Massively parallel analysis of cells and nucleic acids

JULIA SANDBERG

Doctoral Thesis in Biotechnology
Stockholm, Sweden, 2011
Massively parallel analysis of cells and nucleic acids

Julia Sandberg

Royal Institute of Technology
School of Biotechnology
Stockholm 2011
Abstract
Recent proceedings in biotechnology have enabled completely new avenues in life science research to be explored. By allowing increased parallelization an ever-increasing complexity of cell samples or experiments can be investigated in shorter time and at a lower cost. This facilitates for example large-scale efforts to study cell heterogeneity at the single cell level, by analyzing cells in parallel that also can include global genomic analyses. The work presented in this thesis focuses on massively parallel analysis of cells or nucleic acid samples, demonstrating technology developments in the field as well as use of the technology in life sciences.

In stem cell research issues such as cell morphology, cell differentiation and effects of reprogramming factors are frequently studied, and to obtain information on cell heterogeneity these experiments are preferably carried out on single cells. In paper I we used a high-density microewell device in silicon and glass for culturing and screening of stem cells. Maintained pluripotency in stem cells from human and mouse was demonstrated in a screening assay by antibody staining and the chip was furthermore used for studying neural differentiation. The chip format allows for low sample volumes and rapid high-throughput analysis of single cells, and is compatible with Fluorescence Activated Cell Sorting (FACS) for precise cell selection.

Massively parallel DNA sequencing is revolutionizing genomics research throughout the life sciences by constantly producing increasing amounts of data from one sequencing run. However, the reagent costs and labor requirements in current massively parallel sequencing protocols are still substantial. In paper II we have focused on flow-sorting techniques for improved sample preparation in bead-based massive sequencing platforms, with the aim of increasing the amount of quality data output, as demonstrated on the Roche/454 platform. In paper III we demonstrate a rapid alternative to the existing shotgun sample titration protocol and also use flow-sorting to enrich for beads that carry amplified template DNA after emulsion PCR, thus obtaining pure samples and with no downstream sacrifice of DNA sequencing quality. This should be seen in comparison to the standard 454-enrichment protocol, which gives rise to varying degrees of sample purity, thus affecting the sequence data output of the sequencing run. Massively parallel sequencing is also useful for deep sequencing of specific PCR-amplified targets in parallel. However, unspecific product formation is a common problem in amplicon sequencing and since these shorter products may be difficult to fully remove by standard procedures such as gel purification, and their presence inevitably reduces the number of target sequence reads that can be obtained in each sequencing run. In paper IV we barcode specific fluorescent probe was used for target-specific FACS enrichment to specifically enrich for beads with an amplified target gene on the surface. Through this procedure a nearly three-fold increase in fraction of informative sequence was obtained and with no sequence bias introduced. Barcode labeling of different DNA libraries prior to pooling and emulsion PCR is standard procedure to maximize the number of experiments that can be run in one sequencing lane, while also decreasing the impact of technical noise. However, variation between libraries in quality and GC content affects amplification efficiency, which may result in biased fractions of the different libraries in the sequencing data. In paper V cell heterogeneity within a human being is being investigated by low-coverage whole genome sequencing of single cell material. By focusing on the most variable portion of the human genome, polymgaunine nucleotide repeat regions, variability between different cells is investigated and highly variable polymgaunine repeat loci are identified. By selectively amplifying and sequencing polymgaunine nucleotide repeats from single cells for which the phylogenetic relationship is known, we demonstrate that massively parallel sequencing can be used to study cell-cell variation in length of these repeats, based on which a phylogenetic tree can be drawn.
Keywords: Massively parallel sequencing, 454, Illumina, multiplex amplification, whole genome amplification, single cell, polymgaunine, flow-cytometry
© Julia Sandberg

Julia Sandberg (2011): Massively parallel analysis of cells and nucleic acids
Division of Gene Technology, School of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden.

Abstract
Recent proceedings in biotechnology have enabled completely new avenues in life science research to be explored. By allowing increased parallelization an ever-increasing complexity of cell samples or experiments can be investigated in shorter time and at a lower cost. This facilitates for example large-scale efforts to study cell heterogeneity at the single cell level, by analyzing cells in parallel that also can include global genomic analyses. The work presented in this thesis focuses on massively parallel analysis of cells or nucleic acid samples, demonstrating technology developments in the field as well as use of the technology in life sciences.

In stem cell research issues such as cell morphology, cell differentiation and effects of reprogramming factors are frequently studied, and to obtain information on cell heterogeneity these experiments are preferably carried out on single cells. In paper I we used a high-density microewell device in silicon and glass for culturing and screening of stem cells. Maintained pluripotency in stem cells from human and mouse was demonstrated in a screening assay by antibody staining and the chip was furthermore used for studying neural differentiation. The chip format allows for low sample volumes and rapid high-throughput analysis of single cells, and is compatible with Fluorescence Activated Cell Sorting (FACS) for precise cell selection.

Massively parallel DNA sequencing is revolutionizing genomics research throughout the life sciences by constantly producing increasing amounts of data from one sequencing run. However, the reagent costs and labor requirements in current massively parallel sequencing protocols are still substantial. In paper II we have focused on flow-sorting techniques for improved sample preparation in bead-based massive sequencing platforms, with the aim of increasing the amount of quality data output, as demonstrated on the Roche/454 platform. In paper III we demonstrate a rapid alternative to the existing shotgun sample titration protocol and also use flow-sorting to enrich for beads that carry amplified template DNA after emulsion PCR, thus obtaining pure samples and with no downstream sacrifice of DNA sequencing quality. This should be seen in comparison to the standard 454-enrichment protocol, which gives rise to varying degrees of sample purity, thus affecting the sequence data output of the sequencing run. Massively parallel sequencing is also useful for deep sequencing of specific PCR-amplified targets in parallel. However, unspecific product formation is a common problem in amplicon sequencing and since these shorter products may be difficult to fully remove by standard procedures such as gel purification, and their presence inevitably reduces the number of target sequence reads that can be obtained in each sequencing run. In paper IV we barcode specific fluorescent probe was used for target-specific FACS enrichment to specifically enrich for beads with an amplified target gene on the surface. Through this procedure a nearly three-fold increase in fraction of informative sequence was obtained and with no sequence bias introduced. Barcode labeling of different DNA libraries prior to pooling and emulsion PCR is standard procedure to maximize the number of experiments that can be run in one sequencing lane, while also decreasing the impact of technical noise. However, variation between libraries in quality and GC content affects amplification efficiency, which may result in biased fractions of the different libraries in the sequencing data. In paper V cell heterogeneity within a human being is being investigated by low-coverage whole genome sequencing of single cell material. By focusing on the most variable portion of the human genome, polymgaunine nucleotide repeat regions, variability between different cells is investigated and highly variable polymgaunine repeat loci are identified. By selectively amplifying and sequencing polymgaunine nucleotide repeats from single cells for which the phylogenetic relationship is known, we demonstrate that massively parallel sequencing can be used to study cell-cell variation in length of these repeats, based on which a phylogenetic tree can be drawn.
Keywords: Massively parallel sequencing, 454, Illumina, multiplex amplification, whole genome amplification, single cell, polymgaunine, flow-cytometry
© Julia Sandberg
Den här boken tillägnas Ulla som gav en bit av sig själv och räddade min pappas liv.
List of publications

This thesis is based upon the following five original papers, which are referred to in the text by their Roman numerals (I-V). The papers are appended at the end of this thesis.


*Authors have contributed equally to the work.

All papers are reproduced with permission from the copyright holders.

Additional publications

Studying single cells

Decisions, decisions. What to choose?

Mechanical micromanipulation

PCR

Targeted capture by hybridization

Selective Circularization

Patterns

Laser Capture Microdissection

Multiplexity limitations of PCR and some ways around them

The cell

Global amplification of genomic DNA

Ligase chain reaction (LCR)

Microwells

Po

INTRODUCTION

Contents

INTRODUCTION

DNA samples

1. The cell

1.1. DNA encodes life

1.2. RNA links the information

1.3. Proteins do the work

1.4. If it only was this simple

2. Studying single cells

2.1. A short recapitulation

2.2. Acquiring cells

2.2.1. Mechanical micromanipulation

2.2.2. Laser Capture Microdissection

2.2.3. Flow cytometry

2.2.4. Non-contact cell printing

2.3. Compartmentalization and manipulation

2.3.1. Microwells

2.3.2. Patterns

2.3.3. Droplets

2.3.4. Cell manipulation in chips

3. Getting enough - complexity reduction and amplification in DNA samples

3.1. Targeted amplification using polymerase

3.1.1. Polymerase Chain Reaction (PCR)

3.1.2. Real-time, Digital, Single Cell and Long-range PCR

3.1.3. Multiplexity limitations of PCR and some ways around them

3.1.4. Isothermal targeted amplification strategies

3.2. Using ligation for increased specificity

3.2.1. Ligase chain reaction (LCR)

3.2.2. Multiplex Ligation Dependent Probe Amplification (MLPA)

3.2.3. Golden Gate

3.2.4. Tri-nucleotide Threading (TnT)

3.3. Selective Circularization

3.3.1. Padlock probes

3.3.2. Molecular Inversion probes (MIP)

3.3.3. Selector

3.4. Targeted capture by hybridization

3.4.1. PCR-based amplification strategies

3.4.2. Multiple Displacement Amplification (MDA)

3.4.3. Linear amplification techniques

3.5. Decisions, decisions. What to choose?

3.6. Global amplification of genomic DNA

3.6.1. PCR-based amplification strategies

3.6.2. Multiple Displacement Amplification (MDA)

3.6.3. Linear amplification techniques
INTRODUCTION

1. The cell

Last year, the breaking news of the creation of the first-ever synthetic cell was published by Craig Venter, regarded by some as an icon of the world of biotechnology. By computationally designing the DNA code, synthesizing it in the lab and finally transplanting it into a recipient cell to replace its original DNA, Venter’s team created living and proliferating bacterial cells containing only man-made genomic material. To verify that the bacterial cell was in fact carrying the synthetic DNA, they inserted a few watermarks in the form of quotations (such as “What I cannot build I cannot understand” - Richard Feynman) into the DNA of the synthetic cell. Despite the fact that the information within its genome is essentially identical to that of existing bacteria, this is astonishing news and I will leave the discussion about “playing God” outside the scope of this thesis and instead focus on the biological side of this achievement.

Imagine building a car from scratch. The work will be much easier if you know what parts are essential for a car to be able to run, decelerate, turn and stop, when you start building it. However, if you don’t, you will certainly learn a lot about the subject along the way. The same is true when making synthetic cells. Being able to build a living cell opens up opportunities to learn more about what parts of the genomic makeup of a cell are really necessary for life. The current state of knowledge of cell biology in general, and some essential technologies in particular, have reached the stage where creating a synthetic cell is possible. This Chapter contains a brief introduction to the basic components and processes of cells, describing a path between three types of macromolecules - DNA, RNA and protein - that are essential for most known forms of life.
The concept of a cell was first recognized in the mid-nineteenth century, when Matthias Jacob Schleiden and Theodor Schwann defined the cell as “the basic structural and functional unit of living organisms”. This theory was soon extended to include the view that new cells are produced by division of preexisting cells and that tissues and organs are made up of specialized cells that are created by cell multiplication. However, it has taken a long time to establish how the life of a cell works and is regulated, and there is still much left to be determined.

1.1. DNA encodes life

Genetic information instructing the eukaryotic cell how to grow, divide and behave is stored in digital form as a set of long molecules, in the cell nucleus. The characteristic double-helical arrangement of deoxyribonucleic acid (DNA) was deduced by Watson and Crick in 1953. DNA is made up of four different building blocks, nucleotides, each of which is in turn composed of a phosphate group, a sugar molecule and a nitrogenous base. Each block can only match – basepair – to one of the other three nucleotides. Basepairing of a DNA nucleotide to its complementary nucleotide is an essential feature of nature as it enables DNA replication, a DNA copying process that is carried out when a cell is to divide into two.

In order for the DNA code to have an impact in the cell, it needs to be translated into a more functional code. This system of multiple layers allows stability and backup, since the information source in itself is quite inert and the functional elements (RNA and proteins), whether they are correctly synthesized or not, are created and degraded in a continuous flow. This process is carried out in two steps.

1.2. RNA links the information

With the help of a collection of enzymes, a part of the long DNA code is copied into code consisting of a similar polymer molecule, ribonucleic acid (RNA). This process, which is called transcription, is done through the incorporation of RNA nucleotides matching and base-pairing to the DNA template, and getting attached to each other, one base at a time, to form an RNA molecule.

There are many types of RNA molecules with different functions in the cell, a subject to which I will return later. Messenger RNA (mRNA) carries the information from DNA to the protein factories of the cells. A few hundred thousand mRNA molecules are present in a human cell at any time; however, their internal ratios and origins may vary. The newly synthesized mRNA molecules are transported out of the nucleus into the ribosomes in the cell cytoplasm or bound to the endoplasmic reticulum, for a final transformation of the genetic code into the third class of macromolecules – proteins.

The concept of a cell was first recognized in the mid-nineteenth century, when Matthias Jacob Schleiden and Theodor Schwann defined the cell as “the basic structural and functional unit of living organisms”. This theory was soon extended to include the view that new cells are produced by division of preexisting cells and that tissues and organs are made up of specialized cells that are created by cell multiplication. However, it has taken a long time to establish how the life of a cell works and is regulated, and there is still much left to be determined.

1.1. DNA encodes life

Genetic information instructing the eukaryotic cell how to grow, divide and behave is stored in digital form as a set of long molecules, in the cell nucleus. The characteristic double-helical arrangement of deoxyribonucleic acid (DNA) was deduced by Watson and Crick in 1953. DNA is made up of four different building blocks, nucleotides, each of which is in turn composed of a phosphate group, a sugar molecule and a nitrogenous base. Each block can only match – basepair – to one of the other three nucleotides. Basepairing of a DNA nucleotide to its complementary nucleotide is an essential feature of nature as it enables DNA replication, a DNA copying process that is carried out when a cell is to divide into two.

In order for the DNA code to have an impact in the cell, it needs to be translated into a more functional code. This system of multiple layers allows stability and backup, since the information source in itself is quite inert and the functional elements (RNA and proteins), whether they are correctly synthesized or not, are created and degraded in a continuous flow. This process is carried out in two steps.

1.2. RNA links the information

With the help of a collection of enzymes, a part of the long DNA code is copied into code consisting of a similar polymer molecule, ribonucleic acid (RNA). This process, which is called transcription, is done through the incorporation of RNA nucleotides matching and base-pairing to the DNA template, and getting attached to each other, one base at a time, to form an RNA molecule.

There are many types of RNA molecules with different functions in the cell, a subject to which I will return later. Messenger RNA (mRNA) carries the information from DNA to the protein factories of the cells. A few hundred thousand mRNA molecules are present in a human cell at any time; however, their internal ratios and origins may vary. The newly synthesized mRNA molecules are transported out of the nucleus into the ribosomes in the cell cytoplasm or bound to the endoplasmic reticulum, for a final transformation of the genetic code into the third class of macromolecules – proteins.
1.3. Proteins do the work

The next step involves the interpretation of the ribonucleic acid code, which consists of four letters, into the protein code of twenty letters in a process called translation. Different combinations of three consecutive nucleotides in the RNA sequence correspond to different amino acids. Thus, by using an mRNA sequence as template, amino acids can be coupled together in a long chain. Based on the order of the amino acids, proteins are folded into well-defined three-dimensional structures, which are fundamental to their particular functions. Proteins appear in practically all cellular machineries and processes, with widely varying functions, including catalyzing chemical reactions, giving structure and firmness to tissues and organs, or defending the cell against hostile attack. A cell contains proteins in varying concentrations; abundant ones may be present in counts of ~10 million in a cell, whereas rare proteins can be found at only 1-10 copies per cell\(^3\).

The behavior of the cell is largely determined by its protein content. A useful image is to think of the DNA sequence as a cookbook, where the genes are the recipes describing to the cell how to cook up the different proteins. In an oversimplification one could state that the more mRNA coding for a particular gene present at the ribosomes, the more of that protein will be produced. The pattern of protein levels in a cell gives the cell its particular traits and determines the cellular phenotype. This model of informational flow, in combination with DNA replication, is called the central dogma of cell biology and was first proposed in 1958 by Francis Crick, and twelve years later updated in a widely cited article\(^3\).

To some extent this model still holds true, but since then the picture has become much more complex.

1.4. If it only was this simple

The picture of basic cellular processes has become more complicated since the central dogma was first proposed. It is now widely accepted that the regulatory machinery that tightly controls cell behavior acts in response to information input from the surrounding environment and neighboring cells. Tight regulation of transcription rate, transcript degradation, protein translation and the breakdown of proteins are all necessary for cell survival and this regulation plays a crucial role during embryonic development, as different developmental states require different protein combinations in the cell. Biological processes are driven by networks rather than genes. Gene networks are clusters of genes with causal interactions so that their state influences the levels of expression of other genes directly or indirectly, through their RNA and protein expression products\(^5\).

It is not only the protein-coding regions of the DNA that contain biologically meaningful information. The ~20,300 protein-coding genes in the human genome are surrounded by...
large amounts of mostly repetitive DNA which is commonly called non-coding DNA\textsuperscript{11}. In fact, almost all the DNA sequences in the genome, the majority of which does not encode proteins, is transcribed in the cell and various functions for these sequences have been identified\textsuperscript{10}. RNA molecules can play a more direct role in determining cellular characteristics by catalyzing chemical reactions, such as cutting and ligating other RNA molecules\textsuperscript{11}, and small RNA molecules are widely involved in regulation of transcription and translation through a series of mechanisms. Moreover, splicing of mRNAs prior to translation into protein is a common cellular process that gives rise to differences in protein function. Post-translational protein modification is another level at which protein function and thus cell behavior is shaped. Finally, chromatin structure affects how accessible the DNA is to the enzymatic machinery needed for transcription, and through this affects the transcription level. Chromatin structure can be inherited over many cell generations and therefore allows cellular memory\textsuperscript{10}.

It is obvious that there is no unidirectional flow of information as suggested by Crick et al in\textsuperscript{1970}\textsuperscript{9}. Instead there is a much more complicated picture, which requires information flow in several dimensions and with the participation of many types of molecules\textsuperscript{10}. Another interesting issue is the cell-to-cell variation in genetic makeup and phenotype. This will be discussed in the following chapter.
2. Studying single cells

Cells in a particular tissue are not identical. Instead, cells that are identical from a genomic point of view may have considerable variation in gene expression profile and protein levels\(^{13,5}\), giving rise to a heterogeneous collection of cells with different behavior and appearance. Sometimes the genomic DNA sequence varies between neighboring cells within a tissue, a feature that can be used to study tissue and tumor development\(^{13,5}\). Tissues are commonly made up from several cell types. When studying multiple cells, or a mixture of genomic DNA from several cells, the average characteristics of the bulk are usually obtained and information on rare cells may be lost\(^{12}\). Investigating single cells renders a more detailed view and thus is preferable. Traditionally, and in many cases currently, studies have been carried out on a collection of cells; in studies aimed at global analysis and/or investigating large/average differences between samples this approach is still valid.

Cell-cell variability is an interesting aspect to study in many biological contexts. In some diseases, a combination of traits is known to give a cell its particular disease phenotype and obtaining information on cell heterogeneity within a tissue is sometimes crucial. For instance, it has been argued that knowing more about the co-occurrence of several different mutations within the same cell is essential to fully understand the nature and evolution of cancer\(^{13,5}\). In other cases, a cell population may be made up of two subpopulations with distinctly varying traits, making the information obtained dependant upon their proportions in a mixed sample. Circulating tumor cells in cancer patients and fetal cells in the blood of pregnant women are two examples of rare cells that are difficult to study in their natural habitat and therefore need to be isolated for proper characterization of cancer genotype, and risk of metastasis or fetal aneuploidy, respectively. In microbial analysis, many species cannot be cultured in the laboratory making single cell analysis the only option\(^{16}\). By monitoring single cell proliferation over time, the response of a heterogeneous cell population to a particular treatment can be studied. This is particularly useful in cancer research, where understanding the behavior of individual cells within a tumor can provide insights into the mechanisms of tumor growth and metastasis.

For example, circulating tumor cells in cancer patients and fetal aneuploidy, respectively. In microbial analysis, many species cannot be cultured in the laboratory making single cell analysis the only option\(^{16}\). By monitoring single cell proliferation over time, the response of a heterogeneous cell population to a particular treatment can be studied. This is particularly useful in cancer research, where understanding the behavior of individual cells within a tumor can provide insights into the mechanisms of tumor growth and metastasis.
population to various agents can be evaluated\(^1\). Also, multiplication and analysis of single cell clones is a way to increase the genetic material available for analysis, by having the selected cell pass through an evolutionary bottleneck\(^2\).

In order to get information from single cells one need to isolate, study, and sometimes culture then, separately. In this chapter a selection of tools that can be used to single out and isolate particular cells of interest will be described.

### 2.1. A short recapitulation

The field of cell biology has always been dependent on technology development. Development of high-resolution microscopy in combination with specific staining methods in 17th century made it possible to observe tissues and individual cells for the first time\(^3\)-\(^5\). Nowadays, microscopy is used daily in most biological laboratories, to investigate, e.g., the intracellular localization of particular proteins and even the division of living cells, when using live-cell imaging in time-lapse microscopy\(^6\)-\(^8\). Methods such as electrophysiology, where the voltage over a single cell or ion channel is measured by attaching a pipette containing a volt-meter to the surface of a cell, give information about one cell at a time\(^7\),\(^8\).

The same is true for flow cytometry, by which the sizes of, and fluorescent signals from, individual cells in a stream can be monitored (cf. below). These methods have been around for a long time and by their nature give information about the properties of single cells. However, when it comes to biochemical approaches for investigating DNA sequences, gene expression levels or the amount of small molecules or protein in a cell, there is a need for tools offering higher sensitivity\(^9\).

One way to increase sensitivity is through miniaturization. Since human cells are about 10 \(\mu\)m in size and occupy a volume of less than a \(\mu\)l, the standard volumes of 50 \(\mu\)l and more, commonly used in laboratories, can be excessive when studying small numbers of cells. A reduction in reaction volume has been identified as a way to enhance the sensitivity and speed of reactions by allowing shorter diffusion distances and smaller dilution factors\(^10\)-\(^15\). Another obvious advantage of smaller reaction volumes is a decrease in reagent costs. In recent times, the numbers of publications in the fields of genomics, transcriptomics and proteomics describing single cell analysis utilizing miniaturization techniques have dramatically increased. Transcriptomics and genomics will be described in more detail in later chapters, however a few examples will also be presented in this section.

Regardless of the biomolecule of interest, single cell analysis has some general technical features, such as picking up single cells (possibly based on certain traits such as cell surface markers) and putting these cells in a container where the downstream analysis can be carried out. There are several readily available platforms for these types of tasks, a handful of which will be presented here.

population to various agents can be evaluated\(^1\). Also, multiplication and analysis of single cell clones is a way to increase the genetic material available for analysis, by having the selected cell pass through an evolutionary bottleneck\(^2\).

In order to get information from single cells one need to isolate, study, and sometimes culture them, separately. In this chapter a selection of tools that can be used to single out and isolate particular cells of interest will be described.

### 2.1. A short recapitulation

The field of cell biology has always been dependent on technology development. Development of high-resolution microscopy in combination with specific staining methods in 17th century made it possible to observe tissues and individual cells for the first time\(^3\)-\(^5\). Nowadays, microscopy is used daily in most biological laboratories, to investigate, e.g., the intracellular localization of particular proteins and even the division of living cells, when using live-cell imaging in time-lapse microscopy\(^6\)-\(^8\). Methods such as electrophysiology, where the voltage over a single cell or ion channel is measured by attaching a pipette containing a volt-meter to the surface of a cell, give information about one cell at a time\(^7\),\(^8\).

The same is true for flow cytometry, by which the sizes of, and fluorescent signals from, individual cells in a stream can be monitored (cf. below). These methods have been around for a long time and by their nature give information about the properties of single cells. However, when it comes to biochemical approaches for investigating DNA sequences, gene expression levels or the amount of small molecules or protein in a cell, there is a need for tools offering higher sensitivity\(^9\).

One way to increase sensitivity is through miniaturization. Since human cells are about 10 \(\mu\)m in size and occupy a volume of less than a \(\mu\)l, the standard volumes of 50 \(\mu\)l and more, commonly used in laboratories, can be excessive when studying small numbers of cells. A reduction in reaction volume has been identified as a way to enhance the sensitivity and speed of reactions by allowing shorter diffusion distances and smaller dilution factors\(^10\)-\(^15\). Another obvious advantage of smaller reaction volumes is a decrease in reagent costs. In recent times, the numbers of publications in the fields of genomics, transcriptomics and proteomics describing single cell analysis utilizing miniaturization techniques have dramatically increased. Transcriptomics and genomics will be described in more detail in later chapters, however a few examples will also be presented in this section.

Regardless of the biomolecule of interest, single cell analysis has some general technical features, such as picking up single cells (possibly based on certain traits such as cell surface markers) and putting these cells in a container where the downstream analysis can be carried out. There are several readily available platforms for these types of tasks, a handful of which will be presented here.
2.2. Acquiring cells

There are many ways to manipulate cells into certain positions. Manual seeding, causing cells to be randomly placed in response to gravity, is the simplest, and standard for some platforms, but when a more precise selection or positioning of cells is wanted, alternative methods may be used.

2.2.1. Mechanical micromanipulation

Mechanical micromanipulation is a technique whereby cells can be handled on an inverted microscope stage, carrying out extremely small-scale operations through a joystick that hydraulically operates glass micropipettes. Mechanical micromanipulation of cells is an accessible but labor-intensive method\(^\text{28}\) that can be used for manipulating cultured cells or cells from liquid samples such as blood, saliva and semen\(^\text{43,44}\). The method first saw clinical use in in-vitro fertilization (IVF) treatment in cases of "slow sperms", and it is still commonly used in assisted fertilization\(^\text{46}\) which was awarded the Nobel price in Physiology or Medicine in 2010. Disadvantages of mechanical micromanipulation include low throughput and the fact that it subjects the cells to mechanical shearing\(^\text{39}\).

2.2.2. Laser Capture Microdissection

Laser Capture Microdissection (LCM) was invented in 1996 and basically uses a microscope coupled to a laser constructed to collect cells. Cells are selected based on morphology or immunohistochemical labeling, liberated from surrounding tissue by laser cutting and then, either by catapulting or gravitation depending on the system, fall into a well or the lid of a tube or onto a slide\(^\text{38}\). A wide range of tissue preparation methods are possible with this technique, including formalin or alcohol fixation, paraffin embedding and fresh-freezing and living cells in suspension can also be handled\(^\text{39}\) and re-culturing of LCM-collected cells is possible. LCM has been used when investigating neurogenesis by comparing genome-wide transcriptional profiles from mature neurons and progenitor cells in the olfactory system\(^\text{40}\), for picking single cells or chromosomes for cytogenetic\(^\text{41}\) and genetic\(^\text{41}\) analysis. However, it starting with tissue sections, a limitation is the fact that some nuclei may be sliced which means loss of chromosome parts, which will affect the genomic data\(^\text{42}\).

2.2.3. Flow cytometry

When flow cytometry was invented in the 1970s it radically improved the throughput of single cell analysis\(^\text{43,44}\). The principle of flow cytometry is as follows. Labeled cells are hydrodynamically focused in the center of a buffer stream and laser beams interrogate the cells as they flow by. Information on cell size, granularity and presence of up to 17 fluorescent dyes\(^\text{43}\) is acquired separately for each cell, at a high speed of up to ∼200,000 cells per second. For collection of interesting cells a Fluorescence Activated Cell Sorter (FACS) uses vibrations to break the sample stream into droplets and those containing interesting cells receive an electrical charge and are later diverted by an electrostatic deflection system into a desired spot\(^\text{43}\). With current FACS instruments, sorting can be done at a speed of

2.2. Acquiring cells

There are many ways to manipulate cells into certain positions. Manual seeding, causing cells to be randomly placed in response to gravity, is the simplest, and standard for some platforms, but when a more precise selection or positioning of cells is wanted, alternative methods may be used.

2.2.1. Mechanical micromanipulation

Mechanical micromanipulation is a technique whereby cells can be handled on an inverted microscope stage, carrying out extremely small-scale operations through a joystick that hydraulically operates glass micropipettes. Mechanical micromanipulation of cells is an accessible but labor-intensive method\(^\text{28}\) that can be used for manipulating cultured cells or cells from liquid samples such as blood, saliva and semen\(^\text{43,44}\). The method first saw clinical use in in-vitro fertilization (IVF) treatment in cases of "slow sperms", and it is still commonly used in assisted fertilization\(^\text{46}\) which was awarded the Nobel price in Physiology or Medicine in 2010. Disadvantages of mechanical micromanipulation include low throughput and the fact that it subjects the cells to mechanical shearing\(^\text{39}\).

2.2.2. Laser Capture Microdissection

Laser Capture Microdissection (LCM) was invented in 1996 and basically uses a microscope coupled to a laser constructed to collect cells. Cells are selected based on morphology or immunohistochemical labeling, liberated from surrounding tissue by laser cutting and then, either by catapulting or gravitation depending on the system, fall into a well or the lid of a tube or onto a slide\(^\text{38}\). A wide range of tissue preparation methods are possible with this technique, including formalin or alcohol fixation, paraffin embedding and fresh-freezing and living cells in suspension can also be handled\(^\text{39}\) and re-culturing of LCM-collected cells is possible. LCM has been used when investigating neurogenesis by comparing genome-wide transcriptional profiles from mature neurons and progenitor cells in the olfactory system\(^\text{40}\), for picking single cells or chromosomes for cytogenetic\(^\text{41}\) and genetic\(^\text{41}\) analysis. However, it starting with tissue sections, a limitation is the fact that some nuclei may be sliced which means loss of chromosome parts, which will affect the genomic data\(^\text{42}\).

2.2.3. Flow cytometry

When flow cytometry was invented in the 1970s it radically improved the throughput of single cell analysis\(^\text{43,44}\). The principle of flow cytometry is as follows. Labeled cells are hydrodynamically focused in the center of a buffer stream and laser beams interrogate the cells as they flow by. Information on cell size, granularity and presence of up to 17 fluorescent dyes\(^\text{43}\) is acquired separately for each cell, at a high speed of up to ∼200,000 cells per second. For collection of interesting cells a Fluorescence Activated Cell Sorter (FACS) uses vibrations to break the sample stream into droplets and those containing interesting cells receive an electrical charge and are later diverted by an electrostatic deflection system into a desired spot\(^\text{43}\). With current FACS instruments, sorting can be done at a speed of
70,000 cells/second. However, if cell viability is of the essence slower speeds are recommended. Flow cytometry is commonly used to study cell surface markers\textsuperscript{47}, intracellular protein abundance\textsuperscript{49}, binding of peptide libraries displayed on cells\textsuperscript{50} and for investigating ploidy status of tumor cells\textsuperscript{51}. Double FACS sorting has been used as a means to purify single marine cyanobacterial cells free from contaminating DNA in order to sequence the genome\textsuperscript{45}. Despite the requirement for sophisticated instrumentation, FACS instruments are found in most research institutions and hospitals and the technique is routinely used to sort and analyze cells from leukemias patients\textsuperscript{52}.

2.2.4. Non-contact cell printing

An ordinary desktop computer printer can be modified to print cells onto “bio-papers” or membranes, at high resolution, without touching the surface\textsuperscript{53}. Non-contact printing (aka inkjet printing) is preferable when cells are sensitive to touch and inkjet printing has been carried out using a number of different cell types including mouse motor neurons, with high viabilities of $\sim90\%$\textsuperscript{54}. Cell printing is a method still in development and its final application is not yet clear, however it is considered a promising tool for printing three-dimensional structures containing several cell types. Applications would include building organs for transplantation or constructing cellular model systems to study cell-cell interactions\textsuperscript{55}.

Oil-submerged inkjet printing has been used to avoid cell dehydration when printing single-cell arrays on a glass slide\textsuperscript{56} and a cell printing device guided by a CCD camera has been used to print single unlabeled HeLa cells in microwells and on slides with $75\%$ viability\textsuperscript{57}. An obvious disadvantage of cell printing techniques is that cell selection needs to be done in a separate step, prior to printing; however the high throughput and flexibility of printing pattern are advantages of this method.

2.3. Compartmentalization and manipulation

There are a number of ways to compartmentalize the cells of choice, all of which have the advantage of requiring smaller volumes than the $50 \mu l$ - $1 ml$ volumes commonly used in the lab.

2.3.1. Microwells

Dividing cells between separate wells on a chip is an efficient way of analyzing many cells in parallel. There are a number of microwell devices on the market and many more “home-made” alternatives; their varying features make them useful for a range of applications. Well sizes vary from just enough to hold one bacteria ($2-3 \mu m \ O_2$ circular)\textsuperscript{58} or animal cell ($10 \mu m \ O_2$ circular)\textsuperscript{59}, up to those sufficiently large to enable cell culturing for days or weeks ($150-650 \mu m$ sides, at well bottom)\textsuperscript{59,60}.
Glass bottoms enable microscopy analysis and high-content screening, whereas chips with bottoms made from less transparent materials such as silicon have an impact on the types of analysis that are possible\textsuperscript{14,15}. Microfluidic chips can also be made by cutting wells in polyester using a laser\textsuperscript{16}, or molding chips from polyethylene glycol (PEG)\textsuperscript{17,18} and poly(dimethylsiloxane) (PDMS)\textsuperscript{19}. Adherent cell culturing in wells may require coating with extracellular matrix components such as fibronectin for adhesion\textsuperscript{20}, and chemical treatment has been used to enhance seeding efficiency by placing repelling surfaces on the well walls and attracting surfaces at the bottoms of the wells\textsuperscript{21}.

Analysis of single cells in microfluidics has been utilized to screen hybridomas for cells producing antigen-specific antibodies\textsuperscript{21–24} as well as cytokine release from single T-cells\textsuperscript{25}. Due to their wide range of characteristics, microfluidic devices may be used for many types of cell deposition techniques such as flow cytometry, manual seeding and LCM.

2.3.2. Patterns

Patterning of glass slides can be utilized both for culturing purposes and for compartmentalization of cells and reaction buffers for direct analysis of single cells. For culturing of single seeded cells, patterns of cell adhesion molecules such as collagen can be printed using a modified computer printer, resulting in a dense pattern of 2,500 µm spots\textsuperscript{26}. Alternatively, hydrophilic patches created by photolithography, may be used together with a gas permeable oil layer\textsuperscript{27}, both facilitating both separate attachment and growth of cells in an ordered manner. Patterning of extracellular matrix spots has been used to investigate the impact of cell shape on apoptosis rate in fibroblasts\textsuperscript{28}.

Enzymatic reactions at the single-cell level can be done using slides with hydrophilic reaction sites surrounded by hydrophobic circles that enable encapsulation of aqueous reaction solution under a lid of oil preventing evaporation during heating. DNA methylation pattern analysis and DNA fingerprinting in combination with cytogenetic analysis of single cells have been performed using the Ampligrid platform (BC, CA, USA), which allows for 48 reactions, each in an area 1 µm in diameter on one slide\textsuperscript{29–31}. Depending on the pattern, these tools may be suitable for cell seeding by hand, LCM, flow cytometry, or inkjet printing and also allow easy interrogation of cell seeding by microscopy or scanning\textsuperscript{32}.

2.3.3. Droplets

Microreactors in aqueous phase can be encapsulated in oil, as in a vinaigrette, to allow separation of single cells or molecules. Emulsions may be created by dropwise addition of aqueous phase to oil phase while stirring\textsuperscript{33}, vigorous shaking of the two\textsuperscript{34} or, with the advantage of a more homogeneous emulsion, through merging of channels containing oil and aqueous phase in a microfluidic chip\textsuperscript{35}. The latter requires more sophisticated laboratory apparatus, but it opens up additional possibilities, such as stirring, splitting and merging, for the manipulation of droplets in the chip; it also facilitates fluorescence analysis and droplet sorting\textsuperscript{36}. With some additional work, sorting of emulsion droplets in a double emulsion can be done using an ordinary FACS instrument\textsuperscript{37} and this approach has been utilized in enzyme library screening experiments\textsuperscript{38}.

Glass bottoms enable microscopy analysis and high-content screening, whereas chips with bottoms made from less transparent materials such as silicon have an impact on the types of analysis that are possible\textsuperscript{14,15}. Microfluidic chips can also be made by cutting wells in polyester using a laser\textsuperscript{16}, or molding chips from polyethylene glycol (PEG)\textsuperscript{17,18} and poly(dimethylsiloxane) (PDMS)\textsuperscript{19}. Adherent cell culturing in wells may require coating with extracellular matrix components such as fibronectin for adhesion\textsuperscript{20}, and chemical treatment has been used to enhance seeding efficiency by placing repelling surfaces on the well walls and attracting surfaces at the bottoms of the wells\textsuperscript{21}.

Analysis of single cells in microfluidics has been utilized to screen hybridomas for cells producing antigen-specific antibodies\textsuperscript{21–24} as well as cytokine release from single T-cells\textsuperscript{25}. Due to their wide range of characteristics, microfluidic devices may be used for many types of cell deposition techniques such as flow cytometry, manual seeding and LCM.

2.3.2. Patterns

Patterning of glass slides can be utilized both for culturing purposes and for compartmentalization of cells and reaction buffers for direct analysis of single cells. For culturing of single seeded cells, patterns of cell adhesion molecules such as collagen can be printed using a modified computer printer, resulting in a dense pattern of 2,500 µm spots\textsuperscript{26}. Alternatively, hydrophilic patches created by photolithography, may be used together with a gas permeable oil layer\textsuperscript{27}, both facilitating both separate attachment and growth of cells in an ordered manner. Patterning of extracellular matrix spots has been used to investigate the impact of cell shape on apoptosis rate in fibroblasts\textsuperscript{28}.

Enzymatic reactions at the single-cell level can be done using slides with hydrophilic reaction sites surrounded by hydrophobic circles that enable encapsulation of aqueous reaction solution under a lid of oil preventing evaporation during heating. DNA methylation pattern analysis and DNA fingerprinting in combination with cytogenetic analysis of single cells have been performed using the Ampligrid platform (BC, CA, USA), which allows for 48 reactions, each in an area 1 µm in diameter on one slide\textsuperscript{29–31}. Depending on the pattern, these tools may be suitable for cell seeding by hand, LCM, flow cytometry, or inkjet printing and also allow easy interrogation of cell seeding by microscopy or scanning\textsuperscript{32}.

2.3.3. Droplets

Microreactors in aqueous phase can be encapsulated in oil, as in a vinaigrette, to allow separation of single cells or molecules. Emulsions may be created by dropwise addition of aqueous phase to oil phase while stirring\textsuperscript{33}, vigorous shaking of the two\textsuperscript{34} or, with the advantage of a more homogeneous emulsion, through merging of channels containing oil and aqueous phase in a microfluidic chip\textsuperscript{35}. The latter requires more sophisticated laboratory apparatus, but it opens up additional possibilities, such as stirring, splitting and merging, for the manipulation of droplets in the chip; it also facilitates fluorescence analysis and droplet sorting\textsuperscript{36}. With some additional work, sorting of emulsion droplets in a double emulsion can be done using an ordinary FACS instrument\textsuperscript{37} and this approach has been utilized in enzyme library screening experiments\textsuperscript{38}.
A number of different cell types have been successfully studied in emulsion with acceptable or good viabilities\textsuperscript{11}. Drug-screening experiment has been performed in color-coded droplets\textsuperscript{20} and enzymatic signal amplification in droplets has been used for analysis of rare cell-surface markers\textsuperscript{20}. Single-cell multiplexed mutation detection has been done by flow cytometry analysis of beads collected from an agarose-in-oil emulsion\textsuperscript{27}.

### 2.3.4. Cell manipulation in chips

All the functions needed for cell manipulation, chemical experimentation and analysis can be incorporated into a single microfabricated device; this is a cost effective approach which allows rapid screening and has the great advantages of minimal sample dilution and possibilities for automation\textsuperscript{29}. Fluids are directed through an intricate system of channels, inlets and outlets, and cells and biomolecules can be fully or partly isolated in microchannels (volumes ranging from picoliters to tens of nanoliters) or physical traps. A characteristic attribute of microfluidics is laminar flow, which makes diffusion virtually the only spontaneously-occurring mixing process. Hence, if mixing is needed, this issue needs to be approached by incorporating active mixing modules into the chip. Initially, chip fabrication involved silicon and glass, using methods borrowed from the electronics industry. Today, stainless steel and polymers such as PMMA and PDMS are also used\textsuperscript{84}. Cells can be lysed on-chip through electroporation, chemical breaking of the cell membrane, shearing against sharp nanostuctures, acoustic lysis or laser illumination\textsuperscript{75}. Purification of nucleic acids can be done using functionalized silica beads for DNA and polyT beads for RNA purification whereas protein metabolites are commonly analyzed immediately on the chip using e.g. electrophoresis, chromatography or microbioassay\textsuperscript{33,80}. Cell culturing in microfluidics chips is customary, and in a recent paper Kang and coworkers used a multi-layer microfluidic array chip to culture and replate embryoid bodies by simply flipping the chip upside-down\textsuperscript{73}.

Various types of experiments can be performed in microfluidic chips. By capturing a single beating heart cell it was possible to carry out real-time measurement of ionic and metabolic flux, while also observing the cell by microscopy\textsuperscript{49}. For nucleic acid analysis, Quake and coworkers used a microchip to carry out multiplex PCR of 16S rRNA from single bacterial cells for digital analysis of fluorescent signals and further sequencing analysis\textsuperscript{11}. They then developed a similar chip approach for efficient whole-genome amplification in nanoliter volumes and subsequent phylogeny analysis of single bacterial cells from different microbial communities\textsuperscript{11,33}. A haplotype-resolved human genome was also recently obtained using microchips for separate amplification and sequencing of every chromosome from a single cell\textsuperscript{11}.

A number of cell-manipulating tools can be incorporated into microfluidic systems. The use of optical fields to pick and position particles was pioneered in 1960 by Ashkin and coworkers who developed the first optical tweezers device\textsuperscript{70}. Today it is possible to trap, rotate and sort multiple cells based on size or refractive index, by creating a pattern of varying optical intensity over the sample area using lasers\textsuperscript{11,23,85}. Optical methods have been used both in open chips and in enclosed microchambers, however sample heating is a disadvantage\textsuperscript{85}. Electrical fields may also be used to trap and manipulate even charge-neutral cells, a technique called dielectrophoresis, invented in the 1950s\textsuperscript{11}. The resolution is
low in comparison with optical methods and the system is non-flexible because the electrodes are placed in fixed positions. Disadvantages for cell culturing include fluid heating and physical stress on cells. In a recent study, dielectrophoresis was used to capture and pull down one bacterial cell into each well of a chip, followed by lysis to release the cellular contents, and the technique can also be used for particle focusing in a microchip flow cytometer. Acoustic waves can be used for manipulating and capturing cells in cages in chips, and with high reported cell viabilities the technique has been used to cause controlled cell aggregation in the center of microwells for cell-cell interaction studies. Sound waves have also been utilized for cell depletion of a blood sample to enable analysis of the plasma fraction for prostate-specific-antigen detection using an antibody array.

This chapter has described tools for picking and placing cells for culturing and/or analysis, and applications of some of these methods are demonstrated in Present Investigations. In nucleic acids research, which is the focus of this thesis, the next step often involves some kind of amplification, and this will be described in the following chapter.
This chapter is about ways of getting what you want from a DNA sample. In many biomedical settings work often involves analysis and comparison of highly variable regions or positions in the genome, or comparison of candidate genes between healthy and diseased individuals. In these cases it is more efficient to direct the focus of the sequencing towards a small fraction of the genome rather than to sequence it all. With this approach, the available sequencing capacity can be used to obtain high coverage of the selected genes, or to increase the number of assayed individuals. There are a number of approaches for decreasing the complexity of the genome by enriching or amplifying targets of interest from the three billion basepairs that are present in the haploid human genome and that make up the background of the sample. The method of choice will depend on the number of genes or regions as well as the number of samples included in the study.

When starting from scarce material such as single, or a few, cells, amplification in some form is usually necessary either as a stand-alone procedure or as a pre-amplification step. Retaining the relative quantities of components while amplifying is important for obtaining accurate information on, e.g., copy number variation in whole genome analysis, and also because a skewed representation of the original sample would impact on the amount of sequence data needed for coverage of the complete genome. Amplifying a DNA sample while introducing minimal bias and errors is therefore of the essence. Here, the most important methods for target selection, as well as for selective and global amplification of DNA samples, will be described.
3.1. Targeted amplification using polymerase

DNA polymerases are important enzymes in the cell that catalyze the incorporation of nucleotides into a growing strand of DNA, while using another strand as a template. These enzymes are widely used in different amplification strategies in laboratories worldwide in order to obtain multiple copies of a given target region. Different polymerases exhibit different features, especially some have proofreading activity - i.e. an incorrectly incorporated nucleotide can be removed from the end of the extended strand - and some are thermo-stable, meaning that they can work at elevated temperatures.

3.1.1. Polymerase Chain Reaction (PCR)

It all started with the Polymerase Chain Reaction (PCR), that made it possible for the first time to selectively amplify a piece of DNA in vitro, thus allowing the remarkable reduction of complexity from the 3 billion bases, that make up the human genome, down to the size of an ordinary PCR product i.e. a few hundred bases. PCR is now used routinely in almost every biology lab and has enabled countless new techniques and findings since it was first invented.

The history of the invention is interesting in itself. In 1971, Khorana and coworkers described most of the features that are essential to what we know today as PCR, but the article lacked experimental evidence and received very limited attention. Fourteen years later, the method described by Kleppe et al. was once more published, this time, in a more extensive way, by Kary Mullis and colleagues, who later received the Nobel prize for this work. When a thermostable polymerase was introduced, the potential of the reaction really escalated, since higher temperature for annealing and extension markedly increased the specificity, sensitivity and yield of the reaction. Since then, new and improved polymerases with markedly decreased error rates have been isolated or engineered and PCR has become an indispensable tool in most biological sciences.

3.1.2. Real-time, Digital, Single Cell and Long-range PCR

Due to the exponential nature of the amplification in PCR, the amount of PCR product obtained at the end of the reaction does not reflect the initial number of target molecules in a sample, hence end point measurements can only distinguish a positive sample from a negative one. For quantification purposes, real-time PCR may be used where a reporter generates a fluorescent signal that indicates the exponential accumulation of product as the reaction proceeds. For absolute quantification a standard curve for that particular target can be used. Reporters may be fluorescent dyes such as SYBR green, which fluoresce when bound to dsDNA, with the inherent limitation that primer-dimers also give fluorescent signals. For assays based on dsDNA-binding fluorophores, melting curve analysis is a useful way to obtain information on product length and hence primer-dimer formation in the PCR reaction.

Another option is to use labeled oligonucleotides that show a change in fluorescent signal upon target sequence binding. TaqMan probes are the most common detection probes used in real-time PCR today. They are designed to hybridize within the target region and are dually labeled with a 5’ fluorophore and a quencher at the 3’ end. When the TaqMan
probe is hybridized to a template molecule it forms a suitable substrate for the 5’ exonuclease activity of the Taq polymerase, which hydrolyses the probe thus releasing the fluorophore from the quencher, and so a fluorescent signal is emitted. The process repeats in every cycle hence the fluorescence signal reflects the amount of specific product created in the reaction\(^{9,10}\). Another example is Molecular Beacons, which are dually labeled with a fluorophore and its quencher at the two ends; when not bound to the target DNA, the molecule forms a stem-loop structure, which causes quenching of the fluorophore. However, when a Molecular Beacon is bound to its target sequence, quencher and fluorophore are sufficiently far away from each other for fluorescence to occur and an amplification signal is emitted\(^{11}\). The advantage of specific detection probes is that target product formation may be monitored instead of the total mass of dsDNA in the reaction.

Another advantage of using probes for detection is the possibility of amplifying multiple targets in one sample using color-coded probes\(^{12}\). Real-time PCR on cDNA has for a long time been the ‘gold standard’ in expression profiling, where quantification of a particular transcript is based on the number of cycles required for the dye to reach a threshold. The method has also been used for single cell gene expression analysis by immobilization of cDNA onto reusable beads\(^{13}\). Real-time PCR has also proven useful for quantification of libraries for high-throughput sequencing, where knowledge of the number of amplifiable molecules is important before a loading a sample onto a sequencer\(^{14,15}\).

Digital PCR, as implied by the name, relies on binary positive/negative results at the endpoint of the PCR and the reaction is normally done on a massive scale with up to one million reactions in parallel\(^{16}\). By diluting a sample so that each chamber in a microfluidic chip is expected to contain one DNA molecule or less, an absolute readout of the number of amplifiable fragments in a sample per unit volume, as indicated by a fluorescent signal, is obtained\(^{17}\). Analogous to real-time PCR, digital PCR can be used for expression analysis of transcription factors expressed at low levels in single cells\(^{18}\), quantification of single nucleotide polymorphisms and chromosomal copy number aberrations\(^{19}\) and library titration for high-throughput sequencing\(^{20}\). Digital PCR may be slightly more sensitive and better at quantifying low copy number templates than real-time PCR\(^{21}\), however for many applications digital PCR is overkill, and the inaccessibility of the advanced instrumentation needed for digital PCR as compared to real-time PCR makes the latter the method of choice for general users today.

PCR from single cells was first reported in the 1980s and was used early on for genetic analysis of sperm cells\(^{22}\) and to investigate rearrangements of the immune cells that give rise to a diverse immune system\(^{23}\). Increased sensitivity is commonly obtained through decreased reaction volumes, careful optimization of reaction conditions and increased number of reaction cycles\(^{24}\). In the field of prenatal diagnostics the method has played an important role in pre-implantation genetic analysis of a number of X-linked diseases and single gene disorders (e.g. cystic fibrosis), and the approach is in clinical use in several countries today\(^{25}\). Miniaturization has been frequently utilized in recent years to achieve increased sensitivity in amplification from minute amounts of DNA\(^{26,27}\). The level of amplification and the high sensitivity of the analysis that often follows amplification of minute amounts of nucleic acids increases the risk of amplifying background contaminants from sample and reagents. Approaches to address this issue include UV treatment and/or filtering of reagents, and detergent and hypochlorite cleaning of tools and glassware\(^{28,29}\).
Long-range PCR (LR-PCR) was, as the name suggests, developed to increase the length of the amplified fragment, which was previously limited to about 5 kb, by using a two-polymerase system to obtain optimal levels of processive polymerase activity while adding proofreading. Increased primer lengths, lower pH, and the use of additives such as glycerol that would enable shorter denaturation times and prevent nicking are strategies commonly applied to obtain amplification of fragments up to 42 kb long. Fragmentation of LR-amplified fragments followed by library preparation for massive parallel sequencing (MPS) is commonly used and applications of this as well as common amplicon sequencing by MPS will be discussed in more detail in Chapter 6.6.2 (Targeted resequencing).

### 3.1.3. Multiplexity limitations of PCR and some ways around them

For most applications, increased throughput and multiplexing is desirable in order to reduce turn-around time, save money and minimize the amount of input material required. PCR amplification of single loci or fragments is generally robust; however when it comes to multiplexing, limitations arise. Several factors, such as primer-dimer formation, unspecific product formation and amplification bias, limit the degree to which multiplexing can be used.

As the number of targets amplified in parallel increases, so does the risk that two primers will cause primer-dimer formation by priming each other through 3′-end complementarity. Primer-dimers are problematic since due to their short length they are efficiently amplified and because of the competitive nature of the PCR reaction they therefore threaten to outcompete the target for polymerase and reagents, thus taking over the whole reaction. Several protocols addressing primer-dimer formation have been published that point towards careful primer design, titration of primer concentrations, optimization of cycling conditions, increased amount of template, and lower number of amplification cycles as ways to avoid primer-dimer formation.

Another difficulty of multiplex PCR, which is harder to tackle, is unspecific product formation. When targeting several regions in a PCR reaction and hence increasing the number of primers in the tube, the chance that two primers will hybridize to unintended sequences thus creating nonspecific amplification products, increases steeply, making primer design of a highly multiplex reaction a walk in a mine field. Attempts to address this issue have been made by introducing nested PCR, where two sets of primers are used in two consecutive PCR amplification steps. Initial amplification of a region containing the target area is followed by amplification using a second pair of primers, which hybridize to the first PCR product inside the first primer pair and thus amplify the target sequence.

Nested PCR has been an effective approach for lower degrees of multiplexity but for higher levels of multiplexity the above-mentioned issues remain unresolved.

Yet another issue of multiplex PCR is a result of the competitive nature of PCR amplification, which causes unequal amplification of different target regions and even loss of some targets when the degree of multiplexity is increased. These differences may occur due to different primer hybridization efficiencies, product lengths or target sequence secondary structures. Due to the exponential nature of the amplification, differences in the yields of unequally amplified fragments, as well as errors introduced in early cycles, are

---

Julia Sandberg

15
propagated in every cycle. To some extent thorough optimization of reaction conditions and primer design can discourage artifacts from occurring and PCR reactions of up to 1000plex have been reported\textsuperscript{18}. However, for most target regions this would probably not be possible, hence multiplex PCR of more than 10-20 loci is rare\textsuperscript{10,18,19}.

Several approaches have been employed to specifically amplify many targets in parallel by PCR. Introduction of universal amplification tags is a common theme in amplification since it enables amplification of an entire complex library using a single primer pair. The approach can be used in a two-step PCR where each primer carries 3’ universal tail sequences, which are introduced in the first amplification cycles, after which amplification is carried out using one universal primer pair only. This method was first described in 1996, when it was demonstrated by 26-plex amplification\textsuperscript{10}. In a similar two-stage PCR protocol, reduced primer-dimer formation was approached by designing the universal primer pair so that potential primer-dimers were folded into an unamplifiable stem-loop structure\textsuperscript{11}. The concept of introducing universal amplification handles is a common theme in amplification and with the advent of high-throughput sequencing, the concept has become a routine method for amplification.

Yet another method of carrying out massive numbers of single-locus PCRs by physical separation in an emulsion, which was briefly described in Chapter 2. Through compartmentalization of each primer pair into separate emulsion droplets, competition between different primer pairs is prohibited and each PCR reaction is allowed to saturate without competition\textsuperscript{13}. Primer pairs for each target are separately introduced into droplets, which are then pooled to generate a primer library after which gDNA is introduced into the droplets for emulsion PCR\textsuperscript{12}. The approach has been demonstrated by amplification of 4000 targets with a specificity of 98%\textsuperscript{13} and in a test for future parents to determine carriership of a number of recessive genetic disorders\textsuperscript{12}. Emulsions with up to 20,000 different primer pairs designed towards the customer’s regions of interest are now available for purchase through RainDance Technologies\textsuperscript{18}. Emulsion PCR is used for sample preparation in several massive sequencing techniques and will be described in more detail in chapter 4.

Another approach involves complete or partial physical separation of PCR reactions to avoid competition and cross-reaction of multiple primers in the same reaction. Immobilizing PCR primers on a solid support in solid-phase PCR is a way to limit primer-dimer formation and background amplification by physical isolation of each primer pair in a gel matrix or on the surface of beads or chips\textsuperscript{14,15}. The efficiency of solid-phase PCR is usually lower than that of in-solution reactions. A combination of solid-phase PCR and “universal tailing” has been used, in which introduction of a general PCR primer pair in the initial cycles on a solid support is followed by a more efficient in-solution PCR amplification using a universal primer pair. The method is named MegaPlex PCR, and it has been used on 75 target regions in parallel\textsuperscript{14,16}. Bridge amplification, where universal primers are immobilized on a chip surface is used in preparation for sequencing by Illumina and will be described in more detail in chapter 4. PCR reactions may also be separated physically in the chambers of a microchip. The Access Array platform, which was recently released by Fluidigm, enables 48 separate amplifications of choice (PCR or LR-PCR) to be carried out in nanoliter volumes\textsuperscript{14}.

Yet another method of carrying out massive numbers of single-locus PCRs by physical separation is in an emulsion, which was briefly described in Chapter 2. Through compartmentalization of each primer pair into separate emulsion droplets, competition between different primer pairs is prohibited and each PCR reaction is allowed to saturate without competition\textsuperscript{13}. Primer pairs for each target are separately introduced into droplets, which are then pooled to generate a primer library after which gDNA is introduced into the droplets for emulsion PCR\textsuperscript{12}. The approach has been demonstrated by amplification of 4000 targets with a specificity of 98%\textsuperscript{13} and in a test for future parents to determine carriership of a number of recessive genetic disorders\textsuperscript{12}. Emulsions with up to 20,000 different primer pairs designed towards the customer’s regions of interest are now available for purchase through RainDance Technologies\textsuperscript{18}. Emulsion PCR is used for sample preparation in several massive sequencing techniques and will be described in more detail in chapter 4.
3.1.4. Isothermal targeted amplification strategies

In living cells, DNA helicase plays an important role in DNA replication by separating two complementary DNA strands so that the polymerase can start copying the template. In vitro, the enzyme can be used in a reaction analogous to PCR but where heat denaturation has been replaced by enzymatic DNA separation. This method is called Helicase Dependent Amplification (HDA) and it has the great advantage of being isothermal, which makes the demands on equipment low. HDA is therefore a natural choice in the field of development of point-of-care diagnostics devices\(^{139}\). Detection and readout in HDA can be done in real-time as described for real-time PCR, or by using lateral flow or biosensors\(^{139,140}\). By using a thermostable helicase, the reaction can be done at elevated temperature (thermophilic helicase dependent reaction, HDA) with increased specificity as a result\(^{140}\). HDA protocols have been developed for detection of several pathogens including Herpes virus\(^{139}\), gonorrhoea\(^{141}\) and tuberculosis\(^{142}\). The product size of ordinary HDA is limited to 400 bp, presumably because of re-annealing of separated strands before the primers have been extended, or because the helicase interferes with the activity of the polymerase by knocking it off the template. By fusing a polymerase and a helicase together into a helimerase, amplification of products of up to 2.3kb has been demonstrated\(^{143}\).

3.2. Using ligation for increased specificity

DNA ligases are enzymes that can join oligonucleotides that are hybridized adjacent to each other on a DNA strand by forming a phosphodiester bond, provided that the nucleotides at the junction are perfectly matched to their target\(^{144}\). By introducing a selective ligation step into a polymerase-based amplification, an additional level of specificity can be obtained\(^{145}\).

3.2.1. Ligase chain reaction (LCR)

In a reaction similar to PCR, thermostable ligases can be used in amplification approaches with the aim of reducing sample complexity while amplifying one or a few target regions. Here, amplification is done by ligases joining two primers that are hybridized adjacent to each other on the same strand of the target DNA. By placing the last base of the upstream primer over a variable position, LCR allows discrimination of SNPs, since the enzyme will only join perfectly matched primers. In total, four primers are used, hybridizing to either strand in pairs, and as more template is created in each cycle the amplification is exponential\(^{146}\). Although LCR is known to be highly specific, background ligation still occurs. To address this issue Gap-LCR was developed. As the name suggests, a gap between the two primers of each pair is introduced and so a template-dependent extension is required to fill in the gap, before the primers can be joined. By restricting the number of different nucleotides added in the reaction, specificity can be increased further. Readout has been achieved using gel electrophoresis and labeled primers or nucleotides; for real-time monitoring, fluorescent dyes or labeled probes can be used. LCR and variants thereof, most prominently real-time Gap-LCR, have proven useful for genotyping pools of genomic
DNA, and are reportedly good at evaluating the frequency of rare SNPs in a cell population\cite{44}. A variant of the method has recently been used to detect and quantify a disease-causing mutation in a genomic DNA pool\cite{45}. The relatively low amplification efficiency of ligation-based amplification may be addressed by an initial PCR step to achieve greater sensitivity in the assay\cite{46}.

A linear version of this technique, ligase detection reaction (LDR), can be done using one pair of primers hybridizing adjacent to each other on the target strand\cite{47}. In a recently published article, a PCR/LDR approach was used to detect rare point mutations in a human DNA sample. An initial PCR step was followed by LDR with biotinylated 3' primers and 5' primers containing a tag sequence. Detection was done by binding of products to tag-complementary spots and visualizing them by streptavidin-coupled silver staining\cite{48}. Generally, ligation is less efficient at each step than polymerization, and the amount of material obtained from ligation-based strategies is a limitation of these methods.

3.2.2. Multiplex Ligation Dependent Probe Amplification (MLPA)

A way to overcome the low product output of ligation-dependent amplification strategies is to follow them up with a PCR using a universal primer pair. Ligation of one unique primer pair per target, primers that hybridize adjacent to each other, in a linear fashion, allows introduction of universal amplification primers into the 5' ends of ligated primer products. These universal primers enable multiplex PCR amplification of successfully ligated probes. Readout is commonly done by capillary electrophoresis and the unique length of each amplified fragment identifies the targeted regions. The reaction is called multiplex ligation dependent probe amplification (MLPA) and its strength is that by using one universal PCR primer pair the competitive nature of PCR in multiplex amplification is decreased substantially. Since the method is based on ligation, it can also be used for detection of single nucleotide changes. Another benefit is the linearity of the initial ligation step, which enables comparative quantification of the different target regions. This makes MLPA useful for analysis of copy number variation, as demonstrated for up to 50 genomic DNA sequences from as little as 20 ng of gDNA\cite{49}, it and has also been used to detect chromosomal abnormalities in melanoma\cite{50}.

3.2.3. Golden Gate

A very common method for multiplex amplification for genotyping purposes, although not for massive sequencing, is the Golden Gate assay. This assay is based on Allele-Specific Extension\cite{51}, in which two primers that hybridize upstream of, and end in, each SNP position, are used, matching the two different genotypes at that locus. The two allele-specific primers contain a unique amplification handle in their 5' end and the perfectly-matched primer will be extended and then ligated onto a third annealed primer. The third probe is a locus-specific oligonucleotide since it contains a SNP-specific address tag and a universal amplification handle in its 3' end. PCR is carried out with three universal primers, two color-coded forward primers matching one SNP-specific sequence in the Allele-specific oligonucleotide handle each, and one reverse primer that is complementary to the universal handle at the locus-specific oligonucleotide. Analysis is done by hybridizing the sample to a random array of beads with different address tags that are tacked into wells on a chip\cite{52}.

DNA, and are reportedly good at evaluating the frequency of rare SNPs in a cell population\cite{44}. A variant of the method has recently been used to detect and quantify a disease-causing mutation in a genomic DNA pool\cite{45}. The relatively low amplification efficiency of ligation-based amplification may be addressed by an initial PCR step to achieve greater sensitivity in the assay\cite{46}.

A linear version of this technique, ligase detection reaction (LDR), can be done using one pair of primers hybridizing adjacent to each other on the target strand\cite{47}. In a recently published article, a PCR/LDR approach was used to detect rare point mutations in a human DNA sample. An initial PCR step was followed by LDR with biotinylated 3' primers and 5' primers containing a tag sequence. Detection was done by binding of products to tag-complementary spots and visualizing them by streptavidin-coupled silver staining\cite{48}. Generally, ligation is less efficient at each step than polymerization, and the amount of material obtained from ligation-based strategies is a limitation of these methods.

3.2.2. Multiplex Ligation Dependent Probe Amplification (MLPA)

A way to overcome the low product output of ligation-dependent amplification strategies is to follow them up with a PCR using a universal primer pair. Ligation of one unique primer pair per target, primers that hybridize adjacent to each other, in a linear fashion, allows introduction of universal amplification primers into the 5' ends of ligated primer products. These universal primers enable multiplex PCR amplification of successfully ligated probes. Readout is commonly done by capillary electrophoresis and the unique length of each amplified fragment identifies the targeted regions. The reaction is called multiplex ligation dependent probe amplification (MLPA) and its strength is that by using one universal PCR primer pair the competitive nature of PCR in multiplex amplification is decreased substantially. Since the method is based on ligation, it can also be used for detection of single nucleotide changes. Another benefit is the linearity of the initial ligation step, which enables comparative quantification of the different target regions. This makes MLPA useful for analysis of copy number variation, as demonstrated for up to 50 genomic DNA sequences from as little as 20 ng of gDNA\cite{49}, it and has also been used to detect chromosomal abnormalities in melanoma\cite{50}.

3.2.3. Golden Gate

A very common method for multiplex amplification for genotyping purposes, although not for massive sequencing, is the Golden Gate assay. This assay is based on Allele-Specific Extension\cite{51}, in which two primers that hybridize upstream of, and end in, each SNP position, are used, matching the two different genotypes at that locus. The two allele-specific primers contain a unique amplification handle in their 5' end and the perfectly-matched primer will be extended and then ligated onto a third annealed primer. The third probe is a locus-specific oligonucleotide since it contains a SNP-specific address tag and a universal amplification handle in its 3' end. PCR is carried out with three universal primers, two color-coded forward primers matching one SNP-specific sequence in the Allele-specific oligonucleotide handle each, and one reverse primer that is complementary to the universal handle at the locus-specific oligonucleotide. Analysis is done by hybridizing the sample to a random array of beads with different address tags that are tacked into wells on a chip\cite{52}.
Determination of the identity (i.e. what SNP does the address tag code for) of each bead in the array, is carried out by repeated hybridization of fluorescently labeled probes, imaging and probe removal, to obtain a combinatorial code representing the identity of each bead[15]. Decoding is followed by hybridization of amplified sample for genotype interpretation. One color indicates a sample from a homozygote, and dual coloring signifies a heterozygous SNP[12]. Random bead arrays are available through Illumina and presently supports 86-3072 SNPs per sample and 12 to 96 samples in parallel[14].

3.2.4. Tri-nucleotide Threading (TnT)

Tri-nucleotide threading (TnT) is an amplification method that was developed at the Royal Institute of Technology (KTH) in order to increase the degree of multiplexity that is possible[15]. In an initial linear amplification step, polymerase extension bridges a gap between two primers that are hybridized unto the same strand and ligation seals the two probes, of which one is biotinylated. Cleanup, using streptavidin-coupled paramagnetic beads, is followed by universal amplification to amplify ligated threads in parallel. By designing the primer locations so that each gap contains only three types of nucleotides and adding only these to the reaction, the specificity of the assay is increased further. The technique has been used for expression profiling[14] and genotyping from 1200 cells[12] using chip readout; for analysis of short tandem repeats by capillary electrophoresis[12]; and for determination of allele frequencies of two sets of 147 SNPs in a cohort of hundreds of individuals by Roche/454 sequencing[14].

3.3. Selective Circularization

3.3.1. Padlock probes

By joining two specific ligation primers using a linker that allows primer folding, a ligation-based circularization reaction can be obtained with only one probe per target sequence. This ingenious design decreases the number of potential byproducts that can be formed with increasing multiplexity. Furthermore, as the ends are joined together while hybridized to the target sequence, the circular probe is interlocked with its target, hence the name Padlock probes. This allows rigorous washing schemes to be used in situ while containing the probes[14][12][3]. Padlock probes were first described in the literature in 1994 and since then the technique has been refined and utilized in various genotyping and mRNA studies[14]. Using exonuclease degradation of linear non-ligated Padlock probes, cross-reactivity can be kept to a minimum[14]. Amplification of the circularized probes can be done by PCR[14] or Rolling Circle Amplification (RCA). The latter method takes advantage of the high-processivity and strand displacement activity of the enzyme phi29 to create a very long DNA product containing a number of copies of the circularized fragment. Amplification products may be fluorescently labeled for in situ detection or chip readout[14][12] or labeled with magnetic nano-sized beads and detected based on response to a magnetic field in a portable device[14]. Padlock probes have been used for in situ detection and genotyping of mitochondrial DNA[14] and mRNA molecules[14] as well as for analysis of infection in cattle[14] and more recently for CpG island methylation analysis[14].
3.3.2. Molecular Inversion probes (MIP)

By introducing a gap between adjacent primer ends, which required polymerase fill in before ligation, Padlock probes were transformed into Molecular Inversion probes (MIP)\(^{(14)}\). Release of circularized products from the template by cleavage at a unique site in the linker region is followed by PCR amplification with primers that hybridize to the linker part of the circularized fragments. MIPs have been used to capture large numbers of target regions in parallel, thus allowing analysis of SNPs in genomic DNA, and identification of RNA editing sites\(^{(16)}\) or alternative splicing events\(^{(16,17)}\). The methodology has received extensive attention due to its impressive multiplexing capabilities, high specificity and low input DNA requirements\(^{(16)}\). The initial number of 1000 targets in one tube\(^{(16)}\) has since been greatly exceeded through the capture of 10,000 and 50,080 target regions\(^{(16,17)}\). Direct sequencing of MIPs by the introduction of Illumina handles in the final PCR step has been carried out in a slightly modified protocol requiring sub-micrograms of input material. High sensitivity was demonstrated by ~50% of all targets being captured per reaction\(^{(17)}\). In a recently optimized protocol called Long-Padlock, the maximum length of the capture regions was increased from the 191 bp achievable in the ordinary MIP approach to 546 bp. As is commonly seen for enzymatic amplification strategies, the limitation of MIPs appears to lie in bias of capture\(^{(17)}\).

In methods where complex oligonucleotide libraries are required cost is usually a limiting factor. For several liquid-phase capture methods such as MIPs, this problem has been solved by using probe release from a programmable microarray to obtain a vast number of oligonucleotide probes cheaply\(^{(18)}\). Microarrays from vendors such as Affymetrix, Agilent, and NimbleGen contain up to several million of customizable primers between 25 and 200 bp long\(^{(17)}\).

3.3.3. Selector

Selector is a highly specific method, based on circularization, in which the actual targets, rather than the probes, are circularized. Starting with digestion of gDNA, using one or several restriction enzymes, selector probes are hybridized to specific ends of the target; fragments of interest are then folded and circularized through a process of gap fill and ligation. As for previously described strategies involving circularization, uncircularized or erroneously ligated linear fragments are digested by exonuclease\(^{(17)}\) before the final amplification step, which may be PCR with primers specific for the common part of the circularized fragments\(^{(17,18)}\) or RCA\(^{(14,17)}\). Recently, the latter protocol has been shown to reduce the enrichment bias in an experiment when capturing 501 exons using 1883 selector probes in a single reaction\(^{(17)}\).

3.3.2. Molecular Inversion probes (MIP)

By introducing a gap between adjacent primer ends, which required polymerase fill in before ligation, Padlock probes were transformed into Molecular Inversion probes (MIP)\(^{(14)}\). Release of circularized products from the template by cleavage at a unique site in the linker region is followed by PCR amplification with primers that hybridize to the linker part of the circularized fragments. MIPs have been used to capture large numbers of target regions in parallel, thus allowing analysis of SNPs in genomic DNA, and identification of RNA editing sites\(^{(16)}\) or alternative splicing events\(^{(16,17)}\). The methodology has received extensive attention due to its impressive multiplexing capabilities, high specificity and low input DNA requirements\(^{(16)}\). The initial number of 1000 targets in one tube\(^{(16)}\) has since been greatly exceeded through the capture of 10,000 and 50,080 target regions\(^{(16,17)}\). Direct sequencing of MIPs by the introduction of Illumina handles in the final PCR step has been carried out in a slightly modified protocol requiring sub-micrograms of input material. High sensitivity was demonstrated by ~50% of all targets being captured per reaction\(^{(17)}\). In a recently optimized protocol called Long-Padlock, the maximum length of the capture regions was increased from the 191 bp achievable in the ordinary MIP approach to 546 bp. As is commonly seen for enzymatic amplification strategies, the limitation of MIPs appears to lie in bias of capture\(^{(17)}\).

In methods where complex oligonucleotide libraries are required cost is usually a limiting factor. For several liquid-phase capture methods such as MIPs, this problem has been solved by using probe release from a programmable microarray to obtain a vast number of oligonucleotide probes cheaply\(^{(18)}\). Microarrays from vendors such as Affymetrix, Agilent, and NimbleGen contain up to several million of customizable primers between 25 and 200 bp long\(^{(17)}\).
Targeted amplification strategies

- Tailed PCR
- Nested PCR

Ligase Chain Reaction
- 2 primer pairs
- Exponential amplification
- Tri-nucleotide Threading
- Linear amplification
- Universal PCR

Gapped Ligase Chain Reaction
- 2 primer pairs
- Exponential amplification

Multiplex Ligation Dependent Probe Amplification
- 1 primer pair
- Universal PCR

Molecular Inversion probes

Padlock probes

Selector

digested DNA

Cleavage
3.4. Targeted capture by hybridization

Assays based solely on hybridization offer by far the highest multiplexity level and can be done either on a solid support or in solution. Custom-designed planar arrays can be used to enrich for regions of interest from a linker-ligated DNA library, through a simple but protracted hybridization – washing – release by heat protocol, which is followed by universal PCR amplification of enriched library material, prior to sequencing. This was first demonstrated in 2007171,177. One way of increasing the number of samples enriched in array-based strategies is by barcoding individual DNA libraries prior to enrichment178,179; alternatively, arrays may be reused177.

Capture may also be done in solution using biotinylated DNA or RNA180 oligonucleotides and binding to streptavidin-covered beads, target elution and universal PCR amplification. Solution-based capture has lower reagent costs, a lower DNA requirement (i.e. 10-20 µg gDNA for an array171,177 0.5-3 µg in solution181) and is easily automated181. Hence, in-solution capture is currently the method of choice and several vendors, such as Agilent and NimbleGen, provide kits targeting whole exomes, or parts thereof, and also offer custom made kits. Uniformity is generally higher in hybrid capture methods than in amplification based methods181,185.

The method allows identification of insertions, deletions and fusion genes as demonstrated in a recent paper focused on leukemia-associated genes182 and capture has been done on formalin fixed paraffin embedded material184. Since the advent of the technique, many interesting studies using targeted resequencing by hybridization capture have emerged and these will be described in more detail in Chapter 4.6.2 (Targeted resequencing).

3.4. Targeted capture by hybridization

Assays based solely on hybridization offer by far the highest multiplexity level and can be done either on a solid support or in solution. Custom-designed planar arrays can be used to enrich for regions of interest from a linker-ligated DNA library, through a simple but protracted hybridization – washing – release by heat protocol, which is followed by universal PCR amplification of enriched library material, prior to sequencing. This was first demonstrated in 2007171,177. One way of increasing the number of samples enriched in array-based strategies is by barcoding individual DNA libraries prior to enrichment178,179; alternatively, arrays may be reused177.

Capture may also be done in solution using biotinylated DNA or RNA180 oligonucleotides and binding to streptavidin-covered beads, target elution and universal PCR amplification. Solution-based capture has lower reagent costs, a lower DNA requirement (i.e. 10-20 µg gDNA for an array171,177 0.5-3 µg in solution181) and is easily automated181. Hence, in-solution capture is currently the method of choice and several vendors, such as Agilent and NimbleGen, provide kits targeting whole exomes, or parts thereof, and also offer custom made kits. Uniformity is generally higher in hybrid capture methods than in amplification based methods, and this affects the number of reads required to obtain sufficient information on all target regions. However the specificity is low in comparison with that of amplification-based methods181,185.

The method allows identification of insertions, deletions and fusion genes as demonstrated in a recent paper focused on leukemia-associated genes182 and capture has been done on formalin fixed paraffin embedded material184. Since the advent of the technique, many interesting studies using targeted resequencing by hybridization capture have emerged and these will be described in more detail in Chapter 4.6.2 (Targeted resequencing).
3.5. Decisions, decisions. What to choose?

Several studies comparing selected enrichment strategies have been carried out recently. Generally, enzymatic amplification strategies have higher specificity and lower uniformity than hybrid-capture methods. It appears that those regions which are GC rich and/or occur frequently in repetitive regions of the genome are difficult to capture whichever of the methods is used; they may also tend to get lost in downstream processes such as sequencing and mapping. Thus, there appears to be little to gain by combining more than one enrichment method.

The choice of enrichment strategy depends on the application. For amplifying and analyzing a small number of target regions, PCR or multiplex PCR-derived methods may be useful. PCR enrichment in emulsion, which is commercially available, allows a moderate number (>20,000 PCR reactions) of targets and samples. For a larger number of targets in combination with a large number of samples, methods which scale and multiplex effectively, such as selection by circularization, e.g. MIP and solution phase hybrid capture, of which the latter is commercially available for selection (MIPs are commercially available only for genotyping purposes), will serve the researcher well for reducing the complexity of the initial sample.
3.6. **Global amplification of genomic DNA**

In contrast to the targeted amplification and selection methods presented previously, the aim of global amplification of a nucleic acid sample is to maintain the complexity of the starting material unchanged while increasing the amount of nucleic acid. The approach can be used as a preparative step or as a stand-alone protocol prior to, for example, sequencing or microarray analysis.

3.6.1. **PCR-based amplification strategies**

Most early global amplification strategies utilized PCR amplification together with partly or fully random primers. Degenerate Oligonucleotide-Primed PCR (DOP-PCR) uses primers with a defined sequence at each end and a random hexamer sequence in the middle, whereas in Primer Extension Preamplification PCR (PEP-PCR) completely random primers are used for amplification. Despite protocol modifications since the methods were first developed, both techniques have shown limitations in coverage when amplifying scarce material.

The recurrent trick of universal amplification tags can also be applied to global amplification. Initial tag incorporation can be carried out using primers with random 3' ends and a 5' tag sequence either by a few initial PCR cycles or through linear isothermal amplification to obtain overlapping copies of the initial DNA molecule. Another way is ligation of amplification tags to gDNA that has been fragmented physically by shearing or cut by restriction enzymes. A proprietary variant of the approach is available as a commercial kit under the name GenomePlex (Sigma-Additech, MO, USA) and has been used in work reported in a number of publications for single cell analysis by CGH arrays, SNP arrays and massive sequencing.

PCR-based WGA methods are generally affected by secondary DNA structures, which can cause polymerase slippage and dissociation of enzyme from template; this may result in poor coverage, short product length and uneven representation of the template due to amplification bias and other amplification artifacts such as formation of chimeras. Despite this, DOP-PCR and PEP-PCR and methods based on general amplification handles are still used today, for example in the development of approaches such as pre-implantation genetic diagnosis.

3.6.2. **Multiple Displacement Amplification (MDA)**

A major milestone in the field of molecular biology was the discovery of Phi29, a polymerase that showed extraordinary strength when holding on to the DNA template during synthesis. Through its remarkable processivity in combination with random priming, the displacement activity of Phi29 leads to a hyper-branched structure of amplified material with an average product length of >10 kb. The method is called Multiple Displacement Amplification (MDA) and it can be used to isothermally amplify the human genome 1000-fold. Although amplification bias is lower than for PCR-based WGA methods, non-specific synthesis, either from DNA contamination or from endogenously generated DNA such as primer-dimers, is an issue for this means of amplification too.

3.6. **Global amplification of genomic DNA**

In contrast to the targeted amplification and selection methods presented previously, the aim of global amplification of a nucleic acid sample is to maintain the complexity of the starting material unchanged while increasing the amount of nucleic acid. The approach can be used as a preparative step or as a stand-alone protocol prior to, for example, sequencing or microarray analysis.

3.6.1. **PCR-based amplification strategies**

Most early global amplification strategies utilized PCR amplification together with partly or fully random primers. Degenerate Oligonucleotide-Primed PCR (DOP-PCR) uses primers with a defined sequence at each end and a random hexamer sequence in the middle, whereas in Primer Extension Preamplification PCR (PEP-PCR) completely random primers are used for amplification. Despite protocol modifications since the methods were first developed, both techniques have shown limitations in coverage when amplifying scarce material.

The recurrent trick of universal amplification tags can also be applied to global amplification. Initial tag incorporation can be carried out using primers with random 3' ends and a 5' tag sequence either by a few initial PCR cycles or through linear isothermal amplification to obtain overlapping copies of the initial DNA molecule. Another way is ligation of amplification tags to gDNA that has been fragmented physically by shearing or cut by restriction enzymes. A proprietary variant of the approach is available as a commercial kit under the name GenomePlex (Sigma-Additech, MO, USA) and has been used in work reported in a number of publications for single cell analysis by CGH arrays, SNP arrays and massive sequencing.

PCR-based WGA methods are generally affected by secondary DNA structures, which can cause polymerase slippage and dissociation of enzyme from template; this may result in poor coverage, short product length and uneven representation of the template due to amplification bias and other amplification artifacts such as formation of chimeras. Despite this, DOP-PCR and PEP-PCR and methods based on general amplification handles are still used today, for example in the development of approaches such as pre-implantation genetic diagnosis.
decrease in reaction volume from 50 µl to 60 nl, using a microfluidic chip, has been shown to improve specificity and lower the amplification bias, while maintaining the level of amplification, possibly due to reduced competition from background contaminants or to reduced damage to the DNA template\textsuperscript{38}, and similar a microfluidic setup has been used for evaluating nucleic acid contamination in commercially-available MDA reagents\textsuperscript{39}.

MDA has been used to prepare material from single cells for various applications such as CNV analysis by comparative genomic hybridization on, CGH\textsuperscript{39}, cloning and shotgun sequencing in a process called “ploning”\textsuperscript{40}, or as a pre-amplification step prior to PCR amplification of target regions\textsuperscript{41}. However, it has recently been shown that MDA introduces inversions but not translocations in the sequenced material\textsuperscript{42}. Through double WGA of FACS sorted cells, de novo sequencing and genome assembly of single bacteria has been done. Coverage varied 1000-fold however there was no correlation between coverage and GC content\textsuperscript{43}. In a recent publication MDA-assisted pre-implantation diagnosis was used to ensure the birth of a healthy child from parents carrying genes for a recessive kidney disease\textsuperscript{44}. Several kits for φ29-based amplification are commercially available.

### 3.6.3. Linear amplification techniques

Linear amplification strategies have the advantage of repeated copying the original template molecule over and over, instead of making more templates that in turn will be copied. In this way linear amplification avoids the problem of error propagation that are an unwelcome feature of exponential amplification strategies.

One method of linear amplification of dsDNA goes via RNA production using a modified variant of a widespread RNA amplification protocol\textsuperscript{45}. Briefly, by addition of a poly(T)-tail to the 3' end of the template, introduction of a T7 promoter sequence is enabled by extension from a polyA-tagged primer. This promoter sequence can then be utilized for linear amplification through in vitro transcription of the DNA template followed by reverse transcription to convert the resulting RNA amplification products back into DNA\textsuperscript{45-47}. The T7 promoter sequence can also be introduced by ligation\textsuperscript{48}. Single Primer Isothermal Amplification (SPIA) is another elegant strategy in which a general RNA sequence is introduced into the 5' ends of template fragments by using a chimeric random-DNA and tag-RNA primer that is either extended\textsuperscript{49} or introduced by ligation\textsuperscript{48}, thus incorporating priming sites for the forthcoming linear amplification. Chimeric RNA-DNA primers are then used to prime amplification by the strand-displacing polymerases. Cleavage of the RNA part of the introduced primer by RNase H exposes new priming sites, which enables template amplification in a linear fashion.

decrease in reaction volume from 50 µl to 60 nl, using a microfluidic chip, has been shown to improve specificity and lower the amplification bias, while maintaining the level of amplification, possibly due to reduced competition from background contaminants or to reduced damage to the DNA template\textsuperscript{38}, and similar a microfluidic setup has been used for evaluating nucleic acid contamination in commercially-available MDA reagents\textsuperscript{39}.

MDA has been used to prepare material from single cells for various applications such as CNV analysis by comparative genomic hybridization on, CGH\textsuperscript{39}, cloning and shotgun sequencing in a process called “ploning”\textsuperscript{40}, or as a pre-amplification step prior to PCR amplification of target regions\textsuperscript{41}. However, it has recently been shown that MDA introduces inversions but not translocations in the sequenced material\textsuperscript{42}. Through double WGA of FACS sorted cells, de novo sequencing and genome assembly of single bacteria has been done. Coverage varied 1000-fold however there was no correlation between coverage and GC content\textsuperscript{43}. In a recent publication MDA-assisted pre-implantation diagnosis was used to ensure the birth of a healthy child from parents carrying genes for a recessive kidney disease\textsuperscript{44}. Several kits for φ29-based amplification are commercially available.
Primer Extension Preamplification PCR (PEP-PCR)

Introduction of general amplification handles by extension

Isothermal extension or PCR

Linear amplification through in vitro transcription

Repeated priming and extension

Degenerate Oligonucleotide-primed PCR (DOP-PCR)

Introduction of general amplification handles by ligation

Isothermal extension or PCR

Linear amplification through in vitro transcription

Repeated priming and extension

Global amplification strategies

Multiple Displacement Amplification

Introduction of general amplification handles by ligation

Multiple Displacement Amplification

Introduction of general amplification handles by extension

Universal PCR

Ligate adaptors

Reverse transcription

Extension

3'-5' exonuclease activity

RNA

cDNA

Linear amplification with RNA primers

RNase H cleavage to free primer hybridization site

Linear amplification with RNA primers

RNase H cleavage to free primer hybridization site

RNA

cDNA

Linear amplification through in vitro transcription

RNase H cleavage to free primer hybridization site

RNase H cleavage to free primer hybridization site

RNA

cDNA

Global amplification strategies
Choosing a global amplification strategy

Amplification bias and allelic dropout are still issues in current WGA methods when starting from scarce material. Several comparative studies have been published demonstrating that MDA amplification gives a better representation of the initial sample in comparison with other amplification strategies\(^\text{197}\); however, others have found linker-PCR strategies to give a better\(^\text{34}\) or equal\(^\text{209}\) result in terms of dropout rate and amplification bias in comparison with MDA-based methods. The short product length of many PCR-based amplification strategies is a limitation for some applications, and this, together with the ease of use of the MDA approach, makes many labs turn to MDA. However the potential for amplifying degraded DNA is, for some applications, an obvious advantage of PCR-based approaches\(^\text{209}\).

This chapter has described various ways to treat a DNA sample in preparation for, e.g., sequencing analysis. The wide repertoire of methods available reflects the enormous importance of sample preparation and the limitations of available techniques. In the next chapter, current and forthcoming methods for Massively Parallel Sequencing will be described in more detail.
4. Massively parallel DNA sequencing

DNA sequencing can help us to answer fundamental questions about disease and human biology and thus contributes to both the saving of lives and increases human quality of life. The first organism whose genome was completely sequenced was a virus, bacteriophage \( \Phi X174 \). This was accomplished in 1977; an immense amount of work was required to finish that tiny genome of roughly 5000 basepairs.

Today, genomes 600,000 times larger than this (such as the human genome) can be sequenced in a matter of days. The technical revolution that has occurred in the field of gene technology during the last six years has made it possible to rapidly obtain tremendous amounts of sequence data, with a comparable reduction in cost. In fact, the amount of sequence data obtained per day doubles every year. The increase is so rapid that it outstrips Moore’s law (a proposal made in the 1960s, which stated that the amount of available computing power would double every two years for the foreseeable future). The rapid progress in this field has greatly increased the scope of experimentation beyond simply determining the order of bases in a gene or viral genome, making it possible to ask completely new questions.

This chapter is about massively parallel DNA sequencing (MPS). Current and emerging sequencing technologies will be presented, starting with some molecular and technical tools that are used in these platforms. Since today’s technologies were developed on the foundation of those that came before, the chapter begins with a brief recapitulation of the history of the field.
4.1. Deciphering the genetic code

While the double helix structure of DNA was first reported in 1953, it took another fifteen to twenty years for the development of experimental methods for determining DNA sequences. In 1977, two approaches for reading DNA sequences were presented, both of which were based on the generation of ssDNA fragments of varying lengths spanning the template molecule and ending with a certain base, for subsequent size separation and base calling by gel electrophoresis. The method of Maxam & Gilbert relied on base-specific chemical cleavage of DNA\(^{213}\), whereas Sanger used so-called dideoxy nucleotides (ddNTPs) - modified nucleotides that lack a 3’ hydroxyl group and therefore cannot be extended by a polymerase\(^{212}\). In both methods, four separate reactions are performed in order to obtain radioisotope-labeled fragments ending with a specific base, the identity of which varies from reaction to reaction. The Sanger method was slightly easier and less hazardous than that presented by Maxam & Gilbert and therefore became the method of choice. The first step towards the automation of Sanger sequencing was taken by Hood and colleagues in 1986; these workers used color-coded primers for each of the four synthesis reactions, making it possible to analyze the sample during the gel electrophoresis step\(^{214,215}\). Color-coded terminating nucleotides have since replaced color-coded primers, to allow all four synthesis reactions to be carried out in a single tube. Gel electrophoresis in capillaries has further increased the method’s sensitivity, increasing the read length to its current 700-1000 bp\(^{215}\).

Using contemporary techniques, it is possible to run 96 or 384 samples simultaneously.

The Sanger method has dominated the DNA sequencing field for nearly thirty years and has enabled a number of important achievements. The first sequenced genome of a free-living organism (a bacterium) was published in 1995\(^{216}\) and several others followed shortly thereafter. The first human genome sequence was a key achievement of biotechnology and resulted from a race between a public and a private initiative. The two competing teams used different approaches to stitch Sanger reads together into contigs and then assemble these into a consistent whole. The public initiative used a direct mapping approach in which the genome was broken into ordered and overlapping fragments that were cloned and then sequenced, whereas the private initiative used “whole genome shotgun sequencing” of random DNA fragments followed by sequence assembly on the basis of overlapping sequences\(^{216,217}\). The private initiative also introduced paired-end sequencing of both ends of the cloned fragment to facilitate the assembly of sequence reads\(^{217}\). Despite the fact that the private initiative started years later than the publicly funded group, the two groups published their drafts of the human genome on the same week in 2001 and the race was officially a tie\(^{214,215}\). Both shotgun sequencing and paired-end sequencing have since found widespread use. It took 13 years and the efforts of a very large number of people to accomplish the sequencing of the human genome, at a total cost of approximately $3 billion\(^{218}\).

New methods for DNA sequencing were developed during the late 80’s and early 90’s. One such method involves sequencing by the hybridization of labeled template DNA molecules to nucleotide probes that have been immobilized on a membrane or surface. By analyzing the hybridization pattern, it is possible to assemble an entire sequence, although this method is incapable of detecting single-base alterations\(^{219}\). A sequencing method involving cyclical monitoring of DNA synthesis events by detecting luminescence, known as pyrosequencing, was also developed; this technique is discussed in more detail later on in 4.1. Deciphering the genetic code

While the double helix structure of DNA was first reported in 1953, it took another fifteen to twenty years for the development of experimental methods for determining DNA sequences. In 1977, two approaches for reading DNA sequences were presented, both of which were based on the generation of ssDNA fragments of varying lengths spanning the template molecule and ending with a certain base, for subsequent size separation and base calling by gel electrophoresis. The method of Maxam & Gilbert relied on base-specific chemical cleavage of DNA\(^{213}\), whereas Sanger used so-called dideoxy nucleotides (ddNTPs) - modified nucleotides that lack a 3’ hydroxyl group and therefore cannot be extended by a polymerase\(^{212}\). In both methods, four separate reactions are performed in order to obtain radioisotope-labeled fragments ending with a specific base, the identity of which varies from reaction to reaction. The Sanger method was slightly easier and less hazardous than that presented by Maxam & Gilbert and therefore became the method of choice. The first step towards the automation of Sanger sequencing was taken by Hood and colleagues in 1986; these workers used color-coded primers for each of the four synthesis reactions, making it possible to analyze the sample during the gel electrophoresis step\(^{214,215}\). Color-coded terminating nucleotides have since replaced color-coded primers, to allow all four synthesis reactions to be carried out in a single tube. Gel electrophoresis in capillaries has further increased the method’s sensitivity, increasing the read length to its current 700-1000 bp\(^{215}\).

Using contemporary techniques, it is possible to run 96 or 384 samples simultaneously.

The Sanger method has dominated the DNA sequencing field for nearly thirty years and has enabled a number of important achievements. The first sequenced genome of a free-living organism (a bacterium) was published in 1995\(^{216}\) and several others followed shortly thereafter. The first human genome sequence was a key achievement of biotechnology and resulted from a race between a public and a private initiative. The two competing teams used different approaches to stitch Sanger reads together into contigs and then assemble these into a consistent whole. The public initiative used a direct mapping approach in which the genome was broken into ordered and overlapping fragments that were cloned and then sequenced, whereas the private initiative used “whole genome shotgun sequencing” of random DNA fragments followed by sequence assembly on the basis of overlapping sequences\(^{216,217}\). The private initiative also introduced paired-end sequencing of both ends of the cloned fragment to facilitate the assembly of sequence reads\(^{217}\). Despite the fact that the private initiative started years later than the publicly funded group, the two groups published their drafts of the human genome on the same week in 2001 and the race was officially a tie\(^{214,215}\). Both shotgun sequencing and paired-end sequencing have since found widespread use. It took 13 years and the efforts of a very large number of people to accomplish the sequencing of the human genome, at a total cost of approximately $3 billion\(^{218}\).

New methods for DNA sequencing were developed during the late 80’s and early 90’s. One such method involves sequencing by the hybridization of labeled template DNA molecules to nucleotide probes that have been immobilized on a membrane or surface. By analyzing the hybridization pattern, it is possible to assemble an entire sequence, although this method is incapable of detecting single-base alterations\(^{219}\). A sequencing method involving cyclical monitoring of DNA synthesis events by detecting luminescence, known as pyrosequencing, was also developed; this technique is discussed in more detail later on in
this chapter. Shortly thereafter, methods for sequencing by mass spectrometry that make it possible to rapidly analyze short sequences were reported. A common factor in all of these methods, including Sanger sequencing, is the use of a labor-intensive preparative protocol prior to sequencing. The preparation of samples for sequencing typically involves a sequence of template amplification by PCR, cloning into a vector, clone picking and PCR amplification of the target DNA. Methods for the in vivo amplification of templates in agarose gel or on solid surfaces, prior to sequencing first appeared in the early 2000s and paved the way for future efforts towards parallel sequencing. The advent of massively parallel sequencing in 2005 represented the start of a new era, making it possible to amplify numerous templates separately in a single reaction and then sequence them in parallel.

### 4.2. The Massively Parallel Sequencing toolbox

Massively parallel sequencing can be done on solitary or clonally amplified templates, using different chemistries, although there are some common themes that are present in several platforms. The unique combination of techniques used in a given MPS technology determines the type of data that platform can produce.

#### 4.2.1. Sequencing chemistry

**Sequencing by synthesis**

There are various different ways of sequencing a DNA template by synthesizing complementary DNA. By sequentially adding one nucleotide species at a time (and including washing steps between each successive addition step) and monitoring some effect induced by the incorporation of that nucleotide into the growing DNA strand, it is possible to observe the extension of the strand as it proceeds and determine which nucleotide was added. This concept is used in different ways in the Roche/454 and Ion Torrent sequencing systems.

Fluorescently-labeled terminating nucleotides with a removable blocking group that prevents chain extension while it is present are used in a procedure called Cyclic Reversible Termination (CRT). Chain extension with a modified nucleotide is followed by washing and fluorescent imaging, to analyze the incorporation event. The fluorophore is then removed and the newly-incorporated nucleotide is de-blocked, rendering it competent in chain extension and allowing the next cycle to proceed. The different nucleotides may be coded with differently-colored dyes and added as a mixture (Iluminca or labeled with the same dye and added in separate steps, as is done in the Helicos method. The former case is arguably better, since having all four bases present at once should reduce the likelihood of the wrong base being incorporated at any one time – the incorporation of the correct nucleotide is more kinetically favorable than that of a mismatched one. There are a variety of blocking groups that can be used in these techniques. The Illumina method accomplishes 3’ blocking by modifying the 3’ hydroxyl group, which is essential for incorporation during the chain extension reaction. The 454 method accomplishes 3’ blocking by modifying the 3’ hydroxyl group, which is essential for incorporation during the chain extension reaction.

#### 4.2.1. Sequencing chemistry

**Sequencing by synthesis**

There are various different ways of sequencing a DNA template by synthesizing complementary DNA. By sequentially adding one nucleotide species at a time (and including washing steps between each successive addition step) and monitoring some effect induced by the incorporation of that nucleotide into the growing DNA strand, it is possible to observe the extension of the strand as it proceeds and determine which nucleotide was added. This concept is used in different ways in the Roche/454 and Ion Torrent sequencing systems.

Fluorescently-labeled terminating nucleotides with a removable blocking group that prevents chain extension while it is present are used in a procedure called Cyclic Reversible Termination (CRT). Chain extension with a modified nucleotide is followed by washing and fluorescent imaging, to analyze the incorporation event. The fluorophore is then removed and the newly-incorporated nucleotide is de-blocked, rendering it competent in chain extension and allowing the next cycle to proceed. The different nucleotides may be coded with differently-colored dyes and added as a mixture (Iluminca or labeled with the same dye and added in separate steps, as is done in the Helicos method. The former case is arguably better, since having all four bases present at once should reduce the likelihood of the wrong base being incorporated at any one time – the incorporation of the correct nucleotide is more kinetically favorable than that of a mismatched one. There are a variety of blocking groups that can be used in these techniques. The Illumina method accomplishes 3’ blocking by modifying the 3’ hydroxyl group, which is essential for incorporation during the chain extension reaction. The 454 method accomplishes 3’ blocking by modifying the 3’ hydroxyl group, which is essential for incorporation during the chain extension reaction.
the formation of a covalent bond with the incoming nucleotide. In addition, a cleavable fluorescent label is appended to the base. Both modifications are chemically removed in each cycle. In Helicos sequencing, rather than modifying the nucleotide at two positions, a cleavable linker containing a fluorescent dye is incorporated into the base in a position that allows its steric bulk to prevent further extension without requiring chemical modification of the 3'-hydroxy group. In both of these methods, some fragments of the linkers that connected the label/blocking group to the nucleotide remain attached after cleavage, causing so-called "molecular scars" to accumulate in the DNA strand during sequencing.

A third approach used in some single-molecule sequencing methods involves real-time monitoring of synthesis events by observing the incorporation of fluorescently-labeled nucleotides into the growing DNA chain. Real-time analysis requires nucleotides that are capable of undergoing further extension without chemical modification, but are color-coded with fluorescent tags that are released during the incorporation event itself. The fluorophores are bound to the terminal phosphate of the nucleotide; the release of the fluorescent tag generates natural unmodified DNA.

Sequencing by ligation
Ligation of fluorescently labeled oligonucleotides to a complementary template molecule is another way of interrogating a template DNA molecule. Like some of the other methods discussed above, this approach also involves cyclical sequencing using a mixture of color-coded probes that are allowed to hybridize to the primed template. A ligase then forms a covalent bond between the probe and the primer. After washing, fluorescent imaging is used to determine which probe paired up with the template. Probes may encode one or two bases, which are combined with random or universal bases for stabilization. Before the next cycle begins, a free phosphate group needs to be re-generated to allow ligation. This is done either by cleaving off the last nucleotides of the incorporated probe, allowing ligation of the next probe in a chained manner (SOLiD), or by resetting the template by stripping off the whole primer-probe complex after each ligation step. In this case, several probes are used that are complementary to different positions on the target strand. Complete Genomics, Polonator, and others have used this approach, sequencing primers with different intercepts are used in order to start the reaction at different bases.

4.2.2. Template amplification
Most imaging systems are not built to detect single fluorescent events and so clonal amplification of templates prior to sequencing is commonly used. Immobilization on spatially separated positions avoids PCR competition and allows hundreds of thousands to billions of templates to be amplified in parallel.

Emulsion PCR
By compartmentalizing separate PCR reactions in water-in-oil emulsions, template molecules can be clonally amplified on the surface of primer beads. Adaptors containing universal priming sites are introduced onto the ends of the templates by ligation or PCR, and when mixed with beads that have primers on the surface, templates are captured by hybridization; the aim is to produce no more than one copy per bead. Beads are introduced into droplets containing PCR reagents and a small amount of the same

the formation of a covalent bond with the incoming nucleotide. In addition, a cleavable fluorescent label is appended to the base. Both modifications are chemically removed in each cycle. In Helicos sequencing, rather than modifying the nucleotide at two positions, a cleavable linker containing a fluorescent dye is incorporated into the base in a position that allows its steric bulk to prevent further extension without requiring chemical modification of the 3'-hydroxy group. In both of these methods, some fragments of the linkers that connected the label/blocking group to the nucleotide remain attached after cleavage, causing so-called "molecular scars" to accumulate in the DNA strand during sequencing.

A third approach used in some single-molecule sequencing methods involves real-time monitoring of synthesis events by observing the incorporation of fluorescently-labeled nucleotides into the growing DNA chain. Real-time analysis requires nucleotides that are capable of undergoing further extension without chemical modification, but are color-coded with fluorescent tags that are released during the incorporation event itself. The fluorophores are bound to the terminal phosphate of the nucleotide; the release of the fluorescent tag generates natural unmodified DNA.

Sequencing by ligation
Ligation of fluorescently labeled oligonucleotides to a complementary template molecule is another way of interrogating a template DNA molecule. Like some of the other methods discussed above, this approach also involves cyclical sequencing using a mixture of color-coded probes that are allowed to hybridize to the primed template. A ligase then forms a covalent bond between the probe and the primer. After washing, fluorescent imaging is used to determine which probe paired up with the template. Probes may encode one or two bases, which are combined with random or universal bases for stabilization. Before the next cycle begins, a free phosphate group needs to be re-generated to allow ligation. This is done either by cleaving off the last nucleotides of the incorporated probe, allowing ligation of the next probe in a chained manner (SOLiD), or by resetting the template by stripping off the whole primer-probe complex after each ligation step. In this case, several probes are used that are complementary to different positions on the target strand. Complete Genomics, Polonator, and others have used this approach, sequencing primers with different intercepts are used in order to start the reaction at different bases.

4.2.2. Template amplification
Most imaging systems are not built to detect single fluorescent events and so clonal amplification of templates prior to sequencing is commonly used. Immobilization on spatially separated positions avoids PCR competition and allows hundreds of thousands to billions of templates to be amplified in parallel.

Emulsion PCR
By compartmentalizing separate PCR reactions in water-in-oil emulsions, template molecules can be clonally amplified on the surface of primer beads. Adaptors containing universal priming sites are introduced onto the ends of the templates by ligation or PCR, and when mixed with beads that have primers on the surface, templates are captured by hybridization; the aim is to produce no more than one copy per bead. Beads are introduced into droplets containing PCR reagents and a small amount of the same
primer that is found on the bead. Beads that carry the amplified template on their surface are then enriched and either embedded in agarose gel (Polonator)\cite{22}, chemically cross-linked to an amino coated glass surface (SOLID; Polonator)\cite{23,24} or placed in wells of a PicoTiterPlate (Roche/454)\cite{25} for subsequent sequencing.

**Bridge amplification**

Solid phase amplification on a glass slide makes it possible to prepare randomly distributed clusters of clonally amplified molecules on an easily-interrogated flow cell. Adapter-ligated template library molecules are captured by hybridization to the lawn of high-density forward and reverse primers that are covalently attached to the chip surface. An initial extension step starting from the immobilized primer on the surface generates one surface-attached template molecule; residual unbound DNA is removed by denaturation and washing. The attached molecule is then allowed to bend and hybridize to the surface primer corresponding to its outer end sequence, forming a bridge and a priming site for extension. Bridge amplification is carried out in a cyclic manner under isothermal conditions, using formamide for denaturation\cite{26}; the result is a dense lawn of micron-sized clusters on the chip surface. To make the clusters suitable for sequencing, one of the primer sequences on the surface contains a cleavage site; cleavage generates allows clusters containing identical 5’-attached DNA strands that are ready for sequencing.

**Rolling Circle amplification**

Rolling circle amplification (RCA) using the polymerase 429 is described in the preceding chapters of this thesis. Briefly, the high processivity and strand displacement activity of 429 enables it to amplify a circularized template into a very long DNA molecule containing a large number of copies of the original molecules by loop ing round the template while synthesizing\cite{27}. The amplification product then collapses into a sphere, which is referred to as a DNA Nanoball (DNB) in the terminology used by Complete Genomics\cite{28}. The sequencing of clonally amplified target molecules must occur in a synchronized fashion, with all target copies undergoing the same process at the same time, since the addition of multiple nucleotides/probes or incomplete extension of a cluster would result in a de-synchronization of that cluster (leading/lagging strand de-phasing) that would increase slightly in every subsequent cycle. In the long run, de-phasing causes increased background signal levels and results in base call errors, a decrease in sequence quality towards the end of the read, and a limited read length\cite{29,30}.

**4.2.3. Shrinking observation window for single molecule visualization**

The sequencing of solitary template molecules by fluorescent labeling imposes high demands on the imaging system, as it must be able to accurately detect single fluorophores against the background signal originating from the other nucleotides present in the reaction. A few tricks have been used to focus the window of visualization in order to make the signal-to-noise ratio high enough for stable base calling. Polymerases perform optimally at micromolar nucleotide concentrations\cite{31}. At these levels, the background signal from primer that is found on the bead. Beads that carry the amplified template on their surface are then enriched and either embedded in agarose gel (Polonator)\cite{22}, chemically cross-linked to an amino coated glass surface (SOLID; Polonator)\cite{23,24} or placed in wells of a PicoTiterPlate (Roche/454)\cite{25} for subsequent sequencing.

**Bridge amplification**

Solid phase amplification on a glass slide makes it possible to prepare randomly distributed clusters of clonally amplified molecules on an easily-interrogated flow cell. Adapter-ligated template library molecules are captured by hybridization to the lawn of high-density forward and reverse primers that are covalently attached to the chip surface. An initial extension step starting from the immobilized primer on the surface generates one surface-attached template molecule; residual unbound DNA is removed by denaturation and washing. The attached molecule is then allowed to bend and hybridize to the surface primer corresponding to its outer end sequence, forming a bridge and a priming site for extension. Bridge amplification is carried out in a cyclic manner under isothermal conditions, using formamide for denaturation\cite{26}; the result is a dense lawn of micron-sized clusters on the chip surface. To make the clusters suitable for sequencing, one of the primer sequences on the surface contains a cleavage site; cleavage generates allows clusters containing identical 5’-attached DNA strands that are ready for sequencing.

**Rolling Circle amplification**

Rolling circle amplification (RCA) using the polymerase 429 is described in the preceding chapters of this thesis. Briefly, the high processivity and strand displacement activity of 429 enables it to amplify a circularized template into a very long DNA molecule containing a large number of copies of the original molecules by loop ing round the template while synthesizing\cite{27}. The amplification product then collapses into a sphere, which is referred to as a DNA Nanoball (DNB) in the terminology used by Complete Genomics\cite{28}. The sequencing of clonally amplified target molecules must occur in a synchronized fashion, with all target copies undergoing the same process at the same time, since the addition of multiple nucleotides/probes or incomplete extension of a cluster would result in a de-synchronization of that cluster (leading/lagging strand de-phasing) that would increase slightly in every subsequent cycle. In the long run, de-phasing causes increased background signal levels and results in base call errors, a decrease in sequence quality towards the end of the read, and a limited read length\cite{29,30}.

**4.2.3. Shrinking observation window for single molecule visualization**

The sequencing of solitary template molecules by fluorescent labeling imposes high demands on the imaging system, as it must be able to accurately detect single fluorophores against the background signal originating from the other nucleotides present in the reaction. A few tricks have been used to focus the window of visualization in order to make the signal-to-noise ratio high enough for stable base calling. Polymerases perform optimally at micromolar nucleotide concentrations\cite{31}. At these levels, the background signal from
surrounding fluorescently labeled nucleotides is very high. This issue is eliminated in cyclical wash-and-scan approaches by simply removing unincorporated nucleotides/probes before image capture. For real-time sequencing with an optical readout, the issue can be addressed by shrinking the volume of observation beyond what can be obtained by conventional methods such as confocal or total internal reflection microscopy.\textsuperscript{234,242}

The metallic grid in the door of a microwave oven prevents hazardous microwaves from getting out of the oven while still allowing us to peek through the holes to monitor the cooking process. The reason for this is that microwaves, as inferred by the name, have wavelengths in the µm-range, which makes them too large to pass through the holes of the grid. Visual light on the other hand has short enough wavelengths to get through the wholes, and we can therefore see through the grid in the door. This concept is used in Zero-mode waveguides (ZMW), which are the tiny wells used in Pacific Biosciences sequencing technology. ZMWs are 100 nm deep, circular holes (30 nm Ø) in a metal film, placed on a glass surface. When illuminated from the bottom by laser beamlets whose wavelength is greater than the wells’ diameter, this smart construction allows the beamlets to reach 20-30 nm into the well. This distance is just enough to reach the polymerase that sits at the bottom of the well and excite nucleotides that are close to the polymerase, but the high concentrations of labeled nucleotides that are present in the rest of the well are not illuminated by the laser and therefore do not affect the analysis.\textsuperscript{233,235}

Another variant on the same theme involves using a fluorescent nanocrystal (a so-called quantum dot, or Qdot) that is affixed to the polymerase as the source of light. The single stranded DNA to be sequenced is immobilized on a glass slide and the modified polymerase binds to the DNA. UV laser illumination excites the polymerase-linked Qdot, causing it to emit fluorescent light at a lower wavelength. The wavelength emitted by the Qdot matches that required to excite the fluorescently labeled nucleotides. When dye-labeled nucleotides are close enough to the polymerase, the light emitted by the Qdot can in turn excite a nucleotide-specific light signal. This makes it possible to determine DNA sequences by real-time monitoring using a CCD camera, and yields exceptionally high read lengths.\textsuperscript{234,235}

The method was initially called Visigen, but when LifeTech purchased the technology in 2008, it was rebranded as Starlight. More details regarding this approach were presented to a full auditorium at the 2010 AGBT Conference in Florida, but little has been heard about it since.
4.3. Massive sequencing by washing and scanning

4.3.1. 454 – pyrosequencing en masse

The first sign of what was to revolutionize the field of sequencing appeared with the landmark publication describing 454 sequencing in 2005[25]. For the first time, it became possible to sequence hundreds of thousands of amplicons at once. The 454 technology, which was named for the project number under which it was developed, is based on the pyrosequencing approach, which involves sequencing by synthesis and was developed at the Royal Institute of Technology (KTH) in the late 1990s[22,23]. Briefly, each time a nucleotide is incorporated into the growing DNA chain, one inorganic pyrophosphate (PPi) molecule is released. Through an enzymatic cascade, involving ATP sulfurylase and firefly luciferase, this results in the emission of a detectable light signal. By sequentially adding one type of nucleotide at a time, with continuous imaging by CCD cameras, it is possible to determine the sequence of the target DNA[23]. An additional enzyme, apyrase, is used to continuously degrade unused nucleotides in order to keep the background signal low[22]. Pyrosequencing was originally used with PCR amplicons in 96-well sequencing plates and its read length was limited to 40 bp due to the enzyme-inhibiting effects of the by-products that are formed in each cycle and accumulate as the reactions proceed. The 454 platform uses the same sequencing chemistry but the reactions are performed in a fluidic system that makes it possible to perform washing steps in between each cycle to remove byproducts. This has increased the read lengths from the 40 bp of the original pyrosequencing setup to 700 bp at present.

Clonal amplification of fragmented and linker-ligated template molecules on the surface of 20-25 µm sepharose beads is performed within the droplets of an emulsion. DNA beads are enriched and loaded into individual titanium-coated picolitre volume wells that are etched into the surface of a fiber-optic slide. Beads are then embedded in smaller, enzyme-covered beads and pyrosequencing is carried out by sequentially flooding the chip with different reagents while imaging individual DNA-beads using a CCD camera. Because the method uses natural non-terminated nucleotides, the only way to determine the length of homopolymer stretches in the target is to assess the intensity of the light signal. However, the relationship between repeat length and signal intensity is only linear for homopolymeric stretches of up to 5 nucleotides[21,22]. Consequently, insertions and deletions are the most common errors in 454 sequence reads and the error rates for homopolymeric regions and their surroundings are high due to the increased risk of carry-forward (premature incorporation of the nucleotide present in the homopolymer) and incomplete extension[24,23].

Today, the throughput of 454 sequencing is relatively low and its cost per base is higher than other MPS platforms. The power of 454 sequencing is its read length, which puts 454 in a niche of its own for de novo sequencing applications. Currently, 1M sequences with average read length of over 700 bp can be obtained in a 23h-sequencing run[25].
4.3.2. SOLiD – chained ligation in color space

In Sequencing by Ligation and Detection (SOLiD) base interrogation is carried out by cyclic ligation of two-base color-coded probes, using templates that were previously amplified on the surfaces of beads\(^\text{232}\). Library production and amplification are performed using very similar methods to those used in 454 sequencing, but following emPCR (cf. the section on “Template amplification”) and DNA bead enrichment, amplified DNA is chemically modified at the 3’ end to enable random immobilization of the beads on a glass slide for sequencing. By printing at the 3’ end close to the bead surface, the hybridization of color-coded probes to the primer and the ligation of matching probes can be followed by washing and imaging to infer the identity of the known bases in the probe. The identities of the bases in the first two positions of each probe are known and are associated with specific fluorescent dyes. Chemical cleavage with silver ions removes the last three bases and thereby also the fluorescent label. A series of ligation-steps is performed, giving rise to equal numbers of color calls with three unknown bases between each color-call. The extension product is then removed and the template is reset. A new sequencing primer that binds to a site one position along the template is then hybridized and a new round of ligation cycles is performed. There are a variety of different implementations of this method that use different numbers of chained ligations and sequencing primers, using the most recent setups and instruments, read lengths of up to 75 bp can be attained\(^\text{246}\). While SOLiD sequencing has lagged behind the alternatives (with the exception of Helicos) in terms of read lengths since it entered the market, this most recent upgrade greatly increases its general utility.

The two-base encoding confers inbuilt error correction and hence gives higher accuracy than one-base encoding\(^\text{212}\). SNPs give rise to color-changes in two positions rather than just one. However, the technique requires data manipulation in color-space, whereby color-space sequence reads are aligned to a color-space reference genome; this requirement is unique to SOLiD sequencing. Ligation-based methods also reduce problems with homopolymeric regions, which can give rise to insertions or deletions in sequence data with several other platforms. Instead, the most prominent sequencing error is substitution. The accuracy of the method can be increased from 99.94% to 99.99% by switching to a more expensive protocol that uses a second, three-base encoding probe set. This method yields base-calls directly rather than color-calls\(^\text{244}\).

4.3.3. Illumina – reversible terminators at their best

Thanks to a combination of respectable read lengths and high throughput, Illumina sequencing dominates the sequencing market today. The platform utilizes bridge amplification (cf. the section on “Template amplification”) to obtain billions of clusters made up of identical single-stranded template molecules that are densely packed on a glass slide and ready to be sequenced. Illumina is a “sequencing by synthesis” technology, using reversibly-terminated nucleotides (cf. the section on “Chemistry”). Briefly, the sequencing primer is hybridized to a template molecule, a mixture of reversibly terminated color-coded nucleotides is added, and one matching nucleotide is incorporated into each strand. This step is followed by washing, imaging, and de-blocking, and the whole process is repeated cyclically. 4-color reads are obtained using a 4-laser system\(^\text{195}\). Paired-end sequencing increases its general utility.

The two-base encoding confers inbuilt error correction and hence gives higher accuracy than one-base encoding\(^\text{212}\). SNPs give rise to color-changes in two positions rather than just one. However, the technique requires data manipulation in color-space, whereby color-space sequence reads are aligned to a color-space reference genome; this requirement is unique to SOLiD sequencing. Ligation-based methods also reduce problems with homopolymeric regions, which can give rise to insertions or deletions in sequence data with several other platforms. Instead, the most prominent sequencing error is substitution. The accuracy of the method can be increased from 99.94% to 99.99% by switching to a more expensive protocol that uses a second, three-base encoding probe set. This method yields base-calls directly rather than color-calls\(^\text{244}\).
can be done by switching the direction of the clusters on the slide, through an additional bridge amplification step, followed by a second round of sequencing cycles. Illumina sequencing currently offers 100 +100 bp paired-end reads or 150 bp long single end reads from three billion DNA clusters[24]. The most common sequencing errors are substitutions, which occur when DNA clusters become out-of-sync due to double or incomplete extension.

4.3.4. Complete Genomics – outsourced sequencing of DNA nano-balls

The technology used by Complete Genomics is a comparatively inaccessible and not amenable to modification by the customer, since the company only offers a sequencing service. Sample preparation and sequencing is achieved by subjecting fragmented DNA using a number of cutting - ligation and circularization processes to yield, circularized strands of target DNA containing 4-8 different adapter sequences in between short stretches of template sequence. The adapter sequences are then used as starting points for sequencing. The circular template is amplified by rolling circle amplification to obtain numerous copies of the original molecule that are joined in a head-to-toe fashion[19] in long single-stranded DNA molecules. These self-assemble into DNA Nano-balls (DNB) that are about 200 nm in size. The DNA Nano-balls are then positioned on patterned nanoarrays, aligned with the pixels of the imaging camera. Sequencing of each DNB is carried out by an unchained ligation technique called combinatorial probe anchor ligation (cPAL) chemistry[20]. An adapter-complementary anchor primer is hybridized to the DNB and degenerate nonamer (8-unit-probes containing one known, color-coded nucleotide in position one) are hybridized and ligated to the DNB. After washing and fluorescent readout, the extended construct is removed from the template and the reaction is repeated, this time using probes with a known base in position two. The step is repeated five times to identify the bases in positions 1-5 and then a new adapter sequence containing five degenerate bases is used to interogate nucleotides in position 6-10, in a similar manner. Up to ten bases may be read from each adapter-DNA junction, ultimately yielding a read length of 35 + 35 bp from each end of the original DNA fragment. The sequencing method was first demonstrated by sequencing three human genomes[31].

The business model adopted by Complete Genomics has been available since November 2010, and is very different to that used with other MPS platforms. Instead of selling instruments and reagents, Complete Genomics offers a sequencing service that specialized in the sequencing of entire human genomes. Customers send in a DNA sample and the whole process of library preparation, sequencing, assembly and variant calling is done by the company[21]. For $5,000-$8,000 per genome, customers receive sequence data with 40x coverage in less than three months[18]. The low price stems from the low cost of the reagents used and the extreme throughput of the system. Complete Genomics has already sequenced over 1,000 high-coverage genomes for its customers[35].

Due to the closed nature of the business model of Complete Genomics, information on raw sequence read quality has been difficult to obtain. An educated guess would be that the most common errors obtained from unchained ligation, which has no error propagation or phasing, would be substitutions due to hybridization and ligation of the wrong probes, giving rise to miscalls.
4.3.5. Ion Torrent – moving beyond light

The first "post-light sequencing" platform was recently presented and has since then received significant attention. The platform is basically a miniaturized pH meter, which monitors sequencing by synthesis on a large number of DNA templates, in parallel. When a phosphodiester bond is formed in the nucleotide incorporated into a growing strand, a proton is released from the 3'-OH group of the primer strand and rapidly diffuses away from the DNA molecule, which has gained one negative charge unit. Thus, by monitoring the evolution of charge, it is possible to detect nucleotide incorporation events without fluorescent labeling or enzymatic cascades that generate luminescence.

The viability of detecting nucleotide incorporation by monitoring charge perturbation was first demonstrated in 2006 by Pourmand and coworkers, using DNA immobilized on the surface of gold-coated electrodes and a voltage clamp amplifier. Since then the technique was further developed to allow sequencing en masse; a commercial platform named the "Ion Torrent" system entered the market at the end of 2010[250,251]. Sample preparation is quite similar to that used in Roche/454 sequencing, involving emulsion PCR amplification of adapter-ligated templates on 2 µm acrylamide beads, loading of DNA-beads into wells, and the cyclical addition of individual types of natural nucleotides one at a time, with washing steps in between each addition. However, the reactions are carried out in wells in a semiconductor chip, of the kind used in computers and mobile phones; these are very cheap to produce. Since the system relies on the direct detection of voltage changes, no imaging is involved, which avoids the need for expensive optical capture systems and makes Ion Torrent sequencing simpler, much faster, and more scalable than other wash-and-scan approaches. Its throughput is currently still lower than that of established short read platforms (but higher than that of 454), but the manufacturers expect that increasing the size of the semiconductor chip will allow for increased throughput in the near future. The average read length is currently 100 bp, but 200 bp reads have been acquired[252] and as of 2012 an increase to 400 bp-reads is foreseen. According to Ion Torrent, the per-base read accuracy is 99.2%, but as with 454 sequencing, homopolymers are problematic: 5-mers are called with 97.5% accuracy, and the probability of error rises as the repeats get longer[253].

In contrast to 454 sequencing, Ion Torrent does not require any enzymatic degradation of the nucleotides (possibly due to the fact that the relevant patent is owned by Qiagen, who acquired it when they bought the rights to pyrosequencing). The Ion Torrent proof of concept publication claims that the well size is small enough to allow rapid diffusion of nucleotides out of the wells[254]. However, since 454 sequencing and Ion Torrent have many technical similarities, the discrepancy between their maximum read lengths suggests that the accumulation of leftover material does indeed limit the Ion Torrent read length. It is conceivable that this could be ameliorated by using electrical charges to facilitate nucleotide removal while washing.

4.3.6. Helicos – single molecule sequencing in its infancy

Amplification prior to sequencing is problematic because it can introduce errors and bias in the sequenced clone[255,256]. Notably, because amplification is affected by fragment length and GC content[257,258], amplification before sequencing may skew the coverage of the target genome.
The Helicos Genetic Analysis technique was the first single-molecule sequencing platform to become available, this was in 2008258. Single molecule detection makes Helicos exclusive, although it relies on chemistry and procedures that are rather similar to those used in previously-described massively sequencing techniques. Fragmented DNA templates are 3’ polyadenylated, Cy3-labeled and blocked by the incorporation of Cy5 ddTTP, and then captured onto a planar surface by covalently bound “5’ down” poly(dT) probe259. Sequencing of surface-attached templates is achieved by incorporating reversibly-inhibited nucleotides, in a step-wise one-color setup. The patented nucleotide analogs carry an inhibitor and a fluorescent label, both of which are connected to the base by the same cleavable linker; this makes it possible to remove both the inhibitor and the label in a single step that affects only one position on the nucleotide and leaves it ready for the next cycle259. Imaging of single Cy5 fluorophores that have been introduced into the template strands in each cycle is achieved using a Charged Coupled Device (CCD).

The incorporation of unlabelled ‘dark bases’ results in deletion sequencing errors; the accuracy of the sequence data can be increased by increasing the amount of time spent on the sequencing process. This requires a slightly different library preparation protocol that involves adapter ligation prior to 3’ polyadenylation in order to introduce a priming site at the distal end of the fragment. This site is captured in the same way as above; however, by extending the immobilized probes, it is possible to obtain copies of every template that is covalently attached to the glass surface. These copies are primed using an adapter-complementary primer and can be sequenced repeatedly by melting off the extended primer and restarting the sequencing reaction259. The same library preparation procedure also makes it possible to perform paired-end reads by sequencing the hybridized template rather than performing the first copying step. Moreover, if the polymerase is exchanged for a reverse transcriptase, the same setup can be used for the direct sequencing of RNA molecules without the need for erroneous and bias-prone cDNA synthesis and further amplification256,257.

The amplification-free template preparation protocol is an important feature of the Helicos technology and has proven particularly useful for the sequencing of short and damaged/degraded DNA fragments from samples such as paraffin-embedded tissue or ancient remains. It is useful because compromised DNA quality can give rise to elevated bias in library preparations that involve PCR amplification258.

The single-molecule nature of the platform provides an advantage in terms of library preparation time and costs and absence of an error-prone PCR step. However, it is a wash-and-scan based platform, which increases the time and reagent cost of the sequencing reaction. In fact, the single-molecule nature of the Helicos platform works to its disadvantage as seen in its high error rate. Deletion is the most common error type, even with two-pass sequencing, and the substantial error rates (3-7%) for one pass and 2-5% for two passes256,257 must be taken into account, although the impact of this is mitigated to some extent by the high throughput of the instrument (600 M-1 B reads per run258,259). The problems relating to sequencing errors also limit the method’s read length to 32-33 bp258,259. Reads this short present challenges in terms of stitching together a genome, especially in repetitive regions259. Because it has the shortest read length of all contemporary platforms and the highest error rates, along with certain problems with instrument robustness, Helicos is the least popular of the currently-available platforms258.

The Helicos Genetic Analysis technique was the first single-molecule sequencing platform to become available, this was in 2008258. Single molecule detection makes Helicos exclusive, although it relies on chemistry and procedures that are rather similar to those used in previously-described massively sequencing techniques. Fragmented DNA templates are 3’ polyadenylated, Cy3-labeled and blocked by the incorporation of Cy5 ddTTP, and then captured onto a planar surface by covalently bound “5’ down” poly(dT) probe259. Sequencing of surface-attached templates is achieved by incorporating reversibly-inhibited nucleotides, in a step-wise one-color setup. The patented nucleotide analogs carry an inhibitor and a fluorescent label, both of which are connected to the base by the same cleavable linker; this makes it possible to remove both the inhibitor and the label in a single step that affects only one position on the nucleotide and leaves it ready for the next cycle259. Imaging of single Cy5 fluorophores that have been introduced into the template strands in each cycle is achieved using a Charged Coupled Device (CCD).

The incorporation of unlabelled ‘dark bases’ results in deletion sequencing errors; the accuracy of the sequence data can be increased by increasing the amount of time spent on the sequencing process. This requires a slightly different library preparation protocol that involves adapter ligation prior to 3’ polyadenylation in order to introduce a priming site at the distal end of the fragment. This site is captured in the same way as above; however, by extending the immobilized probes, it is possible to obtain copies of every template that is covalently attached to the glass surface. These copies are primed using an adapter-complementary primer and can be sequenced repeatedly by melting off the extended primer and restarting the sequencing reaction259. The same library preparation procedure also makes it possible to perform paired-end reads by sequencing the hybridized template rather than performing the first copying step. Moreover, if the polymerase is exchanged for a reverse transcriptase, the same setup can be used for the direct sequencing of RNA molecules without the need for erroneous and bias-prone cDNA synthesis and further amplification256,257.

The amplification-free template preparation protocol is an important feature of the Helicos technology and has proven particularly useful for the sequencing of short and damaged/degraded DNA fragments from samples such as paraffin-embedded tissue or ancient remains. It is useful because compromised DNA quality can give rise to elevated bias in library preparations that involve PCR amplification258.

The single-molecule nature of the platform provides an advantage in terms of library preparation time and costs and absence of an error-prone PCR step. However, it is a wash-and-scan based platform, which increases the time and reagent cost of the sequencing reaction. In fact, the single-molecule nature of the Helicos platform works to its disadvantage as seen in its high error rate. Deletion is the most common error type, even with two-pass sequencing, and the substantial error rates (3-7%) for one pass and 2-5% for two passes256,257 must be taken into account, although the impact of this is mitigated to some extent by the high throughput of the instrument (600 M-1 B reads per run258,259). The problems relating to sequencing errors also limit the method’s read length to 32-33 bp258,259. Reads this short present challenges in terms of stitching together a genome, especially in repetitive regions259. Because it has the shortest read length of all contemporary platforms and the highest error rates, along with certain problems with instrument robustness, Helicos is the least popular of the currently-available platforms258.
### Helicos

- Single labeled nucleotides Wash & one-color imaging
- Cleave dye & inhibitor
- Next type of dNTP

### SOLiD

- Color-encoded oligos ligated
- Repeat chained ligation seven times
- Ligation of color-encoded oligos
- Repeat for all anchors

### Complete Genomics

- Ligation of color-encoded oligos
- Wash & four-color imaging
- Reset reaction probeset 4
- Sequencing primer with another offset
- Repeat for all anchors
- Repeat with new probeset (total 5)

### Illumina

- Introduce terminated nucleotides
- Wash & four-color imaging
- Cleave dye & terminator
- Next base is read

### Ion Torrent

- H released at dNTP incorporation
- pH change +
- Sequencing primer with another offset
- Repeat chained ligation seven times
- Ligation of color-encoded oligos
- Repeat for all anchors
- Repeat with new probeset (total 5)

### 454

- APS
- Repeat with 20N primer (read 1)
- Repeat for all anchors
- Sequencing primer with another offset

### Ion Torrent

- H released at dNTP incorporation
- pH change +
- Sequencing primer with another offset
- Repeat chained ligation seven times
- Ligation of color-encoded oligos
- Repeat for all anchors
- Repeat with new probeset (total 5)

### Helicos

- Single labeled nucleotides Wash & four-color imaging
- Cleave dye & inhibitor
- Next type of dNTP

### Complete Genomics

- Ligation of color-encoded oligos
- Wash & four-color imaging
- Reset reaction probeset 4
- Sequencing primer with another offset
- Repeat for all anchors
- Repeat with new probeset (total 5)

### Illumina

- Introduce terminated nucleotides
- Wash & four-color imaging
- Cleave dye & terminator
- Next base is read

### Ion Torrent

- H released at dNTP incorporation
- pH change +
- Sequencing primer with another offset
- Repeat chained ligation seven times
- Ligation of color-encoded oligos
- Repeat for all anchors
- Repeat with new probeset (total 5)
4.4. Massive sequencing in real time

4.4.1. Pacific Biosciences – immobilized polymerase in a tiny well

Real-time sequencing of single DNA molecules first became available when PacBio RS started shipping in April 2011. As described above, there are various tricks that can be used to monitor individual DNA polymerases as they synthesize a complementary DNA strand.

Pacific Biosciences uses ZeroModeWaveguides (ZMW) to obtain a nanoscopic sequencing chamber containing one DNA polymerase each, to enable real-time monitoring of DNA polymerization for up to 15,000 bp\textsuperscript{234,242,263,264}.

The polymerase used is a modified variant of Q29, which has been engineered to work more slowly than the natural enzyme. In conjunction with the fact that the reaction is performed at low temperatures, this reduces the natural speed of nucleotide incorporation to about one nucleotide per second, which is sufficiently slow that each incorporation event can be captured on film. Fluorescently-labeled nucleotides diffuse freely into and out of the ZMW; the only nucleotides in the system that are not in motion are those that are being held in place by the immobilized polymerase immediately prior to their incorporation. Because the fluorescent dye is attached to the terminal phosphate group of the nucleotide rather than to the base, it is automatically cleaved off by the polymerase at incorporation and leaves no trace in the growing DNA strand\textsuperscript{263}.

Template preparation is PCR-free and involves DNA fragmentation and end repair, followed by ligation of looped adapters to effect template circularization. The circularization of the template is critical, since it "locks" the polymerase on the template and enables repeated sequencing of the template. Sequencing errors are common in this technology and occur for two reasons. First, if a (correct) nucleotide binds to the polymerase active site but is not incorporated, it may be interpreted as an incorporation event, creating an insertion error in the sequence data. Second, if the incorporation event occurs too quickly, it may be overlooked, resulting in a deletion error. The accuracy for each nucleotide position in a single-end read is only 85\%, which makes re-reading of the circular template very important in obtaining accurate sequences\textsuperscript{263,264}. However, light damage limits the life span of the polymerase; because of this, it is currently only possible to sequence short templates for circular consensus. In other words, one must choose between accuracy and read length. Templates of 1-6 kb currently yield single reads with moderate accuracy\textsuperscript{263}, while templates of 250 bp – 1 kb can be sequenced repeatedly to yield a circular consensus sequence with a lower error rate. Extended read lengths can be achieved using "strobed reads". By turning the laser illumination on and off at certain intervals, the life span of the polymerase is prolonged and a scaffold of linked sequence reads is obtained from a single template. "Strobed reads" can currently be used to obtain sequence information from templates between 6-10 kb; this can be useful in de novo assembly or for resolving repetitive regions. Sequencing is done in a strip of 8 SMRTcells; each SMRTcell contains 150,000 ZMWs, of which one third will contain a single DNA polymerase\textsuperscript{234,324}.

The number of concurrent sequencing reactions is currently limited by the 73k laser beams that are needed to individually illuminate a single ZMW, although it is possible to analyze several SMRTcells sequentially.

4.4. Massive sequencing in real time

4.4.1. Pacific Biosciences – immobilized polymerase in a tiny well

Real-time sequencing of single DNA molecules first became available when PacBio RS started shipping in April 2011. As described above, there are various tricks that can be used to monitor individual DNA polymerases as they synthesize a complementary DNA strand. Pacific Biosciences uses ZeroModeWaveguides (ZMW) to obtain a nanoscopic sequencing chamber containing one DNA polymerase each, to enable real-time monitoring of DNA polymerization for up to 15,000 bp\textsuperscript{234,242,263,264}.

The polymerase used is a modified variant of Q29, which has been engineered to work more slowly than the natural enzyme. In conjunction with the fact that the reaction is performed at low temperatures, this reduces the natural speed of nucleotide incorporation to about one nucleotide per second, which is sufficiently slow that each incorporation event can be captured on film. Fluorescently-labeled nucleotides diffuse freely into and out of the ZMW; the only nucleotides in the system that are not in motion are those that are being held in place by the immobilized polymerase immediately prior to their incorporation. Because the fluorescent dye is attached to the terminal phosphate group of the nucleotide rather than to the base, it is automatically cleaved off by the polymerase at incorporation and leaves no trace in the growing DNA strand\textsuperscript{263}.

Template preparation is PCR-free and involves DNA fragmentation and end repair, followed by ligation of looped adapters to effect template circularization. The circularization of the template is critical, since it "locks" the polymerase on the template and enables repeated sequencing of the template. Sequencing errors are common in this technology and occur for two reasons. First, if a (correct) nucleotide binds to the polymerase active site but is not incorporated, it may be interpreted as an incorporation event, creating an insertion error in the sequence data. Second, if the incorporation event occurs too quickly, it may be overlooked, resulting in a deletion error. The accuracy for each nucleotide position in a single-end read is only 85\%, which makes re-reading of the circular template very important in obtaining accurate sequences\textsuperscript{263,264}. However, light damage limits the life span of the polymerase; because of this, it is currently only possible to sequence short templates for circular consensus. In other words, one must choose between accuracy and read length. Templates of 1-6 kb currently yield single reads with moderate accuracy\textsuperscript{263}, while templates of 250 bp – 1 kb can be sequenced repeatedly to yield a circular consensus sequence with a lower error rate. Extended read lengths can be achieved using "strobed reads". By turning the laser illumination on and off at certain intervals, the life span of the polymerase is prolonged and a scaffold of linked sequence reads is obtained from a single template. "Strobed reads" can currently be used to obtain sequence information from templates between 6-10 kb; this can be useful in de novo assembly or for resolving repetitive regions. Sequencing is done in a strip of 8 SMRTcells; each SMRTcell contains 150,000 ZMWs, of which one third will contain a single DNA polymerase\textsuperscript{234,324}.

The number of concurrent sequencing reactions is currently limited by the 73k laser beams that are needed to individually illuminate a single ZMW, although it is possible to analyze several SMRTcells sequentially.
The long read lengths attained with the Pacific Biosciences method should be useful for finding fused genes and for sequencing bacterial genomes\textsuperscript{203,230}. Moreover, it has been shown that by analyzing the kinetic signature obtained by repeated sequencing of a sample, DNA modifications such as methylation patterns can be determined\textsuperscript{284}. A recently-reported study exploited the flexibility of the Pacific Biosciences sequencing instrument in an interesting way: by exchanging the DNA polymerase for a ribosome and adding labeled tRNAs instead of nucleotides, it was possible to observe mRNA translation in real time\textsuperscript{285}. The flexibility of the platform is appealing and a similar approach could likely be adopted for other enzymes such as reverse transcriptase or cellular receptors, thus it will be interesting to follow the future applications of this platform.

4.4.2. Nanopore sequencing – electrical measurement through a pore

Nanopore sequencing is still in an early stage of development, but offers the possibility of achieving high speeds and long reads (>5 kb)\textsuperscript{286} with unlabeled, unamplified DNA, meaning it could potentially have many significant advantages over contemporary sequencing platforms.

The central concept in nanopore sequencing is to monitor the translocation of a DNA molecule through a nanometer-sized pore in a membrane. By applying an electrical field to the membrane, the negatively charged DNA molecule can be threaded through the pore in sequential order. The identity of the individual bases passing through the pore can be read out in various ways. Nucleotide bases (i.e. dDNA, ssDNA and free nucleotides) passing through a nanopore block the current that naturally flows through the pore in a characteristic way (charge blocking), giving rise to an identifiable electrical signature\textsuperscript{287}. Alternatively, the nucleotides could be labeled with a fluorescent tag that would be read out optically as the nucleotide traverses the pore\textsuperscript{288}. Current approaches to nanopore sequencing use either natural or mutant protein nanopores, or pores manufactured in solid-state membranes.

The translocation velocity through the pore is extremely high; with the widely-used protein nanopore α-hemolysin it is >1 nucleotide/10 ps, and with solid-state membranes it is >1 nucleotide/10 ns\textsuperscript{289}. It can thus be difficult to obtain satisfactory resolution in nanopore sequencing; this issue has been addressed in various ways. It has been proposed that converting the DNA code into a longer (binary) code would be one way of getting around the issue of the single nucleotide resolution of nanopores; this approach could be used in combination with fluorescent detection by the hybridization of color-coded probes to each DNA segment\textsuperscript{290}.

Slowing down the translocation through the nanopore would also make signal interpretation easier. An engineered α-hemolysin nanopore has been modified to briefly bind each nucleotide as they pass through long enough to enable nucleotide identification based on charge blocking. It has been demonstrated that nucleotide identification by using an exonuclease to sequentially chew off bases from a DNA template and ‘spit’ them into the nearby nanopore is possible\textsuperscript{289,270}. The difficulty appears to lie in ensuring that every nucleotide reaches the nanopore. Another protein nanopore (MspA), which is narrow enough to let ssDNA through, but temporarily halts at a double stranded patch, has
recently been engineered and used to distinguish nucleotides separated by double-stranded sections in a promising proof-of-concept study\textsuperscript{271}. While nanopore methods have yet to be fully developed, the possibility of reading DNA sequences without relying on sequencing by synthesis or ligation and without enzymatic cascades or template amplification is very tantalizing, as it would potentially allow for very rapid and inexpensive sequencing; the considerable effort currently being expended on studying and developing this technique is wholly justified.

4.5. Choosing sequencing platform(s)

The sequencing platform of choice usually depends on the application at hand. The enormous sequence capacity offered by short-read platforms such as HiSeq, SOLiD (and to some extent Helicos) enables mutational detection by targeted (i.e. whole exome) resequencing as well as variant discovery by resequencing of whole genomes. The massive number of reads obtained from these platforms also makes them more useful for quantitative expression analysis by RNA sequencing. Helicos, due to its high error rate and short read lengths, is currently not a system of choice in many labs, but it is the only platform that enables direct RNA sequencing. The longer read lengths of the 454 system give it a special niche in de novo sequencing, since long sequence reads are valuable for stitching together new genomes. However, the throughput of this platform is comparatively low and its cost per base is high. A combination of several sequencing platforms may offer the best combination of long and massive reads for whole genome sequencing.
The cost per base of DNA sequencing is continuously decreasing due to industrial competition in combination with technical innovations, which are constantly increasing the data output from one instrument run. However the runtime of established short-read platforms is measured in weeks rather than hours and the large throughput of most MPS platforms does not match perfectly with sequencing projects of moderate size or clinical applications in which rapid results are needed (e.g. preimplantation genetics). To meet the demand for quick and simple sequencing, several scaled-down sequencing platforms have been announced (e.g. 454 GS Junior, Illumina MiSeq), which yield a fraction of the reads (GS Junior >100,000 reads, MiSeq >1.5M reads) of the large instruments in days rather than weeks. The Ion Torrent platform, with its low-cost 2h-runs and comparably low data output is also a suitable alternative for smaller labs or experiments.

Table 1. Overview over sequencing platforms 2011

<table>
<thead>
<tr>
<th>Platform</th>
<th>Chemistry</th>
<th>Platform Chemistry</th>
<th>Single nucleotide</th>
<th>Template amplification</th>
<th>Method</th>
<th>Average Readlength (bp)</th>
<th>Throughput (Gb/day)</th>
<th>Price per base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanger sequencing</td>
<td>Enzymatic chain termination</td>
<td>No</td>
<td>No</td>
<td>PCR</td>
<td>Cloning &amp; PCR</td>
<td>100</td>
<td>384</td>
<td>$0.0006</td>
</tr>
<tr>
<td>Pyrosequencing</td>
<td>Pyrosequencing</td>
<td>No</td>
<td>No</td>
<td>PCR</td>
<td>Cloning &amp; PCR</td>
<td>2-10</td>
<td>96</td>
<td>$0.0003</td>
</tr>
<tr>
<td>Helicos</td>
<td>One-color</td>
<td>No</td>
<td>Ye</td>
<td>100</td>
<td>emPCR</td>
<td>300</td>
<td>1000</td>
<td>$25</td>
</tr>
<tr>
<td>Complete Genomics</td>
<td>Unchained, one-base encoding</td>
<td>No</td>
<td>No</td>
<td>RCA</td>
<td>-</td>
<td>250</td>
<td>1500</td>
<td>$30</td>
</tr>
<tr>
<td>Life Tech/Ion Torrent</td>
<td>H+ release causes voltage change</td>
<td>No</td>
<td>No</td>
<td>emPCR</td>
<td>-</td>
<td>200</td>
<td>0.5</td>
<td>$1.2</td>
</tr>
<tr>
<td>Life Tech/Visigen</td>
<td>Color-coded nucleotides and FRET</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Life Tech/SOLiD</td>
<td>Chained, two-base encoding</td>
<td>No</td>
<td>No</td>
<td>emPCR</td>
<td>40</td>
<td>4000</td>
<td>$20</td>
<td></td>
</tr>
<tr>
<td>Illumina/Solexa</td>
<td>Four-color Cyclic Reversible Termination</td>
<td>No</td>
<td>No</td>
<td>Bridge amplification</td>
<td>100</td>
<td>3000</td>
<td>$25</td>
<td></td>
</tr>
<tr>
<td>Second generation Roche/454</td>
<td>Pyrosequencing</td>
<td>No</td>
<td>No</td>
<td>emPCR</td>
<td>700</td>
<td>1</td>
<td>$0.4-0.6</td>
<td></td>
</tr>
<tr>
<td>First Generation Sanger sequencing</td>
<td>Enzymatic chain termination</td>
<td>No</td>
<td>No</td>
<td>Cloning &amp; PCR</td>
<td>1000</td>
<td>96-384</td>
<td>$0.0006</td>
<td></td>
</tr>
</tbody>
</table>

Real-time detection is defined as unhalted detection of DNA synthesis or translocation through a nanopore. Conventional pyrosequencing is commonly seen as a real-time process, since the activity of the polymerase is monitored in real-time. However, since only one type of nucleotide is present at a time, it does not fall under the current definition of real-time detection platforms. Throughput per day is calculated, using times not including template amplification. Price per base for the different platforms is described by ranking, where $$$$$ indicates highest cost per base, however the bins do not correlate to a linear increase in price.
4.6. Applications of massive sequencing

The ability of massive sequencing to produce humongous amounts of data at a low price enables questions to be asked that were unimaginable not too long ago. The publication of the reference sequence of the human genome one decade ago\(^{16,17}\) was one of the largest scientific accomplishments in modern history. However, to maximize the impact of this landmark achievement on society, it will be necessary to identify genomic variants of medical importance, such as traits connected to disease susceptibility or drug responses.

During the last ten years, a second phase of human genomic science has taken over, with the aim of finding genomic variants linked to inherited diseases and other traits. This can be achieved in various ways, including by investigating genomes as a whole, particular parts or modifications thereof, or by analyzing gene expression. The examination of the genomes of related species can also shed light on questions such as the origin of species.

4.6.1. Whole genome sequencing

De novo sequencing

Initial, de novo, sequencing of unknown genomes, is common today; as of September 2011, the IMG database\(^{123}\) that keeps track of finished and draft genomes, contained 6,891 genomes from various animal, plant and microbial species\(^{123}\) and the figure is continuously increasing. Reference genomes from species with symbiotic value such as the kangaroo\(^{15,16}\), orangutan\(^{27,42}\) and plant panda\(^{37,38}\), as well as species that are important in the food industry such as cacao\(^{37,38}\), wild strawberry\(^{57,58}\), soybean\(^{57,58}\) and potato\(^{57,58}\) are constantly being generated. Obtaining sequence information on whole genomes of related organisms has allowed a number of interesting large-scale evolutionary studies and shed light on phylogenetic relationships in species ranging from marsupials\(^{17,31}\) to vascular plants\(^{201}\) and Hominidae, the great apes\(^{27,35}\).

Although de novo assemblies have been achieved using short sequencing reads alone\(^{281,281}\), bioinformatic assembly is greatly facilitated by long sequence reads\(^{281,281}\), which suggests that a combination of current sequencing methods may be useful for obtaining sequences with a length and depth suitable for de novo genome assembly. Mate-pairs may be helpful in stitching together unknown sequences and getting through problematic repetitive regions and for detecting structural variation between genomes\(^{281,281}\). Mate-pair libraries contain two joined segments that were originally situated several kilobasepairs from each other; the generation of such libraries provides a way to increase the physical coverage of the genome, since the distance between the two ends is known. Mate-pairs are generally created by end-biotinylatin and circulation of kilo-sized gDNA fragments, followed by fragmentation and streptavidin capture of junction fragments\(^{281,281}\).

In the serious EHEC epidemic outbreak in Germany in the summer of 2011, a number of people were infected with a highly contagious infection that killed several people within a few days. The infectious E. coli strain was rapidly isolated and sequenced by Pacific Biosciences\(^{275}\) and Ion Torrent\(^{287,288}\). Pacific Biosciences performed de novo assembly using both circular consensus and single-end long reads at 75x coverage resulting in 33 contigs\(^{281}\). Five Ion Torrent sequencing runs generated enough sequence reads for de novo assembly,
resulting in >3000 contigs\(^282,283\). Within a few weeks of the outbreak, sequence data from both platforms independently showed that the epidemic strain belonged to an \textit{E. coli} lineage that had acquired genes coding for antibiotic resistance and a toxin from another bacterial strain\(^280,289\), demonstrating the power of massive sequencing to rapidly deliver relevant clinical information.

**Resequencing**

As the cost of sequencing continues to fall, sequencing human genomes becomes more accessible and common. Resequencing of human genomes does not require long read lengths, since reads are mapped against a known reference sequence, and all current massive sequencing platforms apart from Pacific Biosciences (which presently has too low a throughput) have been used to resequence human genomes\(^282,283,289,290\). Each platform’s benchmark publications have reported the sequencing genomes of eminent people such as ‘one of the fathers of DNA’, James Watson\(^244\), or the stipulator of Moore’s law, Gordon Moore\(^295\), or anti-apartheid activist Archbishop Desmond Tutu\(^295\); and reference genomes belonging to people of various ethnic origin such as Nigeria\(^229\), Kalahari Desert and a Bantu southern Africa\(^294\), Korea\(^295\) and a Han-Chinese\(^295\) were also quickly established. Coverage usually varies between 7.4x\(^244\) and 36x\(^295\).

The ultimate goal of sequencing masses of human genomes is to, by obtaining plentiful genetic information, be able to distinguish patterns such as variants that are linked to positive/adverse affects of a particular drug treatment, or variants that are strongly or loosely associated with a particular disease. This would enable the design of drug regimes tailored to the unique genetic profiles of individual patients, the implementation of preventive treatment for individuals at elevated risk of developing a particular disease, and the creation of prognostic tools for determining the severity of disease progress. This idea is one of many that is being addressed by the 1000 Genomes Project Consortium\(^9\), a world-encompassing project that has the aim of sequencing a massive number of genomes of individuals from five major population groups, and to use this data to identify all genomic variants that are present at a frequency of 1% or more\(^296\). In a recent publication of the pilot-experiment of this huge task, involving low-coverage, high-coverage and targeted resequencing of a number of individuals, the authors were able to determine that every person differs from the reference sequence by putative loss-of-function mutations in 250-300 genes\(^297\). All data generated by the consortium is publicly available.

In a large-scale approach to decipher the complex questions posed by cancer, the International Cancer Genome Consortium is aiming to genetically characterize 50 human cancer types\(^298\). Both within and outside this project, the genomes of individuals suffering from melanoma\(^289\), leukemia\(^300\) and cancer of breast\(^301\) and lungs\(^302,303\) have been sequenced. These studies, in combination with a number of targeted resequencing studies have together resulted in the identification of more than 1000 potential cancer genes\(^304\).

Shotgun genome sequencing has already been of clinical use, for example in identifying candidate genes related to a Mendelian disease after whole-genome sequencing of a family quartet in which the parents were healthy and both children suffered from Miller Syndrome\(^305\). Moreover, the approach has been used in noninvasive genetic tests, searching for fetal chromosomal aberrations such as the presence of an extra copy of chromosome 21 in children with Down’s syndrome. By shotgun sequencing of cell-free DNA from the mother, duplication events may be identified based on difference in sequence coverage resulting in >3000 contigs\(^282,283\). Within a few weeks of the outbreak, sequence data from both platforms independently showed that the epidemic strain belonged to an \textit{E. coli} lineage that had acquired genes coding for antibiotic resistance and a toxin from another bacterial strain\(^280,289\), demonstrating the power of massive sequencing to rapidly deliver relevant clinical information.

**Resequencing**

As the cost of sequencing continues to fall, sequencing human genomes becomes more accessible and common. Resequencing of human genomes does not require long read lengths, since reads are mapped against a known reference sequence, and all current massive sequencing platforms apart from Pacific Biosciences (which presently has too low a throughput) have been used to resequence human genomes\(^282,283,289,290\). Each platform’s benchmark publications have reported the sequencing genomes of eminent people such as ‘one of the fathers of DNA’, James Watson\(^244\), or the stipulator of Moore’s law, Gordon Moore\(^295\), or anti-apartheid activist Archbishop Desmond Tutu\(^295\); and reference genomes belonging to people of various ethnic origin such as Nigeria\(^229\), Kalahari Desert and a Bantu southern Africa\(^294\), Korea\(^295\) and a Han-Chinese\(^295\) were also quickly established. Coverage usually varies between 7.4x\(^244\) and 36x\(^295\).

The ultimate goal of sequencing masses of human genomes is to, by obtaining plentiful genetic information, be able to distinguish patterns such as variants that are linked to positive/adverse affects of a particular drug treatment, or variants that are strongly or loosely associated with a particular disease. This would enable the design of drug regimes tailored to the unique genetic profiles of individual patients, the implementation of preventive treatment for individuals at elevated risk of developing a particular disease, and the creation of prognostic tools for determining the severity of disease progress. This idea is one of many that is being addressed by the 1000 Genomes Project Consortium\(^9\), a world-encompassing project that has the aim of sequencing a massive number of genomes of individuals from five major population groups, and to use this data to identify all genomic variants that are present at a frequency of 1% or more\(^296\). In a recent publication of the pilot-experiment of this huge task, involving low-coverage, high-coverage and targeted resequencing of a number of individuals, the authors were able to determine that every person differs from the reference sequence by putative loss-of-function mutations in 250-300 genes\(^297\). All data generated by the consortium is publicly available.

In a large-scale approach to decipher the complex questions posed by cancer, the International Cancer Genome Consortium is aiming to genetically characterize 50 human cancer types\(^298\). Both within and outside this project, the genomes of individuals suffering from melanoma\(^289\), leukemia\(^300\) and cancer of breast\(^301\) and lungs\(^302,303\) have been sequenced. These studies, in combination with a number of targeted resequencing studies have together resulted in the identification of more than 1000 potential cancer genes\(^304\).

Shotgun genome sequencing has already been of clinical use, for example in identifying candidate genes related to a Mendelian disease after whole-genome sequencing of a family quartet in which the parents were healthy and both children suffered from Miller Syndrome\(^305\). Moreover, the approach has been used in noninvasive genetic tests, searching for fetal chromosomal aberrations such as the presence of an extra copy of chromosome 21 in children with Down’s syndrome. By shotgun sequencing of cell-free DNA from the mother, duplication events may be identified based on difference in sequence coverage.
between different chromosomes. Several reports have demonstrated positive results with this approach and large-scale clinical studies are currently being launched in Europe and elsewhere. Shotgun sequencing of cell-free DNA from blood is also an approach that can be used to detect rejection of a transplanted organ in a patient. Since rejection causes increased apoptosis of cells in the donated organ, the level of alien DNA in the blood of the acceptor corresponds to the level of attack that the donated organ is suffering from the patient’s immune system.

4.6.2. Targeted resequencing

It has been estimated that only 5% of the human genome is functional, which is a strong argument for specifically directing sequencing capacity to a smaller portion of the genome whilst gaining increased sequence depth. Depending on the scientific question at hand, it may be desirable to sequence specific regions and either some or all protein-coding sequences (exons), using a suitable strategy for genomic partitioning.

The sequencing of PCR-amplified target regions by massive sequencing is well established; the use of amplification primers tagged with platform-specific adapters makes it possible to streamline the process of library preparation. Applications of ultra-deep amplicon sequencing include the interrogation of the subclonal phylogenetic structures of cancer by investigating variance in the most variable part of the antibody heavy-chain; comparing gut microbiomes between lean and obese twin pairs by 16S rRNA sequencing from fecal samples and HLA typing. Barcoding is often useful to increase the number of samples that may be sequenced in parallel, in order to match the ever-increasing throughput of MPS. The platforms themselves offer a number of barcodes that can be incorporated into the adapter sequences. One may wish to add an additional dimension to an experiment i.e. time point, or maximize the number of samples beyond the limitations of the inbuilt identification system, by using home-made barcodes. Barcodes may be introduced by PCR tagging, ligation or a combination thereof.

It has been predicted that about 85% of all disease-causing mutations lie within the protein-coding regions of the human genome, which makes up 1% of the complete human genome. It is often useful to focus on these regions; the practice of doing so is known as exome sequencing, and is particularly useful when studying Mendelian diseases, since these often stem from a mutation in a protein coding gene. In fact, exome sequencing has shown considerable promise for rare Mendelian diseases. By sequencing whole exomes of members of families that suffer from recessive or dominant diseases; it could be used to identify loss-of-function mutations in one of two copies of a gene, which would impact the expression of that gene but would not give rise to a Mendelian inheritance pattern. It seems that resequencing approaches that investigate a number of genes associated with a particular disease are likely to find clinical applications in the near future, as a part of various initiatives. One such project aims to develop a preconception carrier screen for prospective parents in order to determine whether they are a good match. The screen was directed at 448 severe recessive childhood diseases and was}

between different chromosomes. Several reports have demonstrated positive results with this approach and large-scale clinical studies are currently being launched in Europe and elsewhere. Shotgun sequencing of cell-free DNA from blood is also an approach that can be used to detect rejection of a transplanted organ in a patient. Since rejection causes increased apoptosis of cells in the donated organ, the level of alien DNA in the blood of the acceptor corresponds to the level of attack that the donated organ is suffering from the patient’s immune system.

4.6.2. Targeted resequencing

It has been estimated that only 5% of the human genome is functional, which is a strong argument for specifically directing sequencing capacity to a smaller portion of the genome whilst gaining increased sequence depth. Depending on the scientific question at hand, it may be desirable to sequence specific regions and either some or all protein-coding sequences (exons), using a suitable strategy for genomic partitioning.

The sequencing of PCR-amplified target regions by massive sequencing is well established; the use of amplification primers tagged with platform-specific adapters makes it possible to streamline the process of library preparation. Applications of ultra-deep amplicon sequencing include the interrogation of the subclonal phylogenetic structures of cancer by investigating variance in the most variable part of the antibody heavy-chain; comparing gut microbiomes between lean and obese twin pairs by 16S rRNA sequencing from fecal samples and HLA typing. Barcoding is often useful to increase the number of samples that may be sequenced in parallel, in order to match the ever-increasing throughput of MPS. The platforms themselves offer a number of barcodes that can be incorporated into the adapter sequences. One may wish to add an additional dimension to an experiment i.e. time point, or maximize the number of samples beyond the limitations of the inbuilt identification system, by using home-made barcodes. Barcodes may be introduced by PCR tagging, ligation or a combination thereof.

It has been predicted that about 85% of all disease-causing mutations lie within the protein-coding regions of the human genome, which makes up 1% of the complete human genome. It is often useful to focus on these regions; the practice of doing so is known as exome sequencing, and is particularly useful when studying Mendelian diseases, since these often stem from a mutation in a protein coding gene. In fact, exome sequencing has shown considerable promise for rare Mendelian diseases. By sequencing whole exomes of members of families that suffer from recessive or dominant diseases; it could be used to identify loss-of-function mutations in one of two copies of a gene, which would impact the expression of that gene but would not give rise to a Mendelian inheritance pattern. It seems that resequencing approaches that investigate a number of genes associated with a particular disease are likely to find clinical applications in the near future, as a part of various initiatives. One such project aims to develop a preconception carrier screen for prospective parents in order to determine whether they are a good match. The screen was directed at 448 severe recessive childhood diseases and was
developed for implementation in IVF clinics\textsuperscript{47}. In addition, the sequencing of exons of genes involved in nonsyndromic hearing loss (NSHL) has proven useful for diagnosing patients with this disease\textsuperscript{43}.

4.6.3. Expression profiling

The mixture of proteins present in a cell determines its characteristics and functions, and is indicative of normal or pathological cellular conditions. As such, large-scale analysis of the protein makeup of different cell types is interesting in many aspects of research. Mass Spectrometry (MS) is the most widely used tool in global proteomics studies and has been extensively developed over the last twenty years\textsuperscript{256}. However, its reliance on complicated sample preparation procedures, limited dynamic range and the lack of methods for protein amplification mean that it would be useful to be able to complement proteomics data with quantitative studies on mRNA, which is relatively easy to quantify on a large scale. However, because the pathway by which mRNA is used in protein synthesis is regulated by a variety of factors, including the rate of mRNA and protein degradation, and variable translational efficiency, it is not accurate to assume that mRNA levels will always mirror those of the corresponding proteins\textsuperscript{331,337}. Nevertheless, expression profiling has been used extensively to obtain information on cell function and the underlying mechanisms of diseases and cellular processes.

Interest in RNA analysis peaked in the late 1990s, with the arrival of scalable and low-cost hybridization-based microarray technologies\textsuperscript{330,332,335,336}. As massive sequencing spread, microarray analysis of gene expression will more or less be replaced by sequencing-based expression analysis (RNA-Seq) in the near future, which requires no prior knowledge of the target sequence and therefore makes it possible to identify and quantify unknown transcripts such as fusion genes and novel splice variants\textsuperscript{324,327,328}. Briefly, RNA is converted into cDNA and then further into short pieces, known as ‘tags’, that stem from the original transcript. ‘Tags’ are sequenced and reads are mapped to the reference genome. By counting the mapping tags, a quantitative measure of gene expression is obtained with a greater sensitivity and dynamic range than is possible with microarray based analysis\textsuperscript{35}.

Since RNA-Seq was first presented in 2008, it has been further developed in a variety of ways. In the original RNA sequencing studies, in order to find splice-sites, reads were mapped to exon-junction sequences that are unique for particular transcripts\textsuperscript{337}. With the growing read lengths of current platforms, it is now easier to recognize splice-junctions without having to search for specific ones. Also, with the extreme read lengths of emerging single molecule technologies, full-length cDNA/RNA sequencing will likely soon be possible, enabling the identification of splice events in a completely unbiased fashion. The number of classified RNA species has increased greatly in the last year and it is now known that transcripts are also synthesized from the antisense DNA strand. Antisense transcript are believed to primarily be involved in regulatory mechanisms, whereas the sense version of a transcript is more likely to give rise to a functional protein\textsuperscript{44}. Moreover, other types of noncoding RNA species may also be transcribed from both strands of DNA. Information on the original strand that an RNA molecule stems from is therefore useful. There are a number of strand-specific sample preparation techniques available\textsuperscript{332,333,334,335}; alternatively, it is possible to directly sequence first strand cDNA\textsuperscript{332,333,334} or mRNA\textsuperscript{335}.

Since RNA-Seq was first presented in 2008, it has been further developed in a variety of ways. In the original RNA sequencing studies, in order to find splice-sites, reads were mapped to exon-junction sequences that are unique for particular transcripts\textsuperscript{337}. With the growing read lengths of current platforms, it is now easier to recognize splice-junctions without having to search for specific ones. Also, with the extreme read lengths of emerging single molecule technologies, full-length cDNA/RNA sequencing will likely soon be possible, enabling the identification of splice events in a completely unbiased fashion. The number of classified RNA species has increased greatly in the last year and it is now known that transcripts are also synthesized from the antisense DNA strand. Antisense transcript are believed to primarily be involved in regulatory mechanisms, whereas the sense version of a transcript is more likely to give rise to a functional protein\textsuperscript{44}. Moreover, other types of noncoding RNA species may also be transcribed from both strands of DNA. Information on the original strand that an RNA molecule stems from is therefore useful. There are a number of strand-specific sample preparation techniques available\textsuperscript{332,333,334,335}; alternatively, it is possible to directly sequence first strand cDNA\textsuperscript{332,333,334} or mRNA\textsuperscript{335}.

Interest in RNA analysis peaked in the late 1990s, with the arrival of scalable and low-cost hybridization-based microarray technologies\textsuperscript{330,332,335,336}. As massive sequencing spread, microarray analysis of gene expression will more or less be replaced by sequencing-based expression analysis (RNA-Seq) in the near future, which requires no prior knowledge of the target sequence and therefore makes it possible to identify and quantify unknown transcripts such as fusion genes and novel splice variants\textsuperscript{324,327,328}. Briefly, RNA is converted into cDNA and then further into short pieces, known as ‘tags’, that stem from the original transcript. ‘Tags’ are sequenced and reads are mapped to the reference genome. By counting the mapping tags, a quantitative measure of gene expression is obtained with a greater sensitivity and dynamic range than is possible with microarray based analysis\textsuperscript{35}.

Since RNA-Seq was first presented in 2008, it has been further developed in a variety of ways. In the original RNA sequencing studies, in order to find splice-sites, reads were mapped to exon-junction sequences that are unique for particular transcripts\textsuperscript{337}. With the growing read lengths of current platforms, it is now easier to recognize splice-junctions without having to search for specific ones. Also, with the extreme read lengths of emerging single molecule technologies, full-length cDNA/RNA sequencing will likely soon be possible, enabling the identification of splice events in a completely unbiased fashion. The number of classified RNA species has increased greatly in the last year and it is now known that transcripts are also synthesized from the antisense DNA strand. Antisense transcript are believed to primarily be involved in regulatory mechanisms, whereas the sense version of a transcript is more likely to give rise to a functional protein\textsuperscript{44}. Moreover, other types of noncoding RNA species may also be transcribed from both strands of DNA. Information on the original strand that an RNA molecule stems from is therefore useful. There are a number of strand-specific sample preparation techniques available\textsuperscript{332,333,334,335}; alternatively, it is possible to directly sequence first strand cDNA\textsuperscript{332,333,334} or mRNA\textsuperscript{335}.
cDNA synthesis and forthcoming amplification come with a risk of introducing errors and bias due to reverse transcriptase template switching, the enzyme’s preference for longer transcripts and even GC content, and its talent for creating spurious second-strand cDNAs due to its DNA-dependent DNA polymerase activity. Problematic second-strand synthesis and amplification steps can be avoided by instead sequencing the first strand cDNA. This has been done on the Helicos platform by capturing polyadenylated cDNA molecules on the flow cell or by carrying out the actual cDNA synthesis straight on the sequencing flow cell, an approach presented by both Helicos and Illumina. Taken to its extreme, this approach would result in the sequencing of RNA molecules without the prior conversion of RNA into cDNA; this would completely avoid the problems associated with the modest efficiency of cDNA synthesis and its tendency to create false cDNA molecules and introduce bias into the resulting sample. Direct RNA sequencing is currently possible only with the Helicos platform, and is done by capturing 5'-polyadenylated RNA molecules in vitro on the flow cell and then sequencing these in a manner analogous to Helicos DNA sequencing, but with the DNA polymerase exchanged for a reverse transcriptase; less than one 1 ng of RNA is required in this process. The limited read length of the Helicos platform is a shortcoming, since it affects the scope for mapping; in the proof of concept paper only 50% of the reads were >18 bp long. However, a major advantage of the method is the ability to analyze partly-degraded samples due to the lack of amplification steps.

In clinical applications of RNA-seq, as well as stem cell analysis and forensics, methods for the global profiling of minute amounts of RNA are important. Some sample preparation protocols have addressed this issue, enabling RNA-seq of scarce material or even single cells by e.g. using a linear isothermal amplification strategy (SPIA) to amplify double-stranded cDNA, double PCR amplification or by using the template-switching activity of the reverse transcriptase in order to incorporate a cell-specific barcoded adapter and a general adapter sequence in each end of a first strand cDNA molecule. Transcriptome profiling has shed light on vastly different subjects such as the polyadenylation pattern of human cells and embryonic stem cell maturation. In complex diseases a number of alleles that are common among people can cause a slight increase in the risk of contracting a specific disease, and these alleles usually have a minor effect on gene expression. It is thus interesting to study changes in gene expression. An in-depth analysis of expression patterns in combination with structural genomic data from 10 melanoma samples identified 11 novel melanoma gene-fusions. The study also showed that there were a large number of somatic mutations that accumulated in these melanomas, which supports the idea that point mutations are the major driving force behind tumor progression. Tasmanian devils are an endangered species that frequently suffer from an unusual, infectious facial cancer, which spreads through physical transfer of living cancer cells when the animals bite each other (which is apparently common in mating and other encounters). By transcriptome profiling of tumor and a testis from the same devil, it was shown that the tumor first occurred in the nervous-system-associated Schwann cells of the original animal. Through a number of mutations, the tumor then somehow acquired the ability to bypass the immune system of other animals so that when spread by biting, they were able to act like a tissue engraftment, forming a tissue in the new animal.
5. The papers

This thesis is based on five papers that cover the process from picking cells to analyzing their genetic makeup.

In the first paper we present a miniaturized chip and use it to seed single stem cells into separate wells, and then culture and analyze maintained pluripotency and differentiation, in a high throughput manner. Then we move on to massively parallel sequencing. In paper II-IV we develop ways to manipulate template-carrying beads for sequencing by Roche/454 sequencing, in order to exploit the output that one obtains from a sequencing run. This is done (i) by selecting beads that have amplified template on the surface, or (ii) beads that carry a particular target sequence or (iii) barcode sequence tag. In the last paper we use massively parallel sequencing on single cell material to interrogate the cell-to-cell variation in hyper variable repeat regions in the genome.
5.1. Culturing of single stem cells in a microchip (I)

The intriguing ability of stem cells to self-renew or differentiate into numerous cell types (pluripotency) has for instance been demonstrated by the ability of an individual stem cell to generate a functional mammary gland[41] or prostate[42] upon transplantation in nude mice. Stem cell differentiation into distinct cell types occurs through changes in cellular biochemistry, giving rise to specialized cells with different functions of a tissue or organ. In vitro culturing of mammalian adult and embryonic neural stem cells, while maintaining multi-or-pluripotency is obtained by addition of particular growth factors. New growth factors and mixtures thereof, are sought for and extensively tested in large-scale screening studies, for effect in maintaining pluripotency or for directing stem cell differentiation or reprogram mature stem cells into regaining pluripotency[43-44]. The cost of the growth factors used are steep, thus decreased screening volumes have large effects on the cost of the assay. Embryonic stem cells are typically cultured adherently by coating plates with extracellular matrix proteins, while adult stem cells may grow adherently or as free-floating spherical cell balls, so called neurospheres (NS), depending on culturing conditions. A number of microowell devices are available for cell culturing/analysis, with a wide span of well counts and sizes, among which the 96 and 384 well plates are the most commonly used.

In paper I we present a miniaturized stem cell screening chip and use of it for culturing and analysis of mouse and human embryonic stem (ES) cells and adult neural stem cells from mouse forebrain. The high-density microowell chip is glass bottomed to allow easy analysis by microscopy and imaging systems. The walls of the wells are made up from etched silicon and sloped to allow cells ‘rolling down’ to the bottom[32]. The chip is the size of an ordinary microscopic slide and contains 672 individually labeled microwells, with 500 nl volume. The chip layout is carefully selected to match the shortest movement of the plate holder of ordinary flow cytometers (i.e. 1500 µm centre-to-centre distance between wells) to enable directed single cell seeding by FACS. With minor changes, a similar chip had previously been used to studying carcinoma cell heterogeneity[32]. To prevent liquid evaporation while allowing gas exchange and easy interrogation by microscopy, the chip is covered with a semi-permeable PDMS membrane.

The applicability of the chip for customary culturing conditions and assays was demonstrated for the three types of stem cells. Differentiation was stimulated by supplementing differentiation medium to seeded single cells in microwells. After nine days differentiation was verified by antibody labeling for a neuronal differentiation marker. Long-term culturing and clone formation with maintained pluripotency is interesting for investigating stem cell proliferation and was demonstrated in the current microowell chip. Maintained pluripotent state of the two ES cell samples and an adherent culture of adult neural stem cells, was confirmed four days after single cell seeding, by antibody labeling for a pluripotency marker. Adult neural stem cells were also cultured in a fashion that encourages neurosphere formation, which indicates pluripotency. Indeed after four days sphere formation was observed. The microowell chip presented here, facilitates monitoring and manipulation of cells individually or en masse, which is useful in high throughput differentiation screening assays and clonal experiments.

---

5.1. Culturing of single stem cells in a microchip (I)

The intriguing ability of stem cells to self-renew or differentiate into numerous cell types (pluripotency) has for instance been demonstrated by the ability of an individual stem cell to generate a functional mammary gland[41] or prostate[42] upon transplantation in nude mice. Stem cell differentiation into distinct cell types occurs through changes in cellular biochemistry, giving rise to specialized cells with different functions of a tissue or organ. In vitro culturing of mammalian adult and embryonic neural stem cells, while maintaining multi-or-pluripotency is obtained by addition of particular growth factors. New growth factors and mixtures thereof, are sought for and extensively tested in large-scale screening studies, for effect in maintaining pluripotency or for directing stem cell differentiation or reprogram mature stem cells into regaining pluripotency[43-44]. The cost of the growth factors used are steep, thus decreased screening volumes have large effects on the cost of the assay. Embryonic stem cells are typically cultured adherently by coating plates with extracellular matrix proteins, while adult stem cells may grow adherently or as free-floating spherical cell balls, so called neurospheres (NS), depending on culturing conditions. A number of microowell devices are available for cell culturing/analysis, with a wide span of well counts and sizes, among which the 96 and 384 well plates are the most commonly used.

In paper I we present a miniaturized stem cell screening chip and use of it for culturing and analysis of mouse and human embryonic stem (ES) cells and adult neural stem cells from mouse forebrain. The high-density microowell chip is glass bottomed to allow easy analysis by microscopy and imaging systems. The walls of the wells are made up from etched silicon and sloped to allow cells ‘rolling down’ to the bottom[32]. The chip is the size of an ordinary microscopic slide and contains 672 individually labeled microwells, with 500 nl volume. The chip layout is carefully selected to match the shortest movement of the plate holder of ordinary flow cytometers (i.e. 1500 µm centre-to-centre distance between wells) to enable directed single cell seeding by FACS. With minor changes, a similar chip had previously been used to studying carcinoma cell heterogeneity[32]. To prevent liquid evaporation while allowing gas exchange and easy interrogation by microscopy, the chip is covered with a semi-permeable PDMS membrane.

The applicability of the chip for customary culturing conditions and assays was demonstrated for the three types of stem cells. Differentiation was stimulated by supplementing differentiation medium to seeded single cells in microwells. After nine days differentiation was verified by antibody labeling for a neuronal differentiation marker. Long-term culturing and clone formation with maintained pluripotency is interesting for investigating stem cell proliferation and was demonstrated in the current microowell chip. Maintained pluripotent state of the two ES cell samples and an adherent culture of adult neural stem cells, was confirmed four days after single cell seeding, by antibody labeling for a pluripotency marker. Adult neural stem cells were also cultured in a fashion that encourages neurosphere formation, which indicates pluripotency. Indeed after four days sphere formation was observed. The microowell chip presented here, facilitates monitoring and manipulation of cells individually or en masse, which is useful in high throughput differentiation screening assays and clonal experiments.
5.2. **Bead sorting approaches for improved performance in massively parallel sequencing (II-IV)**

In three current massive parallel sequencing platforms, 454, SOLiD and Ion Torrent, template molecules are propagated on the surface of beads in emulsion, to obtain beads covered in thousands to millions of identical template copies, for sequencing. Beads with more than one amplified template on the surface give rise to mixed and unreadable sequencing signals and take up valuable sequencing space. Choosing the optimal DNA-to-bead ratio is a trade-off between reducing the number of mixed beads, while still obtaining sufficient amounts of beads with amplified template on the surface and the value varies between libraries. Poison bead to droplet distribution and poison template to droplet distribution normally renders less than 1/5 of the resulting beads DNA covered\(^{1,49}\), and naturally these need to be enriched prior to sequencing.

In the sequencing protocols, enrichment is done either by coupling DNA-covered beads to large low-density enrichment beads and separating based on density\(^{297}\) or by coupling DNA covered beads to paramagnetic beads and gather these using a magnet\(^{10,28}\). Both methods require substantial manual washing steps, in order to remove empty beads. DNA-covered beads are then loaded into individual wells for sequencing by synthesis or ligations, in a wash-and-scan manner. Although flow cytometry were originally developed for cell analysis, the apparatuses can also be used for sorting and analyzing beads\(^{295,30,31}\). In paper IV flow cytometry has been used in different ways to improve the performance in massive parallel pyrosequencing.

5.2.1. **Library titration and collection of DNA covered beads (II)**

In paper II we used flow cytometry to evaluate DNA libraries and to enrich for DNA covered beads. The enrichment strategies used in SOLiD, Ion Torrent and 454 have varying efficiency, and when empty beads entering the sequencing reaction these waste valuable sequence space. In paper II we demonstrate that by fluorescently labeling DNA covered beads in an emulsion PCR sample and high-speed flow-sorting, DNA covered beads could rapidly be obtained with an immaculate purity. Samples were labeled by using a generic fluorescently labeled probe, or streptavidin-coupled fluorophore targeting a biotin on the non-attached strands. After FACS sorting the fluorescent probes were removed by alkali washing and the beads were sequenced. Sequencing of the FACS sorted beads demonstrated that these beads and the DNA on their surface were intact and that the sequence quality was not affected by the procedure. Moreover, we demonstrated that a simple flow cytometry experiment could replace the laborious and time-consuming titration protocol of the initial Genome Sequencer system.

At the time of the launch of the 454 and SOLiD platforms, the only available way of titrating a DNA library, was by preparing several (usually four) miniature emulsion PCR reactions, with different amount of input DNA and then evaluate the outcome by sequencing. Titration sequencing is a costly, laborious and time-consuming procedure; hence there was a need for simpler protocol(s). In paper II we fluorescently labeled DNA beads in the miniature emulsion PCR reactions and analyzed these by flow cytometry. In the standard protocol titration curves were drawn based on the proportion of loaded beads.

5.2. **Bead sorting approaches for improved performance in massively parallel sequencing (II-IV)**

In three current massive parallel sequencing platforms, 454, SOLiD and Ion Torrent, template molecules are propagated on the surface of beads in emulsion, to obtain beads covered in thousands to millions of identical template copies, for sequencing. Beads with more than one amplified template on the surface give rise to mixed and unreadable sequencing signals and take up valuable sequencing space. Choosing the optimal DNA-to-bead ratio is a trade-off between reducing the number of mixed beads, while still obtaining sufficient amounts of beads with amplified template on the surface and the value varies between libraries. Poison bead to droplet distribution and poison template to droplet distribution normally renders less than 1/5 of the resulting beads DNA covered\(^{1,49}\), and naturally these need to be enriched prior to sequencing.

In the sequencing protocols, enrichment is done either by coupling DNA-covered beads to large low-density enrichment beads and separating based on density\(^{297}\) or by coupling DNA covered beads to paramagnetic beads and gather these using a magnet\(^{10,28}\). Both methods require substantial manual washing steps, in order to remove empty beads. DNA-covered beads are then loaded into individual wells for sequencing by synthesis or ligations, in a wash-and-scan manner. Although flow cytometry were originally developed for cell analysis, the apparatuses can also be used for sorting and analyzing beads\(^{295,30,31}\). In paper IV flow cytometry has been used in different ways to improve the performance in massive parallel pyrosequencing.

5.2.1. **Library titration and collection of DNA covered beads (II)**

In paper II we used flow cytometry to evaluate DNA libraries and to enrich for DNA covered beads. The enrichment strategies used in SOLiD, Ion Torrent and 454 have varying efficiency, and when empty beads entering the sequencing reaction these waste valuable sequence space. In paper II we demonstrate that by fluorescently labeling DNA covered beads in an emulsion PCR sample and high-speed flow-sorting, DNA covered beads could rapidly be obtained with an immaculate purity. Samples were labeled by using a generic fluorescently labeled probe, or streptavidin-coupled fluorophore targeting a biotin on the non-attached strands. After FACS sorting the fluorescent probes were removed by alkali washing and the beads were sequenced. Sequencing of the FACS sorted beads demonstrated that these beads and the DNA on their surface were intact and that the sequence quality was not affected by the procedure. Moreover, we demonstrated that a simple flow cytometry experiment could replace the laborious and time-consuming titration protocol of the initial Genome Sequencer system.

At the time of the launch of the 454 and SOLiD platforms, the only available way of titrating a DNA library, was by preparing several (usually four) miniature emulsion PCR reactions, with different amount of input DNA and then evaluate the outcome by sequencing. Titration sequencing is a costly, laborious and time-consuming procedure; hence there was a need for simpler protocol(s). In paper II we fluorescently labeled DNA beads in the miniature emulsion PCR reactions and analyzed these by flow cytometry. In the standard protocol titration curves were drawn based on the proportion of loaded beads.
with a readable key sequence (% keypass). The key sequence is a 4-base sequence that is common to all template molecules and used for instrument calibration. Since also mixed beads have a readable key sequence, the value of % keypass corresponds to the total proportion of DNA covered beads in the sample. Likewise, labeled emulsion PCR sample analyzed by flow cytometry, determines the proportion of DNA covered beads in the sample. We demonstrated that virtually identical titration curves for a particular library, based on which the optimal DNA:bead ratio can be determined, could rapidly be obtained from the flow cytometric analysis.

5.2.2. Selecting beads with a particular gene fragment (III)
Deep sequencing of hypervariable genetic regions has shown useful for various applications such as studies on tumor development and high-resolution human leukocyte antigen typing. Amplicon sequencing is normally carried out by PCR amplifying target regions while also introducing suitable adapters for massive sequencing as well as identity tags. PCR amplification sometimes causes unspecific product formation, which is problematic as these fragments due to their shorter length, outcompete specific target products in the ordinary PCR. Methods for eliminating unspecific fragments, if sufficient size difference exists, include precipitation onto carboxylic acids coated paramagnetic beads, size separation by gel electrophoresis (grikut) and nowadays also commercial systems for automated gerkut, Pippin Prep (Sage Science) and LabChip XT (Caliper). However, these approaches are work intensive and require sufficient size differences.

In paper III the second exon in Dog leukocyte antigen (DLA) in 34 dogs was investigated by deep sequencing of tagged amplicons. In this paper we took the FACS enrichment approach one step further by specifically labeling beads carrying the target sequence template to remove beads with amplified unspecific product that includes the A/B linker, on the surface. After emPCR of the pooled library, beads were fluorescently labeled with a target-specific probe, combined with random hexamers and the DNA double-strand specific dye SYBRgreen. Target-bearing beads were collected by flow cytometric sorting based on target-specific dye and DNA quantity. The FACS sorting protocol rendered virtually a three-fold increase in target reads as compared to the same emulsion PCR sample enriched according to standard protocol. Naturally, since the majority of the short unspecific product-bearing beads were removed, there was also a vast increase in average read length. An assessment of sequence distribution demonstrated that the 50 bp long probe used for labeling allowed many mismatches and did not impact the representation of different individuals or genotypes in the samples. Gene-specific FACS sorting can thus be used to enhance sequence output of problematic amplicon libraries.

with a readable key sequence (% keypass). The key sequence is a 4-base sequence that is common to all template molecules and used for instrument calibration. Since also mixed beads have a readable key sequence, the value of % keypass corresponds to the total proportion of DNA covered beads in the sample. Likewise, labeled emulsion PCR sample analyzed by flow cytometry, determines the proportion of DNA covered beads in the sample. We demonstrated that virtually identical titration curves for a particular library, based on which the optimal DNA:bead ratio can be determined, could rapidly be obtained from the flow cytometric analysis.

5.2.2. Selecting beads with a particular gene fragment (III)
Deep sequencing of hypervariable genetic regions has shown useful for various applications such as studies on tumor development and high-resolution human leukocyte antigen typing. Amplicon sequencing is normally carried out by PCR amplifying target regions while also introducing suitable adapters for massive sequencing as well as identity tags. PCR amplification sometimes causes unspecific product formation, which is problematic as these fragments due to their shorter length, outcompete specific target products in the ordinary PCR. Methods for eliminating unspecific fragments, if sufficient size difference exists, include precipitation onto carboxylic acids coated paramagnetic beads, size separation by gel electrophoresis (grikut) and nowadays also commercial systems for automated gerkut, Pippin Prep (Sage Science) and LabChip XT (Caliper). However, these approaches are work intensive and require sufficient size differences.

In paper III the second exon in Dog leukocyte antigen (DLA) in 34 dogs was investigated by deep sequencing of tagged amplicons. In this paper we took the FACS enrichment approach one step further by specifically labeling beads carrying the target sequence template to remove beads with amplified unspecific product that includes the A/B linker, on the surface. After emPCR of the pooled library, beads were fluorescently labeled with a target-specific probe, combined with random hexamers and the DNA double-strand specific dye SYBRgreen. Target-bearing beads were collected by flow cytometric sorting based on target-specific dye and DNA quantity. The FACS sorting protocol rendered virtually a three-fold increase in target reads as compared to the same emulsion PCR sample enriched according to standard protocol. Naturally, since the majority of the short unspecific product-bearing beads were removed, there was also a vast increase in average read length. An assessment of sequence distribution demonstrated that the 50 bp long probe used for labeling allowed many mismatches and did not impact the representation of different individuals or genotypes in the samples. Gene-specific FACS sorting can thus be used to enhance sequence output of problematic amplicon libraries.
5.2.3. Normalizing barcode presence (IV)

Barcoding and pooling of multiple samples is often useful to increase the number of samples that may be sequenced in parallel [531-534] and the 454 platform offers numerous short barcodes denoted Multiplex Identity (MID) tags, that are incorporated in the sequencing adapters. Multiplex procedures require accurate quantification of the individual targets before library pooling and emulsion PCR. However, despite careful quantification prior to pooling, highly biased library fractions have been reported [535-536].

In paper IV we used color-coded barcode-specific oligonucleotide probes to specifically label beads based on barcode. To enable sufficient hybridization probes were designed to cover the 10 bp long MID sequence as well as the four 6 bp long key sequence and two bp in the general sequencing primer site. In a proof-of-concept experiment a sample containing four MID libraries in highly biased fractions were MID-specifically labeled and equal amounts of beads from all libraries were collected by FACS sorting. Library equalization was confirmed by sequencing the FACS enriched beads. Beads with more than one amplified fragment on the surface, mixed beads, would most likely (unless carrying several fragments with the same barcode) be carrying more than one fluorescent dye. Since only single-dye labeled beads were collected in the flow-sorting process the proportion of mixed beads was vastly reduced by the protocol. Collection of single template beads, with a large number of amplified templates on the surface (i.e. high fluorescent signal) rendered an increased proportion of high quality sequence reads in the FACS sorted sample.

5.3. Investigating polyG variation between individual cells (V)

Cell heterogeneity is of key importance in many aspects of disease, such as in tumor development, and can also be used to understand how cells organize themselves into complex tissues initiated at the stage of embryogenesis. Over time, cells further away from each other in the hereditability tree become more dissimilar than cells that are more closely related. By studying the level of similarity between the polymers of life (i.e. genes, proteins or parts thereof) in different animal species, phylogenetic maps of interspecies relationships have been drawn since 1960 [537].

When studying cells that are more closely related to each other than those from different species, the markers need to be highly variable or in large numbers in order to carry sufficient information on hereditary relationships. Mutations are more likely to occur in genomic regions with low-complexity such as repeat sequences. Among repeating units, polyguanine nucleotide repeats have a significantly higher mutation rate than other types of repeats, such as polyA- or CA-repeats [538-539]. Phylogenetic fane maps, that describe how various cells within one organism are related to one another, have been drawn by tracking differences in length of highly variable polyguanine nucleotide repeats in single cells from different organs in mice [54]. Discrimination between repeat-lengths was here performed by PCR and fragment analysis. Thus far, fair-mapping based on multiplex fragment length
differences has been used, and the method inherently limits the number of loci that can be investigated. In this study we have approached large-scale analysis of length altering mutations in polyguanine tracts of human T-cells by massively parallel sequencing.

The high variability of polyguanine nucleotide repeats is mainly due to polymerase slippage \(^1\). Since polymerases are used in every in vitro amplification step from DNA to sequencing, this poses a problem for analysis, in particular when studying scarce material or single cells. Moreover, highly biased GC fractions are known to impact PCR amplification in a negative manner \(^2\). The sequencing chemistry used in Illumina is meant to enable homopolymer sequencing with no difficulties \(^3\), however since the technique still involves PCR amplification, there may be limitations. By using a model system, the ability to accurately Illumina sequence polyguanine nucleotide repeats, ranging in lengths from 3-20 nucleotides, was evaluated. A great difficulty in correctly establishing lengths exceeding 10-15 basepairs was demonstrated.

High-coverage whole genome sequence data from a multi-cell sample from a male individual was used to evaluate variability and sequence coverage in polyguanine nucleotide repeats over the genome. It was evident that polyguanine repeat lengths for our individual, determined by Illumina sequence reads were in general shorter than those present in the human reference genome. Cell to cell variability was investigated in low-coverage sequence data from individual T-cells originating from the same individual as the multi-cell sample. A comparison between multi-cell data and single cell data verified maintained quality of the single cell data however, allelic dropout of a portion of the loci was observed, presumably due to the low coverage in combination with the effects of whole genome amplification of single cell material.

In an attempt to focus sequencing efforts to a selected number of polyguanine nucleotide repeat loci, we employed the method Tri-nucleotide threading (TriT). In a new way, denoted Mono-nucleotide threading (MnT), to selectively amplify polyguanine repeats from single cell material. The method uses an initial extension and ligation step in a cyclic manner to introduce universal amplification primers in each amplified fragment in a linear fashion. The universal amplification handles are then used in a universal PCR amplification step, of all threads in parallel. By designing MnT-primers so that they hybridize adjacent to the polyguanine repeats, and by only including dCTPs in the linear amplification reaction, an increased selectivity is approached. By selectively amplifying a set of polyguanine loci from single fibroblast cells, for which the phylogenetic relationship was known, we demonstrated the applicability of using MnT to selectively amplify polyG loci, and construct a phylogenetic tree based on the differences in polyG length between the cells.

5.4. Concluding remarks and future perspectives

Recent advances in biotechnology have enabled completely new avenues in life science research to be explored. By allowing increased parallelization an ever-increasing complexity of cell samples or experiments can be investigated in shorter time and at a lower cost. This facilitates for example large-scale efforts to study cell heterogeneity at the single cell level that may also include global genome analyses. The work presented in this thesis focuses on
massively parallel analysis of cells or nucleic acid samples, demonstrating technology developments in the field as well as use of the technology in life sciences.

The speed of development of the gene technology field in the last five years has been dramatic, with protocol changes and upgrades giving swifter protocols and more and better sequence reads. Since the advent of SOLiD and 454 sequencing, the standard library titration protocols have been changed and library titration in these platforms are nowadays carried out by sequencing but by quantifying the number of enriched ([i.e. DNA covered] beads from the miniaturized PCR reactions. Yet other methods for library titration are available that do not include emulsion PCR. The latter methods offer an advantage over previous procedures by rendering a vast decrease in protocol time\(^{19,20,22,24}\). It is useful to have control over the input prior to sequencing, in order to maximize the output. Flow-sorting of DNA capture beads can be used to collect beads with different features depending on application, as demonstrated in this thesis for the 454 sequencing platform.

Ultra-deep sequencing has shed light on various issues such as cancer development and human leukocyte antigen typing\(^{18,23}\). Gene-specific flow-sorting solves a frequent problem in amplicon sequencing studies, namely the plentiful low quality reads in sequencing data. On the same track is barcode-specific labeling and flow-sorting in which normalization of a pooled sample containing four barcoded libraries was carried out in order to obtain even levels of all four libraries and with increased sequence quality. With current instrumentation the method should be useful for up to 16 libraries in one reaction. Since half of the massive sequencing platforms available today use beads as solid support for template amplification, methods for handling beads in enrichment and selection in order to increase the number of useful reads are of relevance. Today, investment in an expensive flow cytometer for this purpose, can probably not be fully justified. However, a downscaled flow-sorter, specialized for automated bead-enrichment, could come in handy in sample preparation for core laboratories that have a steady stream of samples coming through, in order to maximize the output of each run.

As sequencing continues to develop to offer more data for less money, enabling enormous sequencing initiatives such as 1000 Genomes project\(^{26}\) and the Cancer genomes project\(^{55}\) that are currently being carried out, the knowledge of the connection between certain diseases and genetic signatures will increase. This will give rise to new diagnostic markers for the clinic, and increase the attention for personalized medicine.

Cell-cell variability is interesting in many biological contexts. In some diseases, a combination of traits is known to give a cell its particular disease phenotype and obtaining information on cell heterogeneity within a tissue is then crucial. For instance, it has been argued that knowing more about the co-occurrence of several different mutations within the same cell is essential to fully understand the nature and evolution of cancer\(^{18,19}\). In other cases, a cell population may be made up of two subpopulations with distinctly varying traits, making the information obtained dependant upon their proportions in a mixed sample. Circulating tumor cells in cancer patients and fetal cells in the blood of pregnant women are two examples of rare cells that are difficult to study in their natural habitat and therefore need to be isolated for proper characterization of cancer genotype, and risk of metastasis or fetal aneuploidy, respectively. High density miniaturized wells suitable for high throughput cell screening of individually seeded and possibly clonally expanded cells,
will most likely prove useful in studies focusing on e.g. directed differentiation or reprogramming of stem cells, and may help in making screening studies more accessible.

Whole genome sequencing of single cell material was used to investigate cell-to-cell variation in the most variable portion of the human genome, polyguanine nucleotide repeat regions. Starting from single cells, the challenges were many, including amplification of genomic DNA while introducing minimal bias and investigating the ability to accurately sequence polyguanine nucleotide repeats by Illumina HiSeq sequencing. A number of highly variable regions were identified in multi-cell data from the same individual from which the single cells were collected, and the variability in these loci was confirmed in the single cell data. By selectively amplifying and sequencing a set of polyguanine loci from eight cells for which the phylogenetic relationship was known, we further demonstrated the ability of constructing phylogenetic trees based on massively parallel sequencing of amplified polyguanine loci.

Single cell sequencing approaches using whole genome amplification are highly interesting in areas such as tumor development and stem cell research, nevertheless these approaches come with major technical limitations and are currently limited to low coverage (i.e. around 0%)\(^{16}\). The advent of new sequencing technologies that require less input material and with no inherent amplification steps, may lead to new opportunities in this field within the near future.
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CNV</td>
<td>copy number variation</td>
</tr>
<tr>
<td>cPAL</td>
<td>combinatorial probe-anchor ligation</td>
</tr>
<tr>
<td>dATP</td>
<td>deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>deoxycytidine triphosphate</td>
</tr>
<tr>
<td>ddNTPs</td>
<td>dideoxy nucleotides</td>
</tr>
<tr>
<td>dGTP</td>
<td>deoxyguanine triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DOP-PCR</td>
<td>degenerate oligonucleotide-primed PCR</td>
</tr>
<tr>
<td>ds</td>
<td>double-stranded</td>
</tr>
<tr>
<td>dTTP</td>
<td>deoxythymidine triphosphate</td>
</tr>
<tr>
<td>ES</td>
<td>embryonic stem (cell)</td>
</tr>
<tr>
<td>FACs</td>
<td>fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FRET</td>
<td>fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>Gb</td>
<td>giga bases</td>
</tr>
<tr>
<td>gDNA</td>
<td>genomic DNA</td>
</tr>
<tr>
<td>HDA</td>
<td>helicase-dependent amplification</td>
</tr>
<tr>
<td>kbp</td>
<td>kilo base pairs</td>
</tr>
<tr>
<td>LCM</td>
<td>laser capture microdissection</td>
</tr>
<tr>
<td>LCR</td>
<td>ligase chain reaction</td>
</tr>
<tr>
<td>LR-PCR</td>
<td>long-range PCR</td>
</tr>
<tr>
<td>MDA</td>
<td>multiple displacement amplification</td>
</tr>
<tr>
<td>MIP</td>
<td>molecular inversion probe</td>
</tr>
<tr>
<td>MLPA</td>
<td>multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDSMS</td>
<td>polydimethylsiloxane</td>
</tr>
<tr>
<td>PEP</td>
<td>primer-extension preamplification</td>
</tr>
<tr>
<td>PPi</td>
<td>inorganic phosphate</td>
</tr>
<tr>
<td>RCA</td>
<td>rolling circle amplification</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>real-time PCR</td>
</tr>
<tr>
<td>SBH</td>
<td>sequencing-by-hybridization</td>
</tr>
<tr>
<td>SBT</td>
<td>sequencing-by-identification</td>
</tr>
<tr>
<td>SBS</td>
<td>sequencing-by-synthesis</td>
</tr>
<tr>
<td>SMRT</td>
<td>single-molecule real-time</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOLiD</td>
<td>sequencing by oligonucleotide ligation and detection</td>
</tr>
<tr>
<td>w</td>
<td>single-stranded</td>
</tr>
<tr>
<td>WGA</td>
<td>whole-genome amplification</td>
</tr>
<tr>
<td>ZMW</td>
<td>zero-mode waveguide</td>
</tr>
</tbody>
</table>

---

## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>ChIP</td>
<td>chromatin immunoprecipitation</td>
</tr>
<tr>
<td>CNV</td>
<td>copy number variation</td>
</tr>
<tr>
<td>cPAL</td>
<td>combinatorial probe-anchor ligation</td>
</tr>
<tr>
<td>dATP</td>
<td>deoxyadenosine triphosphate</td>
</tr>
<tr>
<td>dCTP</td>
<td>deoxycytidine triphosphate</td>
</tr>
<tr>
<td>ddNTPs</td>
<td>dideoxy nucleotides</td>
</tr>
<tr>
<td>dGTP</td>
<td>deoxyguanine triphosphate</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleotide triphosphate</td>
</tr>
<tr>
<td>DOP-PCR</td>
<td>degenerate oligonucleotide-primed PCR</td>
</tr>
<tr>
<td>ds</td>
<td>double-stranded</td>
</tr>
<tr>
<td>dTTP</td>
<td>deoxythymidine triphosphate</td>
</tr>
<tr>
<td>ES</td>
<td>embryonic stem (cell)</td>
</tr>
<tr>
<td>FACs</td>
<td>fluorescence-activated cell sorting</td>
</tr>
<tr>
<td>FRET</td>
<td>fluorescence resonance energy transfer</td>
</tr>
<tr>
<td>Gb</td>
<td>giga bases</td>
</tr>
<tr>
<td>gDNA</td>
<td>genomic DNA</td>
</tr>
<tr>
<td>HDA</td>
<td>helicase-dependent amplification</td>
</tr>
<tr>
<td>kbp</td>
<td>kilo base pairs</td>
</tr>
<tr>
<td>LCM</td>
<td>laser capture microdissection</td>
</tr>
<tr>
<td>LCR</td>
<td>ligase chain reaction</td>
</tr>
<tr>
<td>LR-PCR</td>
<td>long-range PCR</td>
</tr>
<tr>
<td>MDA</td>
<td>multiple displacement amplification</td>
</tr>
<tr>
<td>MIP</td>
<td>molecular inversion probe</td>
</tr>
<tr>
<td>MLPA</td>
<td>multiplex ligation-dependent probe amplification</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>MS</td>
<td>mass spectrometry</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>PDSMS</td>
<td>polydimethylsiloxane</td>
</tr>
<tr>
<td>PEP</td>
<td>primer-extension preamplification</td>
</tr>
<tr>
<td>PPi</td>
<td>inorganic phosphate</td>
</tr>
<tr>
<td>RCA</td>
<td>rolling circle amplification</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>real-time PCR</td>
</tr>
<tr>
<td>SBH</td>
<td>sequencing-by-hybridization</td>
</tr>
<tr>
<td>SBT</td>
<td>sequencing-by-identification</td>
</tr>
<tr>
<td>SBS</td>
<td>sequencing-by-synthesis</td>
</tr>
<tr>
<td>SMRT</td>
<td>single-molecule real-time</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>SOLiD</td>
<td>sequencing by oligonucleotide ligation and detection</td>
</tr>
<tr>
<td>w</td>
<td>single-stranded</td>
</tr>
<tr>
<td>WGA</td>
<td>whole-genome amplification</td>
</tr>
<tr>
<td>ZMW</td>
<td>zero-mode waveguide</td>
</tr>
</tbody>
</table>
Acknowledgements

Jag trodde aldrig att den här dagen skulle komma, men nu är den klar, boken. Det här hade aldrig gått utan alla som hjälpt, inspirerat, diskuterat och distraherat under åren som gått.


Tack till min goa exjobbar Peter som är grym trots att du hatar att labba; och min jämförade projektarbeteare Nemo, tur att jag hittade dig. Thanks to my collaborators at CMB Auroreil, Jeff, Marta, Pedro. It’s been great working with you guys! Jag vill också passa på att tacka alla PIs på avdelningen för att vi har så välflangerade och trevliga labb både på albanova och skälla. Tack till Lasse för trädsnack.


Tack till Daniel, Otto, Sara och Johan för att ni tog er tid att läsa ett kapitel eller två. Minst en stjärna i min himmel får ni. Tack Anna, Emma, Nina, Tove & Torun för reder, spenning, vin och smågodis i Kittelfjäll.

Acknowledgements

Jag trodde aldrig att den här dagen skulle komma, men nu är den klar, boken. Det här hade aldrig gått utan alla som hjälpt, inspirerat, diskuterat och distraherat under åren som gått.


Tack till min goa exjobbar Peter som är grym trots att du hatar att labba; och min jämförade projektarbeteare Nemo, tur att jag hittade dig. Thanks to my collaborators at CMB Auroreil, Jeff, Marta, Pedro. It’s been great working with you guys! Jag vill också passa på att tacka alla PIs på avdelningen för att vi har så välflangerade och trevliga labb både på albanova och skälla. Tack till Lasse för trädsnack.


Tack till Daniel, Otto, Sara och Johan för att ni tog er tid att läsa ett kapitel eller två. Minst en stjärna i min himmel får ni. Tack Anna, Emma, Nina, Tove & Torun för reder, spenning, vin och smågodis i Kittelfjäll.
Sen vill jag vill absolut tacka KTH-tjejerna för mycket kul Naama, Yo & Anna det var ju där det började, och andra goda och spridda vänner Elin & Jo, Lobi & Vic, Marica, Reb & Fred, Sunke & Rick, Lena, Cat, Ulrica, Martina, Josefine för att ni bryggar livet utanför laborat.


Jag vill avsluta med att tacka Johan för all hjälp under den här perioden! Tack för att du finns, du är fantastisk. (Nästa år tar jag båten.)
References

53


Cambridge and New York: Cambridge University Press.


Barnes, W. M. PCR amplification of up to 35 an of TCR repertoires selected by antigen in vivo: a high magnitude CD8 response is comprised of very few hematopoietic progenitors by digital RT-PCR.


Heyries, K. throughputs sequencing.


From micrograms to picograms: quantitative PCR reduces the material demands of high dimensional ultrasonic cage for characterization of globin genomic sequences and restriction site analysis of diagnosis of sickle cell anemia. Science 239, 414, (1988).


Academy of Sciences of the United States of America


Zhao, H. Banér, J. Larsson, C.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


Grothues, D., Cantor, C. R. & Smith, C. L. PCR amplification of megabase DNA with tagged random amplification.


