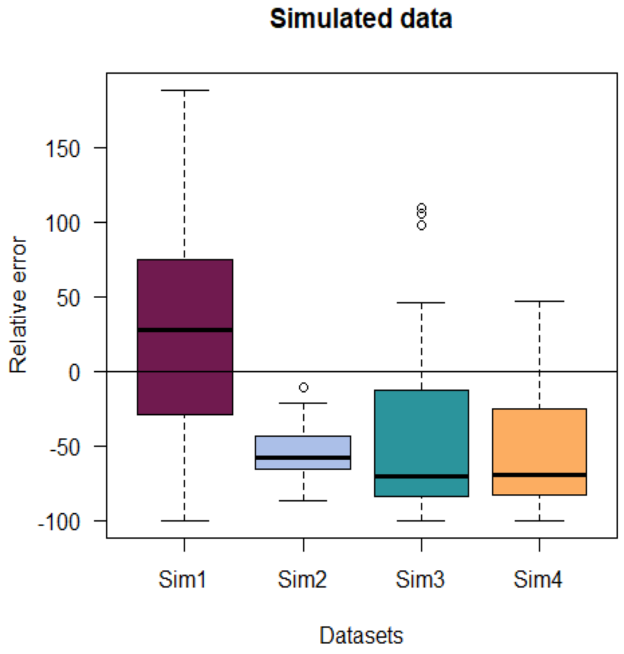


Figure 12.Boxplots of the doublet relative error of DoubetDecon method in the Sim1, Sim2, Sim3 and Sim4 datasets.

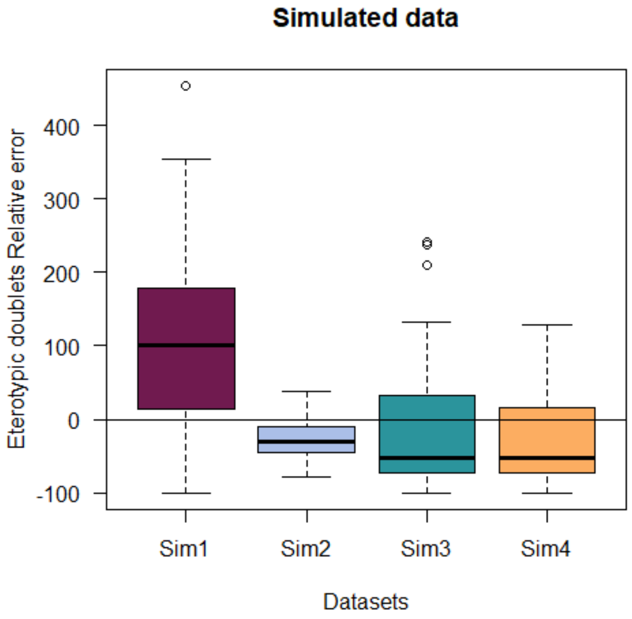


Figure 13. Boxplots of the eterotypic doublet relative error of DoubetDecon method in the Sim1, Sim2, Sim3 and Sim4 datasets.

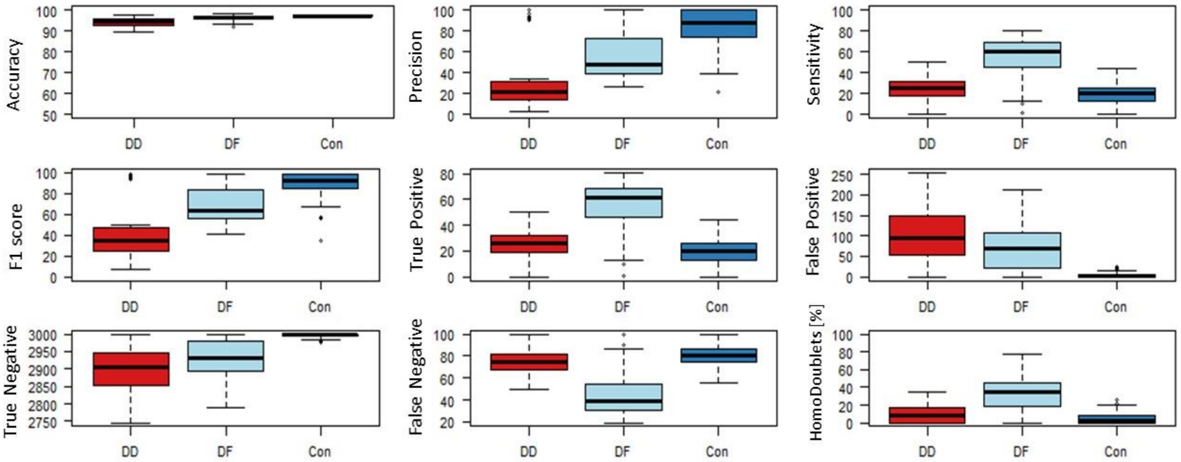


Figure 14. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the Sim1 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives, false negatives and percentage of homotypic doublets detected.

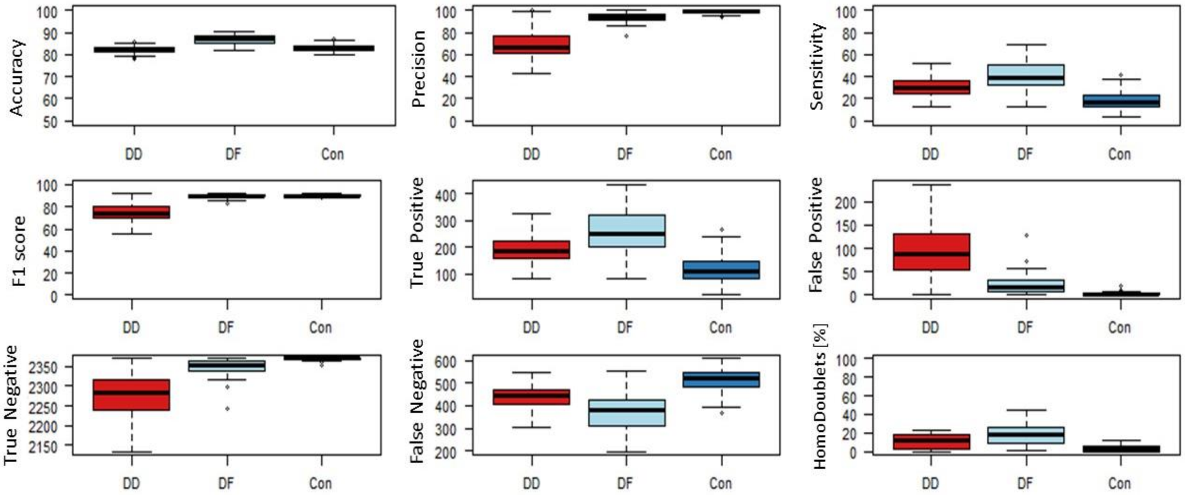


Figure 15. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on Sim2 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, false negatives and percentage of homotypic doublets detected

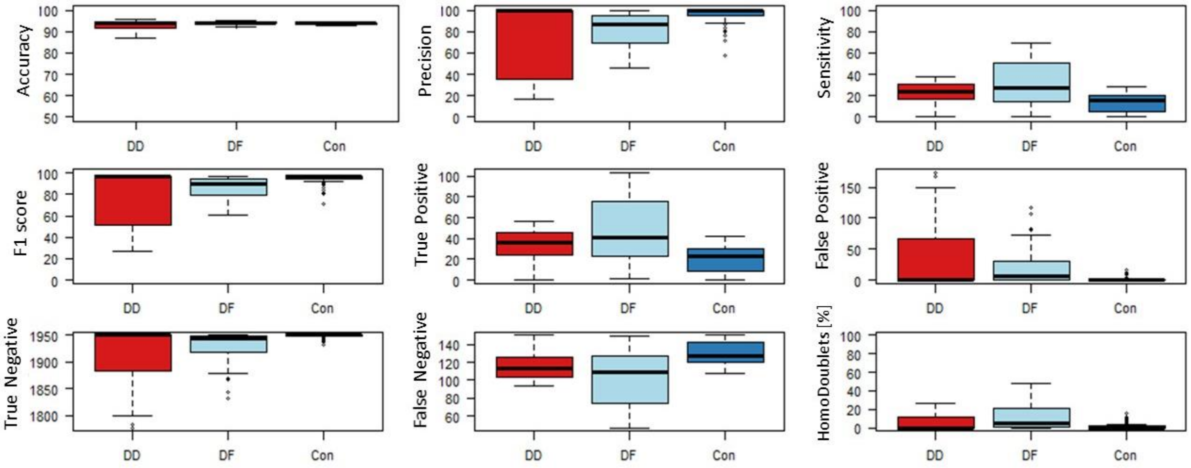


Figure 16. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the Sim4 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives, false negatives and percentage of homotypic doublets detected.

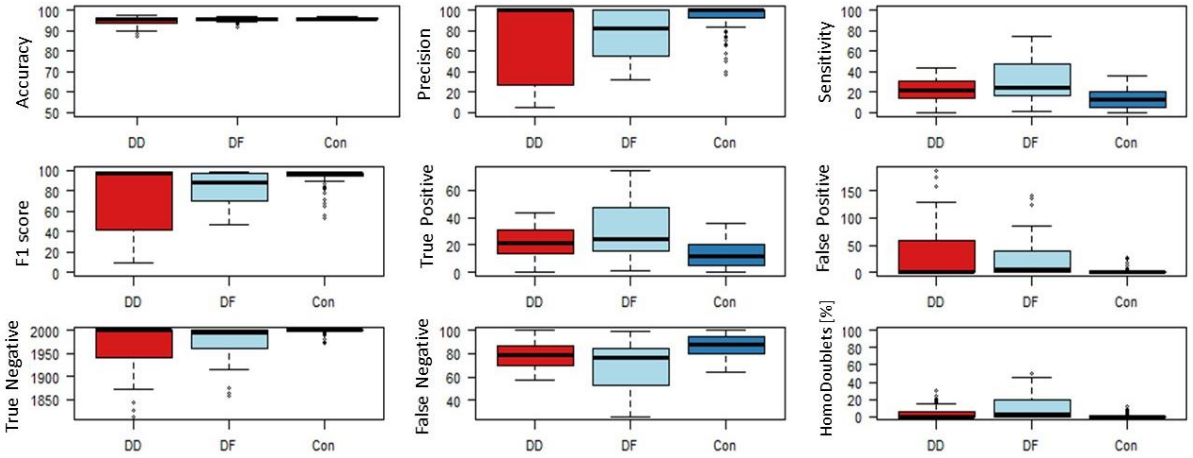


Figure 17. Boxplots of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) on the Sim3 datasets. The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives, false negatives and percentage of homotypic doublets detected.

Table 6. Average values of the performance metrics scored by DoubletDecon (DD), DoubletFinder (DF) and Consensus (Con) in the four simulations (Sim1, Sim2, Sim3, Sim4). The performance metrics considered: accuracy, precision, sensitivity, F1 score, number of true positives, false positives, true negatives, false negatives and percentage of homotypic doublets detected. The shaded cells identify the method that has the best performance for each metric.

Statistic	Simulation	DD	DF	Con
Accuracy	Sim1	94.47	96.35	97.25
Accuracy	Sim2	82.18	86.77	82.72
Accuracy	Sim3	94.81	95.63	84.55
Accuracy	Sim4	92.91	94.12	93.75
Precision	Sim1	35.69	56.75	83.66
Precision	Sim2	71.87	93.15	98.77
Precision	Sim3	64.76	76.99	84.55
Precision	Sim4	72.31	81.57	96.23
Sensitivity	Sim1	25.64	54.74	19.56
Sensitivity	Sim2	29.54	40.46	18.01
Sensitivity	Sim3	21.38	30.03	12.76
Sensitivity	Sim4	22.87	31.37	13.29
F1 Score	Sim1	45.14	69.10	88.98
F1 Score	Sim2	75.91	89.76	90.02
F1 Score	Sim3	74.81	83.58	93.70
F1 Score	Sim4	76.76	86.52	94.85
TP	Sim1	0.83	1.77	0.63
TP	Sim2	6.20	8.50	3.78
TP	Sim3	1.02	1.43	0.61
TP	Sim4	1.63	2.24	0.95
FP	Sim1	3.13	2.19	0.15
FP	Sim2	3.02	0.73	0.06
FP	Sim3	1.45	1.04	0.08
FP	Sim4	1.58	0.97	0.06
FN	Sim1	2.40	1.46	2.59
FN	Sim2	14.80	12.50	17.22
FN	Sim3	3.74	3.33	4.15
FN	Sim4	5.51	4.90	6.19
TN	Sim1	93.65	94.59	96.02
TN	Sim2	75.68	78.27	78.94
TN	Sim3	93.79	94.20	95.16
TN	Sim4	91.28	91.88	92.80
HomoDoublets	Sim1	10.15	32.15	5.56
HomoDoublets	Sim2	11.25	18.28	3.44
HomoDoublets	Sim3	3.81	9.51	1.18
HomoDoublets	Sim4	4.36	10.75	1.77

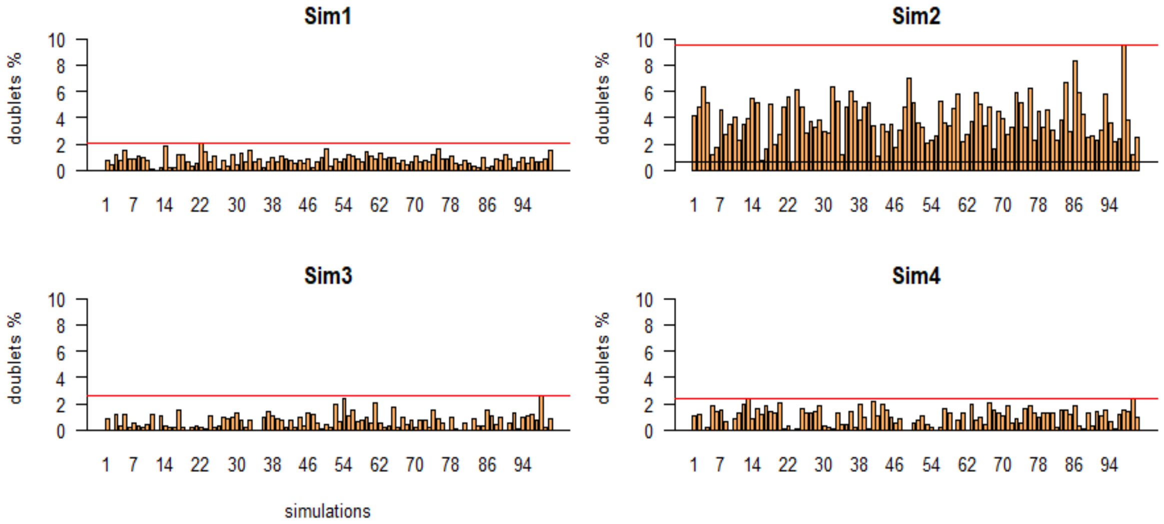


Figure 18. Barplots that represent the percentage of doublets found by the Consensus method (i.e. the fractions of doublets in common between DoubletDecon and DoubletFinder) in the four simulated datasets.The horizontal line represents the average. The horizonal red lines represent the max values and the black one the minimum value.

4.3 Case study: analysis of the tumour microenvironment

As a case study to test the capabilities of the three doublet-detection methods on a real dataset, we considered the study by (Jerby-Arnon et al., 2018), where 7186 cells from the tumour microenvironment of patients with melanoma cancer were subjected to scRNA-seq. For this dataset, a gold standard for benchmarking single/doublet classification was not available, but only cell-type classes (Figure 21).

We pre-processed the data with our Seurat-based pipeline and identified 20 clusters (Figure 19-Figure 20). Figure19b shows the cellular composition of each cluster. It is possible to observe that some clusters are mainly composed of a single cell type: CD8^{+} T cells in cluster 0, B cells in cluster 2, endothelial cells in cluster 16, cancer associated fibroblasts (CAF) in cluster 17, macrophages in cluster 7, and tumour cells in clusters 4, 8, 10, 12 and 18. The mapping of the cell type subpopulation onto the t-SNE embedding allows a better appreciation of the correspondence between the original cell classification (Figure 21) and the clusters obtained in the analysis (Figure 20), which confirms the robustness of our pre-processing approach.

By looking at the singlet/doublet classification performed by DoubletFinder, three main clusters were identified as “contaminated” by doublets, i.e. 7, 16, and 17, which might preferentially correspond to homotypic doublets from macrophages, endothelial cells, and CAFs, respectively (Figure 19d and Figure 23). 8% of the cells in cluster 7 are classified as doublets also by DoubletDecon but, in general, the two methods obtained a very low agreement in this dataset like in subsampled/simulated datasets (Figure 19c and Figure 22). This resulted in a very limited percentage of doublets per cluster identified by the Consensus method (Figure 19e and Figure 23). Overall, DoubletDecon, DoubletFinder, and Consensus identified 0-38, 0-171, and 0-3 doublets per cluster, respectively. For instance, DoubletDecon identified 14% of cells in cluster 11 as doublets, whereas DoubletFinder only 1%. However, this cluster contains different T cell subpopulations that are characterised by similar expression profiles (see also Figure 21). Clusters 11 and 19 group together cells with different transcriptional profiles like B and T cells, which might represent true doublets of cells that were interacting in the tumour microenvironment. For DoubletDecon classification, these clusters contain 14% and 1,4% doublets, respectively, whereas for DoubletFinder cluster 11 contains 1% of doublets while cluster 19 is considered without doublets. Finally, both methods identified several doublets (3% for DoubletDecon and the same for DoubletFinder) in cluster 20, which can reveal previously interacting cells macrophages, B cells, and CD4$^+$ T cells. However, the low agreement between the methods (3% consensus doublets, Figure 19e) prevents a clear interpretation. Finally, the methods identified doublets in clusters characterised by the presence of unknown cells (e.g. 5, 14, 17, 20), which might indeed be doublet cells showing a hybrid transcriptional profile.

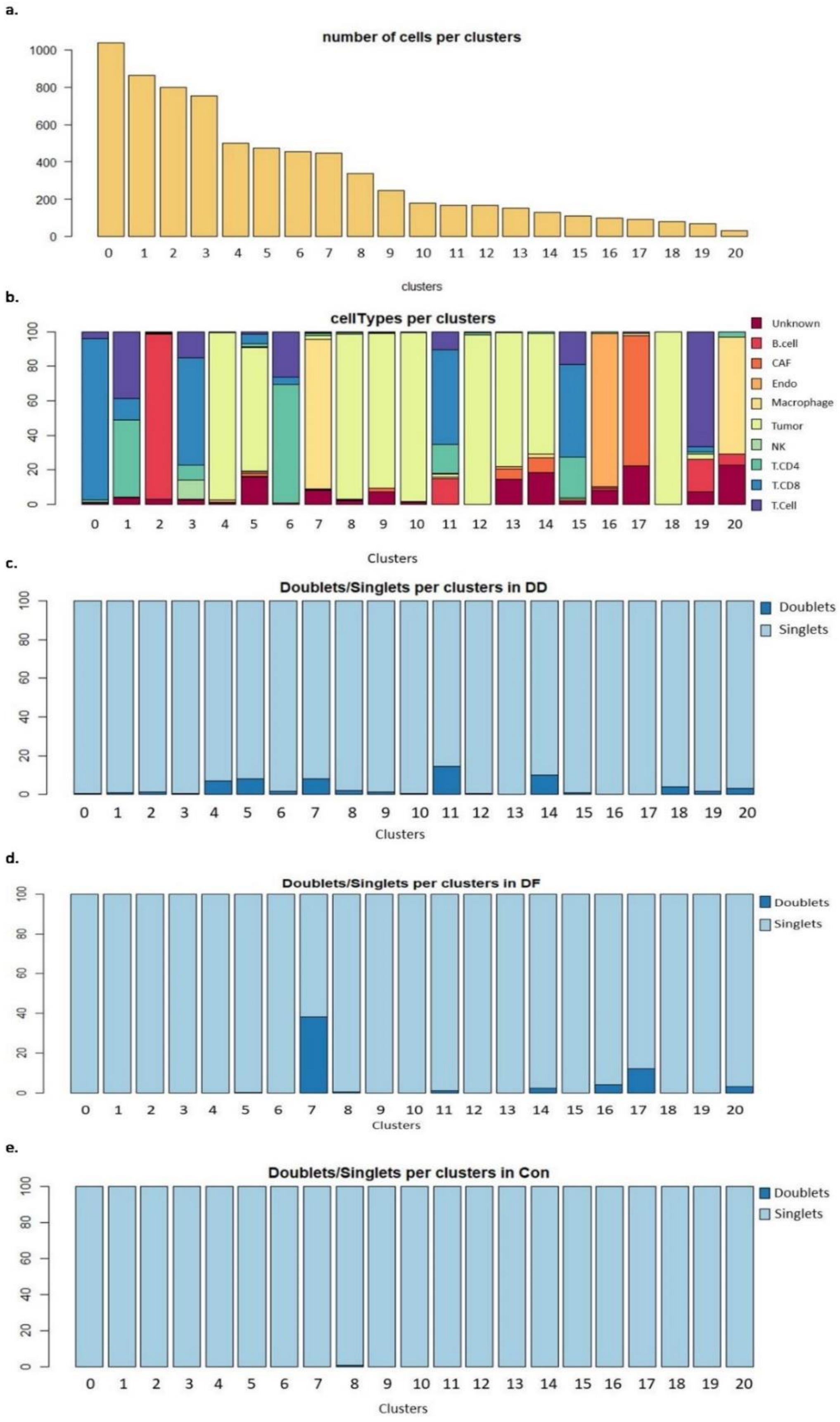


Figure 19. *Analysis of the data from the study of Jerby-Arnon at al., 2018. A) Barplot of the number of cells per cluster; b) Stacked barplot of the cellular composition of each cluster. Stacked barplots of the doublets/singlets per clusters identified by DoubletDecon (DD) (c), DoubletFinder (DF) (d) and Consensus (Con) (e).*

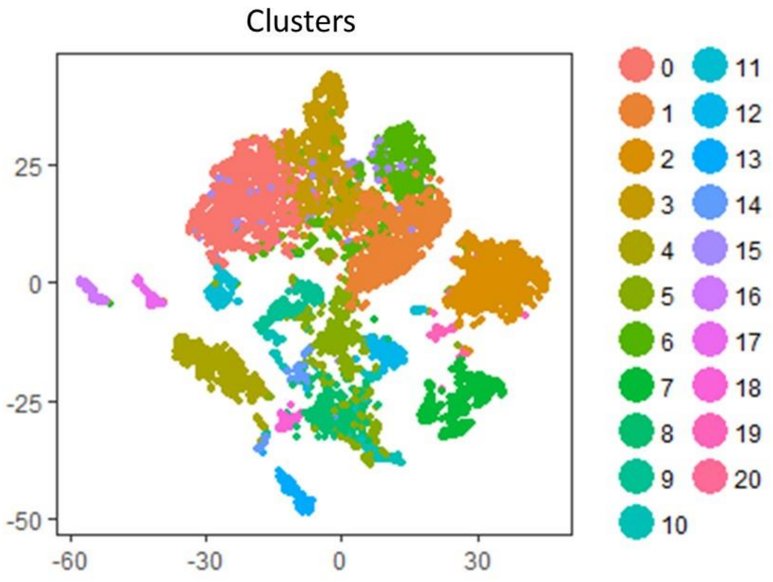


Figure 20. t-SNE plot of the data from Jerby-Arnon at al., 2018 based on the cluster identified through Seurat analysis.

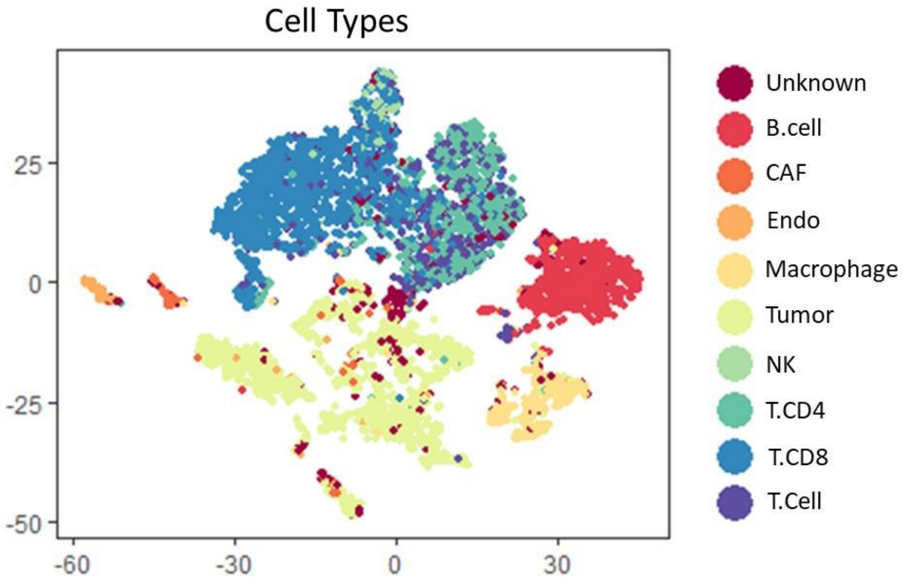


Figure 6. t-SNE plot based on the cell Types from Jerby-Arnon at al., 2018.

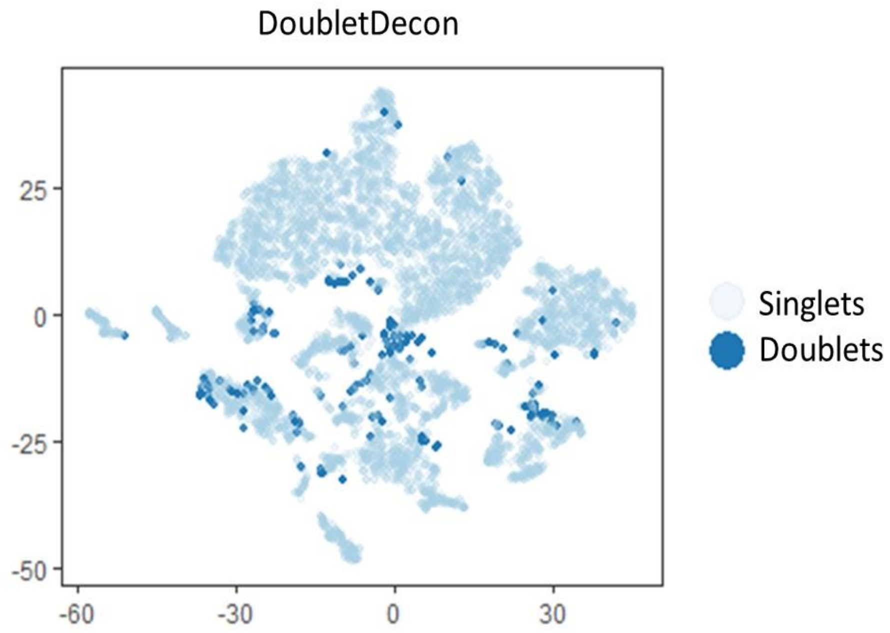


Figure 22. t-SNE plot based on doublet/singlets cell classification carried out by DoubletDecon.

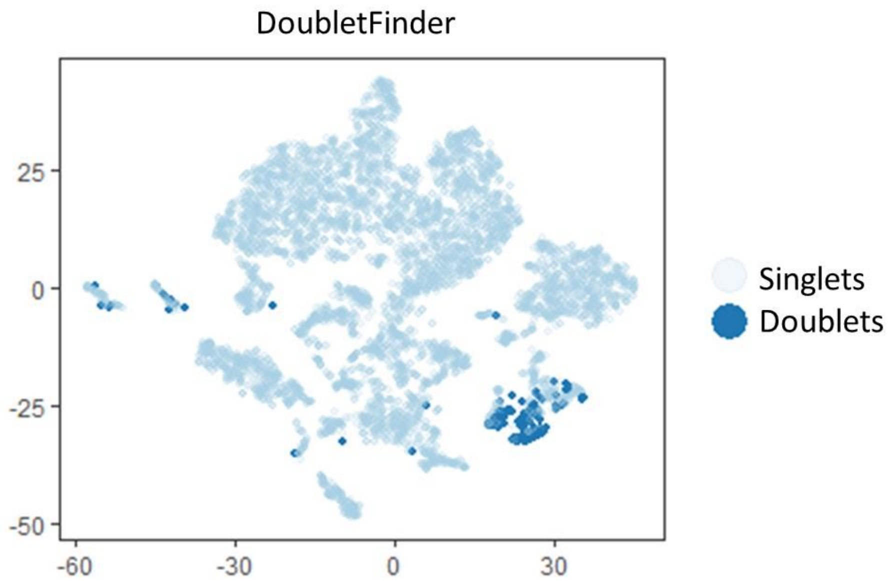


Figure 23. t-SNE plot based on doublet/singlets cell classification carried out by DoubletFinder.

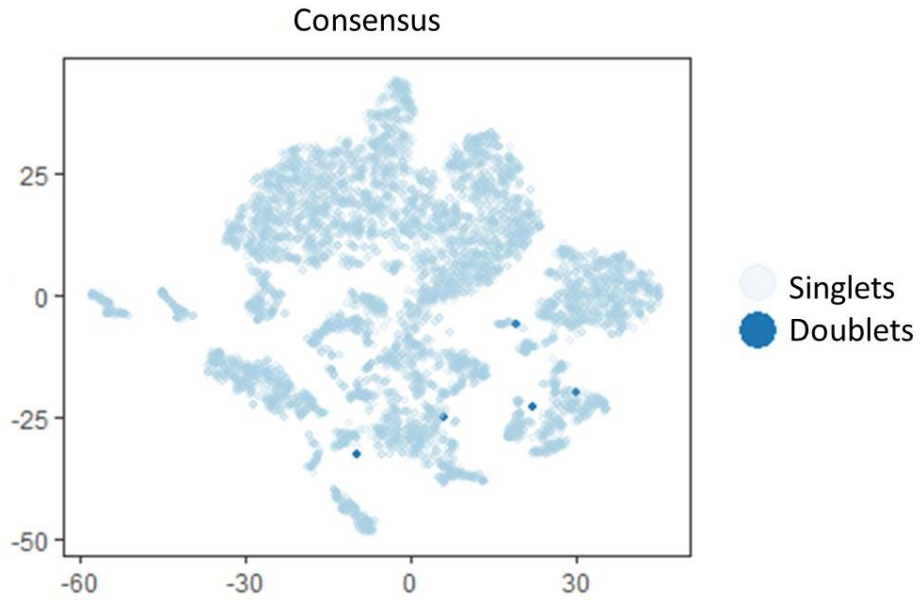


Figure 24. t-SNE plot based on doublet/singlets cell classification carried out by Consensus.

Discussion

To benchmark the performance of R-based methods for the detection of doublets from single-cell RNA-seq data, we generated 7 scRNA-seq datasets (using a subsampling or simulated approach, see Chapter 3.1), for a total of 16,300 single-cells. Before applying doublet classification methods, we implemented and applied a full pipeline for the pre-processing of the single-cell expression data based on Seurat package. The data were first normalised and scaled to exclude possible confounding factors and the genes with high variability across the single cells were detected. Subsequently we applied a dimensionality reduction method, PCA, to build a new space that allows for the representation of the multivariable nature of the data in a space with a relatively smaller dimension. The new representation, possible due to the study of the principal components, is useful to identify the structure of the data. Next, we calculated PCA scores that identify which are the significant PCs that should be considered for the analysis. Next, we used a non-linear dimensional reduction technique to discover the underlying manifold of the data in order to place similar cells together in a low-dimensional space. The last pre-processing step was the clustering of the data into groups based on genes similarity.

For doublet detection, we considered two R-based methods: DoubletDecon and DoubletFinder. DoubletDecon is a computational method that applies a deconvolution-based strategy and classifies the cells according to the similarity of the cell profile to individual cells or synthesised doublets. DoubletFinder, in contrast, classifies the cells using simulated doublets and a nearest-neighbours search to find the fraction of these cells near each real data. This is used to define a doublet score that, given a known number of expected doublets, helped to classify the cells. While the number of expected doublets for DoubletFinder has to be specified by the user. For the purpose of this research, we used the number of doublets identified by DoubletDecon as number of expected doublets for DoubletFinder. This choice was made in order to better compare the cells classified by these different methods and to help in the scoring of the classification efficiency starting from the same conditions. In addition, we implemented a third approach, that we called Consensus, that conservatively identifies as doublets only the cells that are classified as doublets by both DoubletDecon and DoubletFinder.

DoubletDecon can provide a good estimation the real number of doublets in the datasets, even if in some datasets the value is underestimated. The relative error tends to be large.

The analysis of the generated datasets revealed that DoubletFinder, when fed with the expected doublet rate estimated by DoubletDecon, could identify true doublets with high sensitivity and accuracy. Moreover, it allowed the identification of the largest fraction of homotypic doublets (i.e. doublets originating from cells with very similar transcriptomes), which are challenging to identify from scRNA-seq data. DoubletDecon, instead, obtained a high number of FP and FN, which severely affected its performance, especially in terms of precision and F1 score. DoubletDecon is useful to pass the estimation of the number of doublets to DoubletFinder, but it is less accurate on the doublets profile identification.

Despite the common doublet rate imposed, the final cell classification in terms of singlets and doublets strongly differed for the two methods, likely due to the high number of erroneous classifications of DoubletDecon. As a result, the Consensus approach, resulted in a very limited number of doublet calls, thus obtaining very low FP-rates, but also low TP-rate, which might be of limited utility in a real world analysis (i.e. most of the true doublets are not identified), except when a conservative approach is needed.

After benchmarking, we applied the method to a scRNA-seq study with 7,186 cells from the tumour microenvironment of 33 melanoma patients (Jerby-Arnon et al., 2018). The methods identified doublets in clusters characterised by the presence of unknown cells, which might be indeed real doublets showing a hybrid transcriptional profile. However, the fraction of estimated doublets, for these and the other clusters, is rarely concordant between the two methods, in accordance with the results obtained on simulated/subsampled data. This strongly affects the applicability of the Consensus approach, which might result too conservative for most of the applications. In this case study, the doublets classified by DoubleFinder tend to group together in clusters, while those from DoubletDecon classification are distributed all over the space in different clusters.

By comparing the divergent results of DoubletDecon and DoubleFinder, we can see that DoubletDecon identified more doublets in tumour-cell clusters and in T-cell clusters. However, both might be affected by false positives due to the large heterogeneity of tumour cell transcriptional profiles, and by the similarity of the expression signatures of T cell subpopulations, respectively. Nevertheless, tumour cells are usually more tightly connected and more difficult to dissociate as the single cells than immune cells, so they might also reflect a certain number of true positives, especially when confirmed also by the Consensus method. Many doublets identified from DoubletFinder belong to single-cell type clusters, which might confirm its capability of identifying homotypic doublets.

Conclusions

Recent years have been marked by the development of numerous single-cell RNA sequencing methods that have led to the improvement in the study of gene expression and provided new insight into cell states and cell types. Single-cell transcriptomics studies have improved the knowledge of cell types variability and cell states during cell cycle. They enable the identification of new unknown cell population and the analysis of the kinetics of the transcripts. Although they improve the analysis of the biological cell expression patterns, they are characterised by different challenges such as the sparsity of the data, dropout events or error in cells capture.

In particular the capture of two cells together, the doublets problem, is a widely studied problem and, every day, new and innovative experimental end computational methods are produced to identify and classify these cells. Computational methods could be useful for doublets identification since they can be applied to data from large-scale studies (with high number of cells) and conducted with different technologies. This thesis research attempts to assess the performance of two computational methods, DoubletDecon and DoubletFinder, using simulated and real data to test them both. We focused on two methods based on R to have a common processing pipeline and impact in the same way downstream analysis. In addition, we implemented the Consensus approach to explore the possibility of improving the doublets classification accuracy by considering only the concordant results from DD and DF.

To benchmark the method, we created several sets of scRNA-seq datasets, using a subsampling or simulated approach, with precise features, for a total of 16,300 cells. We also implemented a unified pipeline in R that includes first the data pre-processing and then the identification, to ease the analysis.

At the end of the analysis we discover that the two methods lead to different results, i.e. most of the cells considered as doublets by the two methods are different. The doublets for which they concord are really few and the Consensus method is very conservative, so creation is not always useful. Even if the cells that it identifies are almost certainly real doublets, since they are defined as doublets by both methods, they are too few to be of interest for most of the analyses.

DoubleDecon provides a good estimation of the number of doublets in the data, but not of their identity. In fact, it is characterised by a high number of FP and FN. DoubletFinder can be tested with the expected number of doublets estimated by DD, but obtains better results in term of accuracy, precision and F1 score. Considering the analysis results we suggest DoubletFinder for doublets classification since it provides an overall better performance.

Despite not being expected, DoubletFinder is also able to identify a good number of homotypic doublets, whereas DoubletDecon is not. It is a significant result, since the identification of doublets which arise from transcriptionally similar cells is even more challenging than the other type.

In order to improve the detection of doublets further analysis should be carried out using more of the developed tool and more benchmark datasets both simulated and real data to better study which cells are captured together more often. The simulated data could aid to systematically explore whether some pairs of cell types are more difficult to identify due to their similar expression profiles.

References

Bach, K., Pensa, S., Grzelak, M., Hadfield, J., Adams, D. J., Marioni, J. C., & Khaled, W. T. (2017). Differentiation dynamics of mammary epithelial cells revealed by single-cell RNA sequencing. *Nature Communications*, *8*(1), 2128. https://doi.org/10.1038/s41467-017-02001-5

Bloom, J. D. (2018). Estimating the frequency of multiplets in single-cell RNA sequencing from cell-mixing experiments. *PeerJ*, *6*, e5578. https://doi.org/10.7717/peerj.5578

DePasquale, E. A. K., Schnell, D. J., Valiente-Alandí, Í., Blaxall, B. C., Grimes, H. L., Singh, H., & Salomonis, N. (2018). DoubletDecon: Cell-State Aware Removal of Single-Cell RNA-Seq Doublets. *BioRxiv*, 364810. https://doi.org/10.1101/364810

Gong, T., & Szustakowski, J. D. (2013). DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, *29*(8), 1083–1085. https://doi.org/10.1093/bioinformatics/btt090

Griffith, M., Walker, J. R., Spies, N. C., Ainscough, B. J., & Griffith, O. L. (2015). Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLOS Computational Biology*, *11*(8), e1004393. Retrieved from https://doi.org/10.1371/journal.pcbi.1004393

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., … Yanai, I. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biology*, *17*(1), 77. https://doi.org/10.1186/s13059-016-0938-8

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., … Morgan, M. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Amit, I. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of

Tissues into Cell Types. *Science*, *343*(6172), 776 LP – 779. https://doi.org/10.1126/science.1247651

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M.-J., Melms, J. C., … Regev, A. (2018). A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell*, *175*(4), 984-997.e24. https://doi.org/https://doi.org/10.1016/j.cell.2018.09.006

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., … Ye, C. J. (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, *36*(1), 89–94. https://doi.org/10.1038/nbt.4042

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, *161*(5), 1187–1201. https://doi.org/https://doi.org/10.1016/j.cell.2015.04.044

Lafzi, A., Moutinho, C., Picelli, S., & Heyn, H. (2018). Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nature Protocols*, *13*(12), 2742–2757. https://doi.org/10.1038/s41596-018-0073-y

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/https://doi.org/10.1016/j.cell.2015.05.002

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*, *8*(4), 329-337.e4. https://doi.org/https://doi.org/10.1016/j.cels.2019.03.003

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells.

Nature Methods, *10*, 1096. Retrieved from https://doi.org/10.1038/nmeth.2639

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

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A., & Mikkelsen, T. S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *BioRxiv*, 3236. https://doi.org/10.1101/003236

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., … Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*, 865. Retrieved from https://doi.org/10.1038/nmeth.4380

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., … Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, *19*(1), 224. https://doi.org/10.1186/s13059-018-1603-1

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, *352*(6282), 189 LP – 196. https://doi.org/10.1126/science.aad0501

Wang, Y. J., Schug, J., Lin, J., Wang, Z., Kossenkov, A., & Kaestner, K. H. (2019). Comparative analysis of commercially available single-cell RNA sequencing platforms for their performance in complex human tissues. *BioRxiv*, 541433. https://doi.org/10.1101/541433

Wolock, S. L., Lopez, R., & Klein, A. M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, *8*(4), 281-291.e9.

https://doi.org/https://doi.org/10.1016/j.cels.2018.11.005

Zappia, L., Phipson, B., & Oshlack, A. (2018). Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLOS Computational Biology*, *14*(6), e1006245. Retrieved from https://doi.org/10.1371/journal.pcbi.1006245

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049