Methods to Prepare DNA for Efficient Massive Sequencing

Sverker Lundin

Royal Institute of Technology
School of Biotechnology

Stockholm 2012

Royal Institute of Technology (KTH)
School of Biotechnology
Division of Gene Technology

Science for Life Laboratory
SE-171 21 Solna
Sweden

Printed by E-PRINT
Oxtorgsgatan 9-11
SE-111 57 Stockholm
Sweden

ISBN 978-91-7501-548-4
TRITA-BIO Report 2012:22
ISSN 1654-2312

Sverker Lundin (2012): **Methods to Prepare DNA for Efficient Massive Sequencing**, Division of Gene Technology, School of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden

ABSTRACT

Massive sequencing has transformed the field of genome biology due to the continuous introduction and evolution of new methods. In recent years, the technologies available to read through genomes have undergone an unprecedented rate of development in terms of cost-reduction. Generating sequence data has essentially ceased to be a bottleneck for analyzing genomes instead to be replaced by limitations in sample preparation and data analysis. In this work, new strategies are presented to increase both the throughput of library generation prior to sequencing, and the informational content of libraries to aid post-sequencing data processing. The protocols developed aim to enable new possibilities for genome research concerning project scale and sequence complexity.

The first two papers that underpin this thesis deal with scaling library production by means of automation. Automated library preparation is first described for the 454 sequencing system based on a generic solid-phase polyethylene-glycol precipitation protocol for automated DNA handling. This was one of the first descriptions of automated sample handling for producing next generation sequencing libraries, and substantially improved sample throughput. Building on these results, the use of a double precipitation strategy to replace the manual agarose gel excision step for Illumina sequencing is presented. This protocol considerably improved the scalability of library construction for Illumina sequencing. The third and fourth papers present advanced strategies for library tagging in order to multiplex the information available in each library. First, a dual tagging strategy for massive sequencing is described in which two sets of tags are added to a library to trace back the origins of up to 4992 amplicons using 122 tags. The tagging strategy takes advantage of the previously automated pipeline and was used for the simultaneous sequencing of 3700 amplicons. Following that, an enzymatic protocol was developed to degrade long range PCR-amplicons and forming triple-tagged libraries containing information of sample origin, clonal origin and local positioning for the short-read sequences. Through tagging, this protocol makes it possible to analyze a longer continuous sequence region than would be possible based on the read length of the sequencing system alone. The fifth study investigates commonly used enzymes for constructing libraries for massive sequencing. We analyze restriction enzymes capable of digesting unknown sequences located some distance from their recognition sequence. Some of these enzymes have previously been extensively used for massive nucleic acid analysis. In this first high throughput study of such enzymes, we investigated their restriction specificity in terms of the distance from the recognition site and their sequence dependence. The phenomenon of slippage is characterized and shown to vary significantly between enzymes. The results obtained should favor future protocol development and enzymatic understanding.

Through these papers, this work aspire to aid the development of methods for massive sequencing in terms of scale, quality and knowledge; thereby contributing to the general applicability of the new paradigm of sequencing instruments.

Keywords: DNA, Massive sequencing, Next Generation Sequencing, Library Preparation, Barcoding, Multiplexing

Ὁ βίος βραχύς, ἡ δὲ τέχνη μακρή

– Hippocrates

List of publications

This thesis is based on the following five papers, which are referred to in the text by their roman numerals. The papers are included at the end of the thesis.

I

Lundin, S., Stranneheim, H., Pettersson, E., Klevebring, D. and Lundeberg, J. (2010) Increased throughput by parallelization of library preparation for massive sequencing. *PLoS One*, **5**, e10029.

II

Borgstrom, E., *Lundin, S.* and Lundeberg, J. Large scale library generation for high throughput sequencing. *PLoS One*, **6**, e19119.

III

Neiman, M., *Lundin, S.*, Savolainen, P. and Ahmadian, A. Decoding a substantial set of samples in parallel by massive sequencing. *PLoS One*, **6**, e17785.

IV

Lundin, S., Gruselius J., Nystedt B., Lexow. P., Käller, M. and Lundeberg, J. Hierarchical molecular tagging to resolve long continuous sequences by massively parallel sequencing*, submitted*

V

Lundin, S., Hegge, FT., Foam, N., Pettersson, E., Käller. M., Wirta, V., Lexow. P. and Lundeberg. J. Endonuclease specificity and sequence dependence of Type IIS restriction enzymes*, submitted*

Outline

Introduction 1

PREFACE **1**
Scope 2

Chapter 1

THE CODE OF LIFE **5**
Omes and Omics 5
The master molecule 5
Photo 51 6
The RNA tie club 8

Chapter 2

PROTEIN TOOLS FOR DNA SCIENCE **11**
Ligases 11
Restriction Enzymes 12
Polymerases 13
Exonucleases 14
Reverse Transcriptase 15
Other enzymes active on DNA 15

Chapter 3

EARLY DNA SEQUENCING **19**
Electrophoresis 19
DNA Sequencing 19
The Human Genome Project 21

Chapter 4

MASSIVE SEQUENCING **25**
Classifying Sequencing Technologies 25
Early Technologies 26
Established Technologies 26
Newer Technologies 29
Emerging technologies 32
Present limitations 33

Chapter 5

PREPARING DNA FOR MASSIVE SEQUENCING **37**
On libraries 38
Amplification 38
Starting material 39
Fragmenting DNA 40
Modification 41
Targeted enrichment 43
Selection and Purification 44
Quantification 44
Long Insert Paired-End Libraries 45
Specialized libraries 47

Investigation 48

AUTOMATION (I-II) **48**
Papers I and II 48

LIBRARY DEVELOPMENT (III-V) **49**
Paper III 49
Paper IV 49
Paper V 51

Outlook 52

Acknowledgements 53

References 55

Introduction

PREFACE

Aristotle – the first western biologist – made some remarkable discoveries. Perhaps most notably, he described the cuttlefish's unusual method of reproduction (1). This process, now known as hectocotylization, is a sophisticated process where the cephalopod has developed a specialized arm for sperm transfer. In some species the arm is even wrenched off and transferred to the female, and a new arm is grown for the following season (2). Aristotle's description of hectocotylization was widely disbelieved and was not fully verified by modern biologists until 1959 (3).

Aristotle displayed amazing scientific tenacity and also some of the earliest signs of empricism – the systematic use of observations to obtain knowledge. However, he largely lacked proper tools to make quantitative measurements. For instance, he concluded that men have higher body temperatures than women, based on the reasoning that heat is linked to strength and that females in general are weak (4). Measurements of temperature was not possible until the 16th century, when Galileo Galilei constructed one of the world's first thermometers. Galileo was one of the leading figures during the scientific revolution, and made many significant technological contributions. In ancient Greece the divide between *ἐπιστήμη* (*episteme* – knowledge of facts) and *τέχνη* (*techne* – craftmanship) ran deep. Although Aristotle was interested in observations, he was not an inventor. Galileo's inventions enabled him to make observations no one had ever made before. Moreover, he made it absolutely clear that technology is a generator of knowledge and as an integral part of science as episteme. Today, the thermometer is as crucial as it is elementary in any scientific laboratory, and the lack thereof caused Aristotle to make grave errors in his theories. Although not a biologist per se, Galileo was also involved in the development of telescopes and microscopes, which soon enabled other researchers to make the first observations of cells (5). Ever since this period, the development of new methods and technologies has continued to enable new ventures for scientists, thereby driving discovery. Today massively parallel sequencing machines are transforming the study of DNA, cells, health and life. This thesis is dedicated to these technological achievements.

Scope

The experimental work presented in this thesis has been focused on developing methods to prepare DNA samples for analysis using high-throughput sequencing instruments, which consequently marks the main theme of the thesis. Its scope is limited to techniques that specifically and directly target and process DNA for sequencing, and the historical background is written with this in mind. Although a great deal of the literature on sample preparations for massive sequencing concerns methods for isolating and manipulating RNA, these are not emphasized here. Sample extraction techniques such as methods for lysing cells and purifying DNA from lysates are also beyond the scope of this thesis. Finally, prior to the advent of the massive sequencing technologies in use today, a lot of work was done involving the use of various hybridization-based approaches for detection. These have been used for purposes such as genotyping and expression profiling. Despite their widespread use, however, they have no direct bearing on the parallel development of sequencing and are therefore not examined. However, hybridization-based methods are discussed as a means of preparing DNA for sequencing.

Life is not a thing or a fluid any more than heat is. What we observe are some unusual sets of objects separated from the rest of the world by certain peculiar properties such as growth, reproduction, and special ways of handling energy. These objects we elect to call 'living things'.

– Robert Morison

Chapter 1

THE CODE OF LIFE

Life is managed by DNA, RNA and proteins. These groups of molecules constitute a self-replicating machinery enclosed by a lipid bi-layer membrane, together forming a cell that allows reaction conditions to be locally controlled. Information (typically) flows from DNA to RNA to proteins, making DNA the most basic blueprint of life. DNA itself is built up from four basic subunits: adenine, guanine, thymine, and cytosine. In genes, triplets of these bases code for one of 20 amino acids to be added into a growing polypeptide chain via RNA template intermediates to form a protein. The finished proteins promote and perform many different reactions inside or near the micro-compartment depending on their amino acid sequences, and the functioning of cells are determined by the active set of proteins and RNA molecules.

Omes and Omics

In 1986, during an international meeting in Betheseda, USA, geneticist Thomas Roderick and some colleges sat down at a bar to discuss starting a new journal. One of the topics concerned the name of the journal, and Roderick proposed to call it *Genomics* (6). The journal was started the following year focusing on the new scientific endeavor to study entire genomes (in contrast of genetics, which focuses on the study of specific genes).

The word gene comes from the Greek word *γένος* meaning family or origin and was first coined by Johannsen in the early 1900s (7). The basics of heredity were first described by Mendel in 1865 (8), and the word chromosome dates back to 1889 (9). Today, *chromosome* is used to describe a large consecutive piece of DNA containing many genes and surrounding nucleotides. The word genome is a combination of gene and chromosome stemming from a textbook written by Hans Winkler in 1920 (6), and refers to an organism's entire cellular genetic material, which can consist of multiple chromosomes. Thus, in order of size and complexity: Gene < Chromosome ≤ Genome.

Genomics is the study of genomes and is the most mature field of the -*omics* era. The omics-sciences can be described as the scientific attempts to study all the constituents of a complicated mixture collectively, and this development has been enabled very recently by groundbreaking technical achievements. For example the term *proteome* was coined by analogy to "genome" to encompass all the proteins, and *proteomics* was used in 1995 (6) to describe the study of all proteins in parallel, largely enabled by advancements in mass spectrometry. Other *omics* have followed in the footsteps of genomics, such as transcriptomics (RNA), metabolomics (metabolites) and epigenomics (genomic modifications that do not involve changes in the DNA sequence).

The master molecule

The 20th century has been called the century of the gene (10), due to a series of advances that began with the rediscovery of Mendelian inheritance by Tschermak, de Vries and Correns, continued with establishing the structure of DNA

(11), and culminated in the sequencing of the human genome (12,13). In the first half of the 20th century, the chemical nature of the gene had not been definitively settled. Even though the concept of a gene (not using that term) had been described by Gregor Mendel in the 19th century, along with many hereditary mechanisms, and following that, the chromosomes had been intensively studied – most scientists where convinced that proteins held the hereditary information. Solving the puzzle of this master mechanism was to become an intense competition involving some of the greatest scientists in the 20th century.

The discovery of DNA is attributed Friedrich Miescher when he in 1869 isolated the genomic content of cells and concluded that it was an organic acid with a rich phosphorus content (10). Since it came from the nucleus of the cell, Miescher termed it nuclein. Richard Altmann was able to isolate protein-free nuclein, which he named nucleic acid in 1889 (10). In 1910, Albrecht Kossel was awarded the Nobel Prize for his work on nucleic acids (14). By this point, it was clear that nucleic acids contained two purines (A - adenine and G - guanine) and two pyrimidines (T - thymine and C - cytosine). Phoebus Levene contributed some significant discoveries concerning the early studies of DNA composition, and postulated that A, G, C and T occurred at equal proportions, called the tetranucleotide hypothesis (15). At the time, scientists did not realize that the seemingly small size of the DNA molecules they were isolating was due to the extraction conditions used, causing DNA fragmentation. Because of this the size of DNA was greatly underestimated. Levene thought the molecular weight of DNA was 1.5 kDa (10), whereas we know now that the weight of human chromosome 1 has a molecular mass of approximately 150 TDa, making it 100 million times larger than was originally thought. Based on the limited number of components and the conceived small size of nucleic acid, proteins were thought to be the genetic material. This dispute would not be completely settled until the structure of DNA was solved in 1953, but some strong arguments in favor of DNA had been published before that point. In 1928, Frederick Griffith noted that if non-virulent pneumococci was injected together with virulent but heat killed cells, many of the animals became infected with living virulent cells (16). Although the implications of this result were not fully understood at the time, this was a clear indication towards DNA. Oswald Avery continued this research and showed in 1944 that the transformation from non-virulent to virulent could be achieved by injecting purified DNA purified from the virulent strains (17). In 1950, Chargaff published what was to become Chargaff's rule, which stated that the amounts of A and T in the genome are equal, as are those of G and C (18). All of these findings argued in favor of DNA.

Photo 51

Solving the structures of biomolecules was a hot topic in the 1950s. Two of the rivals involved in this competition were Linus Pauling and William Lawrence Bragg – both primarily focused on proteins. Pauling was one of the most influential scientists of the 20th century and is the only person two receive two unshared Nobel Prizes (Chemistry in 1954, and the Peace Prize in 1962) (19,20). Pauling had solved the structure of silicates in the late 20s (21) as well as the protein alpha-helix in 1951 (22). Bragg led the Cavendish lab at Cambridge where James Watson and Francis Crick worked, and had together with his father invented X-ray crystallography – a tool to study the arrangements of atoms in crystals – in the early 1900s. *For their services in the analysis of crystal structure by means of X-rays* they received the Nobel Prize in physics in 1915 (23) (23), and Lawrence, only 25 years old, is still the youngest person ever to receive the prize.. Lawrence, at the time only 25 years old, remains the youngest person to have ever received the prize. In 1953, Pauling proposed that the structure of DNA consisted of a three-stranded helix with the bases pointing outwards (24). However, another paper published in the same year would contest that view. In fact, the structure of DNA can be traced back to a single picture acquired in another lab by Rosalind Franklin at King's College in Cambridge. Franklin had after many years of X-ray crystallography training in Paris, arrived in Cambridge to work together with Maurice Wilkins on the structure of DNA. Franklin and Wilkins did not get along, and so work for Franklin proceeded much in isolation. When Wilkins one day, beyond Franklin's knowledge, brought her results to the attention of James Watson, Watson was able to conclude that DNA consisted of a double helix with the phosphates on the outside, rather than a triple helix. Watson and Crick had previously hypothesized a much similar structure of DNA as the one Pauling published in 1953, which Franklin had strongly argued against (25). The now-famous photo 51 that Wilkins brought to Watson (Figure 1), significantly aided Watson and Crick in publishing the structure of DNA in 1953.

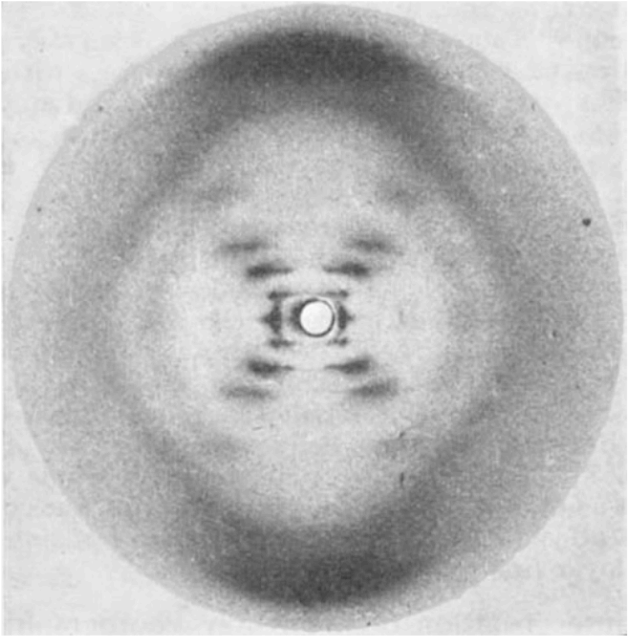

Figure 1. Photo 51. Reprinted by permission from Macmillan Publishers Ltd: Nature 171: 740-741, copyright 1953.

Franklin died young (in 1958) from ovarian cancer, likely related to the hundreds of hours she spent exposed to X-rays while working to decipher the structure of DNA, and the Nobel Prize was awarded to Watson, Crick and Wilkins in 1962 (26). The paper by Watson and Crick is now regarded as marking the birth of quantitative biology and is considered to have settled the debate on the holder of genetic material, as well as providing the first understanding of how DNA could be replicated. Because the basis of life is replication, and the basis of replication is base pairing, the principal structure of DNA had rocketed biology into a new era. The power of a single photograph, enabled by X-ray crystallography that had emerged in the 1910s, truly transformed an entire field of science at a stroke.

The RNA tie club

After Watson and Crick's paper in 1953, much interest turned to working out how only four DNA bases could encode the rich diversity of proteins. Walter Gamow, a Russian physicist entered the race, and started an RNA tie club in 1954. He invited 20 prominent researchers to join as members (one for each amino acid) and also 4 honorary members (one for each nucleotide) to work on solving the riddle of RNA (it was known that protein synthesis occurred at ribosomes and that ribosomes were RNA). Each member received a black wool tie embroidered with a green and yellow helix, and each member was assigned a specific amino acid. Members included Watson (proline), Crick (tyrosine), Feynman (glycine) and Brenner (valine). Eight of the members already were, or were to become Nobel laureates. In 1958, Crick for the first time explicitly stated an idea that he termed the central dogma: *this states that once 'information' passed into protein it cannot get out again*, and he also wrote that he should *be surprised if the main features of protein synthesis are not discovered within the next ten years*. He turned out to be right, but it was not any member of the RNA tie club that solved it. Despite Gamow's efforts, the person to actually crack the code was Marshal Nirenberg, a scientist working at the NIH. Gamow had proposed that three bases coded for an amino acid (two bases would only be sufficient for 16 amino acids, and 20 were known). In 1961, Nirenberg showed that when polyuridylic acid was added to mashed *E. coli* cells the peptide that was produced was polyphenylalanine (27). So it would seem, the code for phenylalanine was UUU. Crick further expanded on this in a paper arguing that amino acids in fact are encoded in triplets and provided the first evidence for phase-shifts (28). Nirenberg and Leder resolved more triplet codes in 1964 (29) and in 1968, Nirenberg, Holley and Khorana received the Nobel Prize for their interpretation of the genetic code (30). Holley was awarded his share of the prize for the discovery of tRNA (31) and Khorana for the invention of methods for synthesizing nucleic acids (32).

Being a molecular biologist before the discovery of DNA as the genetic holder is hard to imagine. Seven years before Watson and Crick's seminal paper, Erwin Schrödinger (the world-famous quantum physicist) had published a book titled "What is Life" (33). He argued that life is not a separate quality among the living, but can be broken down into separate physical quantities. However, without the structure of DNA this was not possible to do. Heritability was a known phenomenon but could not be understood. Franklin's, Wilkins', Watson's and Crick's discovery of DNA made this leap when the concept of base-pairing was elucidated. What then emerged and has continued to develop since, is a quantitative way to study biology. Nowadays, the code of life can be measured, compared, and used for predictions in massive scale. However, many more technological breakthroughs were needed to make this possible.

The active agent in the expressed yeast juice appears rather to be a chemical substance, an enzyme, which I have called "zymase". From now on one can experiment with this just as with other chemicals.

– Eduard Buchner

Chapter 2

PROTEIN TOOLS FOR DNA SCIENCE

Molecular biology is the area of science that investigates the molecular interactions necessary for cellular life. Primarily, the area of focus is DNA, RNA and proteins; but also lipids, carbohydrates and metabolites are being studied. Most of the tools available for studying these macromolecules are themselves derived from living cells. The tools are isolated from their natural occurrence, characterized (and in some cases altered), and then produced and distributed for use in vitro. Most such tools are enzymes. Enzymes are a group of proteins that catalyze chemical reactions within the cell. They have evolved over millions of years to become an integral part of cellular life. The word enzyme was introduced by Wilhelm Kühne from the Greek word *εὔζυμος*, meaning *to leaven*, in 1877 (34), and Eduard Buchner later received the Nobel Prize in Chemistry in 1907 for his discovery of cell-free fermentation using an enzyme he called *zymase* (35). From a technical standpoint, enzymes are extremely useful substances in vitro as they efficiently catalyze reactions involving biomolecules, enabling the formation of new products useful for analysis.

The following sections describe the major contributions of enzymes for the molecular biology toolbox, each marking a substantial technological advancement in the field. The section also includes a brief compilation of different enzymes that are commercially available. It should be stressed that this compilation is intended to provide a concise but not exhaustive list of to the choices that are available for manipulating DNA in vitro.

Ligases

In 1967, four research groups almost simultaneously indentified an enzyme capable of covalently joining DNA strands. It had previously been known that cells contained the ability to repair DNA damage and many researchers were seeking for an enzyme that made this possible. Early in 1967, Martin Gellert showed that the ends of the bacteriophage λ genome could be covalently bound together to form DNA circles using crude extracts of *E. coli* cells (36). In the same year, the enzyme responsible was isolated by four independent research groups (37-40). The isolation of the *ligase* enabled researchers to form recombinant DNA, i.e. to synthetically combine DNA sequences at will. The combination of this technology with restriction enzymes would represent an immensely important step for method development. Today, ligases have been isolated from many different sources. The T4 DNA ligase is perhaps the most widely used, but many different ligases are commercially available (see Table 1). In general, ligases require either ATP or NAD+ for ligation to occur (ATP-dependent ligases have viral or archaeal origin, while NAD-dependent ligases originate in bacteria (41)). In addition, the 5'-hydroxyl terminus of the oligonucleotide sequence to be ligated must be phosphorylated, as this terminal phosphate will form the phosphate bond to the 3'-hydroxyl terminus being ligated.

Ligase	Reaction Tempertature	Heat inactivation	Substrates	Dependency	Supplier	Comment
T4 DNA Ligase	16-37	65/10	B,N,DR,RR	ATP	NEB	Most common choice
T7 DNA Ligase	25	65/10	N	ATP	NEB	Used if blunt end ligation is to be avoided
T3 DNA Ligase	25	65/10	B,N	ATP	NEB	High tolerance for NaCl
T4 RNA Ligase 1	37	65/15	ss, RR,DR	ATP	NEB	ssRNA ligase
T4 RNA Ligase 2	37	No	N, DR,RR	ATP	NEB	dsRNA ligase
E. coli DNA Ligase	16	65/20	N	NAD+	NEB	Used if blunt end and RNA-DNA ligation is to be avoided
Taq DNA Ligase	45-65	No	N	NAD+	NEB	Thermostable ligase
Circ-Ligase	60	80/10	ss,DR	ATP	Epicentre	ssDNA ligase, used to circularize ssDNA oligos
Ampligase	28-85	No	N	NAD+	Epicentre	Thermostable ligase. May ligate sticky ends at low temperature

B = *blunt*
N = *single stranded nicks*
ss = *single stranded*
RR = *will ligate* RNA *to* RNA
DR = *will ligate* DNA *to* RNA

Table 1. A collection of different ligases with their respective properties (42,43).

Restriction Enzymes

The Nobel Prize in Physiology or Medicine was awarded to Werner Arber, Daniel Nathans and Hamilton O. Smith in 1978 *for the discovery of restriction enzymes and their application to problems of molecular genetics* (44). Each winner had made a distinctive contribution. Arber isolated the first restriction enzyme from *E. coli* and termed it *restriction* for its ability to restrict infections by certain bacteriophages (45,46). The enzyme isolated by Arber was EcoBI, a Type I restriction enzyme recognizing a specific sequence but cleaving DNA randomly. The first Type II enzyme was isolated by Smith (47) later to be named HindII, and was shown to recognize and cut at GTY|RAC (48). Nathans made the first technological leap to what would lay the foundation of modern biotechnology. In 1971, Danna and Nathans published a paper outlining the basics of restriction fragment mapping (49). This illustrated many uses of restriction enzymes, soon pairing them with plasmids and ligases to enable researchers to freely choose the DNA to be inserted into *E. coli* cells (50). Although not usually appreciated as such, restriction enzymes were one of the cornerstones necessary for DNA sequencing. Both Sanger and Maxam and Gilbert relied on them to develop their respective sequencing technologies (50) (see Chapter 3).

Today, thousands of different restriction enzymes have been described with hundreds of different specificities (51,52). In nature, restriction enzymes are part of a restriction-modification system as a protection against infection. Each restriction enzyme has a specific recognition sequence and is believed to have a corresponding methyltransferase partner that recognize the same sequence. The host DNA is methylated by the methyltransferase, protecting it from cleavage, while invading DNA (such as that originating from bacteriophages) is *restricted* by being cleaved into pieces. These systems are classified into four categories: Type I, Type II, Type III and Type IV (53). Type I restriction enzymes are multi-subunit protein complexes recognizing specific sequences but cleaving DNA randomly. Type II are the most multifaceted and also most utilized type of restriction enzyme. They recognize specific sequences and cleave at or close to the recognition sequence. A notable subgroup of Type II restriction enzymes is the Type IIS (S for shifted) restriction enzymes, which cleave the DNA at a shifted distance from the recognition sequence. Type III are ATP-dependent, consists of two subunits, and require two copies of the recognition sequence for cleavage.[i] Type IV restriction enzymes only recognize methylated targets. A list of representative enzymes from all classes is presented in Table 2.

Enzyme	Type	Recognition and Cleavage	Cleavage	Comment
EcoBI	I	TGANNNNNNNNTGCT	Random	First restriction enzyme
KpnI	II	GGTAC|C	3' 4 nt	
EcoRI	II	G|AATTC	5' 4 nt	
MseI	II	T|TAA	5' 2 nt	Short recognition sequence
AluI	II	AG|CT	blunt	Short recognition sequence
NotI	II	GC|GGCCGC	5' 4 nt	Long recognition sequence
HindII	II	GTY|RAC	blunt	First type II enzyme
DpnI	IIM	G$_{6m}$A|TC	blunt	Recognize only methylated targets
MmeI	IIS	TCCRAC(N)$_{21}$|	3' 2 nt	Shifted cleavage
FokI	IIS	GGATC(N)$_{9}$|	5' 4 nt	Shifted cleavage
EcoP15I	III	CAGCAG(N)$_{25}$|	5' 2 nt	Shifted cleavage
McrBC	IV	R$_{m}$C(N)$_{40\text{-}3000}$R$_{m}$C	NR	Unknown cleavage

3' 4 nt = 4 nucleotides 3' (top strand) overhang
NR = not reported
| = point of cleavage
R = purine
m = methylation

Table 2. A collection of different restriction enzymes and their respective properties (52).

Polymerases

Polymerases are enzymes responsible for polymerizing nucleic acids complementary to an oligonucleotide template. The first polymerase isolated was DNA polymerase I. This landmark discovery was accomplished by Arthur Kornberg in 1956 (54), who was awarded the Nobel Prize in 1959 along with Ochoa for studies on nucleic acid synthesis (55) (Ochoa for work on RNA (56)). Each cell has different polymerases specialized for different functions, e.g. replication or repair (57), reflecting the opportunities for isolation of suitable in vitro catalysts. The discovery of the DNA polymerase came short in time after Watson and Crick published the structure of DNA, and although the potential replication mechanism was outlined merely from the model Watson and Crick had presented, some doubts lingered in the community (58). The discovery of the polymerase was to shed these doubts. Even more important, it allowed researchers to replicate DNA in vitro which was a crucial step towards sequencing and exponential amplification.

[i] Type IIS and Type III enzymes are further characterized in Paper V in this thesis.

In 1984, a polymerase from phage *ϕ29* was described *(59)*. The *ϕ29* polymerase displayed a remarkable processivity and strand displacement activity and could because of this be used to isothermally amplify primed circular DNA into long continuous linear concatemers (60). The ϕ29 polymerase has since become very useful for amplifying circular or minute samples (61). Another groundbreaking polymerase discovery came in 1988, when the first thermostable DNA polymerase was isolated from a thermophile bacteria, *Thermus aquaticus*, giving rise to the still widely used Taq polymerase (62).

The utility of polymerases in molecular biology cannot be overestimated. Without polymerases, very few nucleotides would have been sequenced, and tremendous effort is going on across the world to develop new and improved polymerases for specialized purposes. In principle, the features of DNA polymerases can be divided into four categories: fidelity, thermostability, strand displacing activity and 5' exonuclease activity. *Fidelity* is the polymerase's ability to accurately replicate DNA, and is often related to proof-reading activity. Proofreading means that the polymerase contains a 3'-5' exonuclease that removes incorrectly incorporated nucleotides. However, other modifications can also affect fidelity such as buffer composition, and high-fidelity is not always desirable, such as when performing error prone procedures to assay effects of base alterations (63). *Thermostability* is important for controlling DNA hybridization and denaturation while maintaining polymerase activity. Highly stable enzymes last longer under extreme experimental conditions, but not all experiments benefit from high temperature. The *strand-displacement* activity ensures that dsDNA encountered during extension does not intervene with replication. In some protocols this is beneficial, in others not. Enzymes with *5'-3' exonuclease activity* degrade encountered 5' ends of duplex DNA during replication. This can be a very useful feature to exploit (64). Table 3 shows a set of polymerases with different properties.

Polymerase	Fidelity (errors per million bases)	3'-5' exo (proof-reading)	5'-3' exo	Strand displacement	Thermo-stability	Primary uses	Reference
Taq	285	-	+	+	++	PCR	NEB/Polbase
Pfu	1.3-2.2	+++	-	+	+++	PCR	Polbase
E. Coli DNA Polymerase I	9	++	+	+	-	Nick Translation	NEB/Polbase
Klenow Fragment DNA Polymerase I	18	++	-	++	-	End repair	NEB
Klenow Fragment 3´→ 5´ exo–	100	-	-	+++	-	Labelling	NEB
phi29 DNA Polymerase	3	++++	-	+++++	-	Amplification	NEB/Polbase
Bst DNA Polymerase	NR	-	+	++++	+	Strand displacement	NEB
Bst DNA Polymerase, Large Fragment	NR	-	-	++++	+	Strand displacement	NEB
T4 DNA Polymerase	< 1	++++	-	-	-	End repair	NEB
T7 DNA Polymerase	15	++++	-	-	-	Long templates	NEB
KAPA HiFi	0.28	+++	-	+	+++	PCR	KAPA
Phusion HiFi	< 0.44	+++	-	-	+++	PCR	NEB

NR = *not reported*

Table 3. A collection of some common DNA polymerases and their respective properties (65-67).

Exonucleases

The exonucleases are another useful group of enzymes that recently has been suggested for use in nanopore sequencing (68), and are also commonly used for library preparation. The first exonuclease to be characterized was Exonuclease I from *E. coli* in 1960 (69,70). It was found to degrade ssDNA in the 3' to 5' direction and is still used today to remove ssDNA. Exonuclease III was isolated in 1964 (71) and was shown to degrade dsDNA from the 3' termini.[i] Exonucleases have varying demands for their substrates and can therefore be used to specifically remove certain unwanted nucleotide polymers during experiments, such as ssDNA or non-circular DNA. Some common exonucleases are presented in Table 4.

Enzyme	Direction	Substrate	Activity without 5'-phosphate	Phosphothioate cleavage	Initiate from nick
Lambda Exonuclease	5'->3'	ds, (ss)	(+)	-	-
T7 Exonuclease	5'->3'	ds, (ss)	+	-	+
Exonuclease III	3'->5'	ds, (ss)	+	-	+
RecJf	5'->3'	ss	+	-	-
Exonuclease I	3'->5'	ss	+	-	NR
Exonuclease T	3'->5'	ss	+	-	NR
Exonuclease V	both	ss, (ds)	+	-	-
T5 Exonuclease	5'->3'	ds, ss	+	+	+

NR = not reported
ds = dsDNA
ss = ssDNA
() = greatly reduced rate

Table 4. Specificities of some common exonucleases (72).

Reverse Transcriptase

Very few sequencing techniques allow for the direct sequencing of RNA. The first step in preparing an RNA library for sequencing is therefore usually to convert it to DNA. The first step to achieve this was taken by Baltimore and Temin & Mizutani in 1970 during studies of RNA viruses (73,74). Verma and Kacian demonstrated the use of a reverse transcriptase enzyme in vitro to synthesize cDNA from mRNA (75,76). Although slightly outside the scope of this thesis, reverse transcriptases deserve mentioning since their discovery have been essential for the functional understanding of genes and genomes, and their existence is largely responsible for the large scale transcriptomics of today.

Other enzymes active on DNA

I will also briefly mention some of the enzymes that have been adapted for massive sequencing preparation protocols and are mentioned in Chapter 5. Site-specific *recombinases* are enzymes recognizing certain motifs in two duplex DNA strands and then break, shuffle and join them to one another, thereby recombining them. Cre-lox recombinase (77) is presently used by the 454 paired-end protocol (see Chapter 5). Recombinases (sometimes also referred to as integrases, resolvases and invertases) have the potential of increasing efficiency and specificity of circularization protocols. *Transposases* are enzymes re-

[i] Exonuclease III is utilized in Paper IV in this thesis to prepare DNA for massive sequencing.

sponsible for inserting transposable elements into the genome (transposons). A version of the Tn5 transposase (78,79) has recently been adapted for massive sequencing (80)(see Chapter 5). Transposases have the benefits of both fragmenting and ligating adapters at the same time, increasing throughput and efficiency. However, they also have recognition preferences that could potentially introduce sequence bias. Other highly utilized enzymes include polynucleotide kinase (for phosphorylating 5' ends), alkaline dephosphatase (for dephosphorylating 5' ends), and nucleases (such as S1, mung bean, DSN, and DNase I for cleaving DNA).

Together, the isolated and characterized DNA enzymes constitute an ever-growing toolbox for the controlled manipulation of genes and genomes to facilitate analysis. As further examined in chapter 3, one such endeavor was to decode the entire human genome.

To learn to read is to light a fire; every syllable that is spelled out is a spark.

– Victor Hugo

Chapter 3

EARLY DNA SEQUENCING

Electrophoresis

A tremendous technological breakthrough in the beginning of last century, which was to become very important to DNA science, was a new technology to separate a mixture of macromolecules. Arne Tiselius received the Nobel Prize in 1948 (81) for his contributions to protein science using an instrument he had invented and termed the *electrophoresis apparatus* (82). This technology led to insights in RNA hydrolysis by Markham and Smith in 1952 (83) and was further developed in 1955 by Smithies, who introduced starch gels as medium for protein separation (84). The use of polyacrylamide gels for DNA analysis was first described by Danna and Nathans (49), agarose gels were introduced in 1972 (85,86), and pulsed field electrophoresis enabled separation of entire chromosomes in the 80s (87). A PhD graduate of Tiselius, Stellan Hjerten, has been called the *father of capillary electrophoresis* for his early adaptation of this technique in 1967 (88). Jorgensen and Lucacs described a micron sized capillary technique in 1981 (89) which later was adapted for separation of DNA (90), and Swerdlow and Gesteland demonstrated the first capillary DNA sequencing in 1990 (91,92).

It is as hard to imagine a molecular biology laboratory not using the electrophoresis principle as it must have been for the early pioneers to imagine that this technique half a century later would be used to read through the entire human genome. The basis of the technique when applied to DNA is to take advantage of the inherent negative charge of the DNA phosphate backbone, making it migrate in a medium when an electrical current is applied. Higher voltage, leads to faster migration. When a gel matrix is used, the speed is determined by the size of the DNA molecule: shorter molecules will migrate faster through the pores of the gel. This process, known as sieving, facilitate the size separation of complex mixtures. When coupled with a reagent specifically attached to the DNA, such as radiolabeled nucleotides or a DNA binding fluorophore, the length of the DNA fragments in an unknown sample can be visualized and determined. This technique is still one of the most fundamental assays of any molecular biology lab, and can also be used preparatively to extract molecules with specific lengths[i] – as well as for sequencing.

DNA Sequencing

Frederick Sanger started as a PhD student under Albert Neuberger in Cambridge in 1940. He did his thesis on the metabolism of lysine but soon ended up working on the amino acid composition of insulin. He spent some time as a guest researcher in Tiselius' lab in Uppsala in 1945, where he became acquainted with some early separation techniques using starch. He was a pioneer in sequencing (of proteins as well as DNA) and established the complete se-

[i] A new way of size selecting DNA molecules is investigated in Paper II

quence of insulin in 1955 (93). This was the first determined sequence of any protein and earned him the Nobel Prize in 1958 (94).

Sanger and coworkers also sequenced the first nucleic acid ever, alanine tRNA, in 1969 (31). The methods of RNA sequencing were much like protein sequencing by partial hydrolysis, separation and labeling. Before the introduction of restriction enzymes, DNA molecules were so long that preparing DNA for sequencing was the main difficulty – the read-out, once there, could potentially be much easier as compared to proteins since there are only four bases to label compared to 20 amino acids (94). The first DNA sequence to be sequenced was the 12 bp cohesive ends of the bacteriophage lambda genome, which were sequenced partially by Wu and Kaiser in 1968, and completed by Wu and Taylor in 1971 using *E. coli* polymerase I and radiolabeled nucleotides (95,96). However, it was the work of Frederick Sanger that would bring greater throughput to DNA sequencing.

Plus and minus

The technology that enabled throughput was electrophoresis. With the advent of polyacrylamide gel electrophoresis, sequences of up to hundreds of bases could be separated with single base resolution. The first take on this approach was Sanger's so called plus and minus method (97). In the first paper, a primary reaction using a synthetic 10 bp primer, restriction fragments of the phiX genome, and DNA polymerase I was used to create as many different lengths of radiolabeled products as possible. The reaction mixture was then split in 2 x 4 reactions, four reactions in which one of each respective base was omitted (the minus) and four in which only one of each respective base was added (the plus). The minus would show which base was to come, and the plus would show which base had last been incorporated. The length of the fragments was then determined by gel electrophoresis, and each position could be read. For homopolymers, lengths were estimated from band spacing which became increasingly difficult for increasing lengths (98).

Maxam-Gilbert

The Maxam-Gilbert sequencing method is still the only purely chemical sequencing method that has been widely used (although restriction enzymes were used to create suitable targets, this was not a theoretical limitation). The method is based on a series of chemical treatments, partially removing specific nucleotides of an end-labeled DNA fragment (99). Subsequent polyacrylamide gel electrophoresis would then reveal which base had been at the end of that fragment length. Since the method did not suffer from the same difficulties with homopolymers as the plus and minus method, it was initially preferred (98).

Dideoxy method

In 1977 Sanger published what became known as the *dideoxy method* (100). This method mitigated the problems the plus-minus method had in getting equal signal from different bands over a long fragment. The new method was based on using nucleotide analogs, such as dideoxynucleotides (ddNTPs), that the polymerase would incorporate, but could not extend from (chain-terminators). By supplementing the reaction with one chain-terminator at a time, both the length of the fragment and the identity of the last incorporated nucleotide could be determined in a single step. Moreover, all lengths of a homopolymer where represented. In this first publication Sanger and colleges reported that up to 200 bases could be sequenced with reasonable accuracy. It was this method that would be used to sequence the human genome.

The Human Genome Project

One must realize the magnitude of absurdity to discuss sequencing the human genome in 1985. The largest genome sequenced was the bacteriophage lambda genome with 48,000 bp (101) and it would be 10 more years until the first bacterium was sequenced (102). The human genome is nearly 5 orders of magnitude larger than that of lambda genome and yet, in that year, a workshop was held at the University of Santa Cruz on the topic: *Can we sequence the human genome?* (103). The human genome project (HGP) was officially initiated in 1990 and declared complete in 2003. The project ended up as a competition between a private initiative run by Celera Genomics and headed by Craig Venter, and a public initiative headed by Eric Lander. The contribution to science and mankind was the remarkable accomplishment of ordering the three billion letter long string of A, G, C and Ts to reflect the chemical composition of a representative human genome (the genomes of several individuals were actually used as input material). The technological achievements were equally remarkable. In 1990, the weekly throughput of sequencing instruments was 25,000 bases. In 2000 it was 5 million, and in 2011 it was over 250 billion (104). Three of the most significant technical revolutions all originated in the 80s: the polymerase chain reaction, automated and capillary sequencing, and shotgun library preparations. These technologies emerged before the project, and enabled its beginning and completion.

PCR

Khorana, who developed techniques for synthesizing nucleic acids, described the initial steps necessary for replicating DNA in vitro in 1971 (105), but it was Mullis et al. who demonstrated all the necessary steps experimentally in 1986 (106) and later received the Nobel Prize in 1993 (107). When coupled to the discovery of the Taq polymerase in 1988 (62) this became an extremely potent amplification technique that completely transformed DNA science, making it possible to amplify and analyze DNA from minute samples containing only traces of material.

Automated Capillary

The next big leap for the Sanger dideoxy method came in 1985 with a publication describing the labeling of primers with fluorophores (*dye primers*) rather than using radiolabeled nucleotides (108). This led to the first paper on automated DNA sequencing in 1986 (109). Both of these papers came from Leroy Hood's lab at Caltech in collaboration with ABI. By using four different fluorophores, each of the 4 base terminating reactions could be combined before gel separation, allowing for the simultaneous read-out of all bases. Shortly thereafter, the ABI 370A DNA sequencer was released (98). The next improvement was to attach the fluorophore to the ddNTPs (called a *dye terminator*) rather than the primer, which allowed all four reactions to proceed simultaneously. Automated capillary electrophoresis progressed during the 90s leading up to the launch of the first automated instrument in 1996 – the ABI 310.

Shotgun

Shotgun sequencing emerged in the early 80s with the sequencing of mitochondrial and phage genomes (101,110). It was a much quicker way of decoding nucleotide sequences than direct sequencing but required more work to assemble the pieces together. A great breakthrough was made when Venter et al. demonstrated the use of this approach in bacterial sequencing (102), and this was the strategy applied for the human genome by Celera Genomics.

Cloning

Molecular cloning played a key role in the preparation of DNA samples during the human genome project, which is noteworthy given that it is *not* a key factor of the following chapters regarding methods of massive sequencing. Cloning was used to construct libraries prior to the advent of massive sequencing and it began with the transformation of *E. coli* to antibiotic resistance, shown by Cohen et al. in 1972 (111). Cohen further described the use of plasmids produced in vitro in 1973, marking the beginning of an era of cloning (112). Plasmids can harbor up to approximately 10 kb of inserted material. Cosmids/fosmids are based on bacteriophage lambda (cosmids, 1978 (113)) and the bacterial F-plasmid (fosmids, 1992 (114)), respectively, are usually inserted into *E. coli*, and can harbor material in the size range of 40 kb. Yeast artificial chromosomes (YACs, 1987 (115,116)) can carry 100 to 2,000 kb, and bacterial artificial chromosomes (BACs, 1992 (117,118)) around 100-300 kb (118). A milestone leading up to the human genome project was the production of cosmid libraries in the early 80s (119), and BAC libraries providing BAC-end sequencing in the late 90s (120).

Today, after much improvements and perfection, Sanger based sequencing is still the 'gold standard' and the method of choice for validation of sequences obtained using other methods, since it can achieve read lengths of 1000 bp with a raw accuracy of 99.999% (121) – a performance which to date is still unrivaled. Features that have improved by alternate methods are throughput and cost.

Machines take me by surprise with great frequency.

– Alan Turing

Chapter 4

MASSIVE SEQUENCING

Most people, at least in the public, probably thought of the human genome as an end-point of discovery – but in fact it was a beginning. Even though we have a good understanding of how genes code for proteins, and the human genome allowed us to roughly categorize those genes, the wealth of information contained in nucleic acids turned out to be far more daunting than ever had been anticipated. If the 20th century was the century of the gene, the 21st century will likely be the century of the genome. This chapter covers the technological breakthroughs in sequencing, from HGP to the present.

Classifying Sequencing Technologies

No clear consensus has yet been established regarding how to classify sequencing technologies. Many different factors weigh in to which technologies should be grouped together, and may vary in importance depending on the biological question at hand. Some examples of factors that may be relevant are *performance*, such as throughput, accuracy, cost, speed, read length and paired-end insert size; *technology,* such as detection features (e.g. real-time or single molecule), chemical or physical basis of sequencing (e.g. sequencing by synthesis, ligation or non-enzymatic sequencing), chemical or physical basis of detection (e.g. fluorescence or non-light based detection); or even the *business model* associated with the technology – i.e. whether it can be implemented using a commercially available instrument or can only be obtained as a service. A common but discouraged method of classification is to divide sequencing technologies into generations. The wide palette of different factors weighing in to define a generation make this terminology ambiguous. To not unduly prioritize any one factor, the following sections are divided into early technologies (i.e. predecessors to more common technologies today), established technologies (i.e. technologies that are widely used today), and newer technologies (i.e. technologies that were developed after the well established technologies). The term massive sequencing is used to denote any technology that substantially increases the throughput of nucleic acid decoding as compared to the sequencing chemistry used to complete the human genome.

In general, massive sequencing technologies provide a miniaturization of reaction volume, which enables parallelization (these techniques are for this reason often referred to as massively parallel sequencing). One of the great achievements of capillary sequencing was the parallelization of the capillaries into a 96 well format, and automate the entire process from gel-injection to fluorescence readout. Since then, further miniaturization of the reaction vessels and automation of the reaction chemistries have permitted additional increases in throughput enabling millions and billions of reactions to occur simultaneously.

Early Technologies

Pyrosequencing

Pyrosequencing was first described in 1998 and was the first new sequencing technology to find widespread use in over 20 years. Ronaghi, Uhlén and Nyrén published a method where the pyrophosphate released during DNA synthesis is coupled to a reaction that generates a light signal, which in turn serves as the basis for detection. By cycled additions of one base at a time, the light signal reveals if the added base was incorporated thereby enabling decoding. The process involved three enzymatic reactions. First, sulfurylase converts the pyrophosphate to ATP. Second, luciferase uses the ATP to generate a light signal. Third, apyrase breaks down the remaining ATP and non-incorporated nucleotides (122). This three-component reaction enabled real-time sequencing (since the light is emitted when the nucleotides are incorporated). This technique, where the bases are called as they are synthesized by a polymerase is called sequencing-by-synthesis, and is still today the most widely utilized sequencing technology.

Polonator

In 2005, two *next generation* sequencing methods were described. One, from 454 Life Sciences, was used in the first commercial next generation sequencing instrument, the GS20 (123). The other was the polonator, or polony-sequencing (124), which would not be as widely spread but lead to the development of other commercial platforms such as SOLiD and Complete Genomics. The "next" generation aspect of both of these technologies was their miniaturized and highly parallelized reaction volumes. The polonator technology began in 1999 with a publication of Mitra and Church demonstrating the formation of PCR colonies in a polyacrylamide matrix, termed polonies (125). In 2003, these authors further demonstrated that sequencing (8 bp read lengths) could be performed within these colonies using reversibly dye-labeled nucleotides (126). In 2005, Shendure et al demonstrated a new sequencing concept in the polonies, namely sequencing-by-ligation (124). An emulsion PCR (emPCR) was used to amplify the material and couple it to beads, and the polyacrylamide gel matrix was used to create a monolayer of the beads used as solid support for sequencing. The templates on the bead surface were decoded using degenerate nonamers with one locked base position coupled to one of four fluorophores, each representing the base that was locked in the nonamer. If the ligase incorporated the nonamer, the fluorophore would reveal which base had been in that nonamer. By cycling through different nonamers with alternating locked positions, they showed that T4 DNA Ligase could discriminate between mismatches up to six bp away in the 5'-3' direction and seven bp in the 3'-5' direction, resulting in 13+13 bp reads with a distance of approximately 1 kb. Each 13 bp read consisted of 6 + 7 bp with internal gap of 3-4 bp stemming from the library. In this way they demonstrated the ability to sequence up to 30 Mb per run and sequenced an *E. coli* genome.

Established Technologies

454

In an early version of the pyrosequencing protocol, the apyrase had not been added to the enzymatic mixture, and so intermediate washing steps were required (127). The addition of apyrase in 1998 obviated the need for washing and solid support for the template. To increase read length however, both washing and apyrase became necessary, and more importantly, by returning to solid support via emPCR (128-130) extensive parallelization was enabled (123). After the library has been coupled to emPCR beads, the beads are loaded on a picotitre plate (131) with millions of small wells capable of holding one emPCR bead. Through the use of a fluidics system, reagents can be delivered and washed away from the wells, which facilitates longer read lengths than the early pyrosequencing protocol could attain. The GS20 machine delivered 20 million bases per run with read lengths of 100 bp. Subsequent upgrades delivered 250 bp, then 400 bp, and current read lengths are approaching 700 bp with the GS FLX+ system delivering 1 million reads and 0.7 Gb of throughput per run (132). Traditionally, the strength of the 454 system has been long read lengths, albeit with relatively high cost per base. Also, 454 sequencing faces inherent problems with homopolymers due to the non-linearity of the light signal intensity associated with multiple incorporations. With the continued development of longer reads for cheaper platforms, the interest in 454 sequencing has begun to dwindle.

Solexa/Illumina

The second, and currently most widespread technology, to enter the market was the Illumina (formerly Solexa) sequencing system (133). Its automated bridge amplification, low cost, paired-end approach, and continued improvements in read length have facilitated the uptake of this technology in the community. The bridge amplification uses two amplification primers grafted onto the surface of a flow cell on which denatured templates (the library) are allowed to hybridize. Extension over the template covalently couples the templates to the flow cell surface, which are then denatured to generate free 3' ends that can hybridize to another primer grafted to the flow cell (thereby forming a bridge) allowed to extend again (134). By cycling hybridization, extension and denaturation in this way, clonal clusters are formed over the surface of the flow cell. These clusters are primed with sequencing primers and sequenced. Illumina uses reversibly dye-terminated nucleotides (135) that terminate synthesis when incorporated, but after imaging can be reinstated as functional 3' ends to enable the next round of synthesis. Paired-end sequencing is achieved by keeping the clusters oriented so that the sequencing reaction proceeds down towards the flow cell surface, after which the cluster is reversed by initiating another bridge cycle followed by the removal of the former sequencing template strand. By using the information contained by the spacing of the ends created during library preparation, deviations in mapping to the genome can be used to discover structural variants.[i] Illumina has consistently scaled their instrument throughput and improved on read lengths. Today, the system achieves 600 Gb per run, and 2x100 bp read lengths over 11 days of sequencing using a HiSeq 2500 instrument. The HiSeq 2500 also has a rapid mode where less data is acquired more quickly. In this mode, 180 Gb can be acquired in 39 hours with 150 bp reads according to company specifications (136). There is also a smaller version of the instrument available called the MiSeq. As the rapid mode for the HiSeq, this system also delivers a smaller number of reads more quickly. The MiSeq is cur-

[i] The automation of Illumina sequencing library preparation and paired-end spacing is further investigated in Paper II

rently stated to produce 250 bp reads, with a throughput of 15 million clusters (= 7.5 Gb paired-end) per 39 hour run (137).

SOLiD

Sequencing by oligo ligation detection (SOLiD), developed by Applied Biosystems (now Life Technologies), is a sequencing system that is related to the polonator in that it exploits the mismatch sensitivity of a ligase to decode the bases of the library templates (138,139). SOLiD, as described by Shendure in 2005, utilizes degenerated oligonucleotides for ligation. Specifically, SOLiD uses octamer di-base probes in which two positions are locked and coupled to one of four fluorophores. The two locked positions are followed by three degenerated nucleotides and then three universal nucleotides, i.e. nucleotides that will base pair with any other nucleotide. The primer oligo (primer n) is ligated to the di-base probes but is not stripped away as in the earlier sequencing-by-ligation protocol. Instead, the three universal nucleotides (and the fluorophore) are cleaved off, and the next cycle interrogates the bases following that. In this way a series of template bases are interrogated – two bases each cycle, followed by a three-base gap. After a certain number of cycles the primer-oligo ligation product is stripped away and a new primer that is one base shorter than the former primer is added (primer n-1), now probing a new base each cycle due to the one base primer offset. In total five primers with different offsets are used, which enable interrogation of each of the five bases the cleaved di-base probes extends over. However, each base is probed twice since there are two locked positions in the di-base probes (140). This generates another type of output data known as color space, in which each specific base is associated with two fluorophore signals (*colors*) coupled to it. The upside of this is that a single incorrectly detected fluorophore will usually lead to a detectable sequencing error, whereas a true base variation will change two colors. The downside of color space or 2-base encoding is that it is less intuitive to work with, and imposes special demands on the data analysis pipeline.

Like 454 and Polony sequencing, SOLiD uses emPCR to amplify the libraries and achieve solid-phase coupling. EmPCR has traditionally been a labor-intensive bottleneck in the preparation of libraries for sequencing. A recent upgrade enables the SOLiD instrument use something similar to bridge amplification, called wildfire. In this method, denatured template strands are hybridized to a grafted oligo on the surface of the FlowChip and extended to attach covalently to the surface. The temperature is then elevated, causing the hybridized templates (not the covalently attached 3' ends as the case for Illumina) to hybridize to another grafted primer and extend once more in a process called *walking*. The covalently attached templates from the first extension is replicated via a primer in solution hybridizing to the distal 3' end, enabling local *wildfire* amplification. The amplification automatically terminates when it reaches the edge of another cluster since the templates will run out of grafted primers. The wildfire amplification protocol is stated to reduce the sample preparation time from 8 hours to 2 hours (141).

A SOLiD run using the latest 5500xl instrument with wildfire yields 120 Gb according to the manufacturer's preliminary specifications (141). Paired-end sequencing is at the time of writing this thesis not available but projected at the end of 2012. EmPCR-based 5500xl sequencing can yield 75+35 bp reads and roughly 150 Gb in 8 days of sequencing (142).

Newer Technologies

Helicos

The first amplification-free single molecule sequencing system was Helicos' Heliscope, which was described in 2008 (143). In that publication, the M13 phage genome was sequenced with raw error rates of 3-7%. The Heliscope chemistry is based on asynchronous single molecule sequencing, where reversibly labeled nucleotides are incorporated via a polymerase in a stepwise manner, one dNTP at a time, with intermediate washing steps. The sample preparation include fragmentation of the DNA, denaturation, and polyadenylation of the fragments with terminal transferase. A labelled ddTTP is then added last to the polyA stretch to identify the fragment on the glass cover slip with grafted poly-T primers. These primers serve for polymerase extension, and the label enables template identification on the cover slip. Before sequencing starts, dTTPs are used to fill in any free adenines from the polyadenylation. Also present are reversible dVTP (not T) terminators, that block further extension once incorporated. Before sequencing, these blocks are removed and labeled reversible terminators are added in cycles (144). The methods current throughput is in the range of 800 million reads per run, with 35 bp reads (28 Gb)(142). One interesting adaptation of this technology was in direct RNA sequencing, where RNA was directly captured using the poly-T primers and directly sequenced with an error rate of 4% (145). Because of the short read length compared to other systems, and the high error rates stemming from only detecting single molecules, Helicos sequencing has not become widely spread. Helicos sequencing instruments are no longer sold, but the sequencing method is commercially accessible as a service.

Pacific Biosciences

Pacific Biosciences (PacBio) published a paper describing their single-molecule real-time DNA sequencing method in 2009 (146) and commercially launched instruments in 2011. The method is based on a zero-mode waveguide (ZMW) technology, which facilitates zeptoliter detection volumes in the bottom of a well on a chip of 150,000 wells (147). At the bottom of the wells, a *ϕ*29 polymerase bound to the DNA template is deposited inside the detection volume, and sequencing begins with fluorescently labeled nucleotides being excited in the active site of the polymerase as they are incorporated. The four different fluorophores of each respective nucleotide generate signals only when they enter the active site of the polymerase, which thus enables the continuous monitoring of incorporation events. When the nucleotide has been incorporated, the fluorophore is automatically cleaved off (since it is bound to the terminal phosphate) and the signal fades. Although the *ϕ*29 polymerase has been shown to incorporate over 70,000 native nucleotides (60), constant electron excitation inside the active site of the polymerase affects its stability, and so read lengths are limited to 1,000-2,500 bp on average. However, longer reads occasionally occur (above 5 kb). Since this is a single molecule real-time sequencing technology, it is amplification-free. This decreases the representational bias compared to other systems (such as GC and AT bias) but single-molecule monitoring is also strongly affected by stochastic effects, such as when a nucleotide enters the active site but is not incorporated. This generates high but stochastic error profiles and around 85% raw read accuracy. The high error rate can be counteracted by limiting the molecule length and creating pseudo-circular constructs for the ϕ29 polymerase to loop through many times. The read length is therefore a tradeoff against accuracy in this technology. Current instruments can reach 50,000 reads in about 1 hour (148). The cost per base is higher than that for other technologies (142), and the high error profiles are cumbersome to work with, but the long reads have can be advantageous in applications such as de-novo sequencing. There is also the potential to detect methylations from the increased time of nucleotide incorporation to such template residues (149). Recently, a different loading strategy based on magnetic beads has been introduced to reduce the amount of sample required and to increase the number of long fragments entering the wells (150).

Ion Torrent

Ion Torrent (Life Technologies) introduced a sequencing technology based on non-optical detection in 2011 (151). Like 454, the method is based on detecting a byproduct from native nucleotide incorporation. In this case it is the proton of the hydroxyl group on the 3' end of the DNA strand that serves for detection, and this approach was originally described by Pourmand et al. in 2006 (152). By coupling the library to beads and amplifying the templates via emulsion PCR, enough protons can be released during nucleotide incorporation to detect a change in pH. Cyclical addition of native nucleotides one at a time thus enables the detection of base incorporation events. Changes in pH are directly converted into electrical signals on a semiconductor chip, obviating the need for imaging (151). This technology was implemented in the commercial Personal Genome Machine (PGM) in 2011, which has been upgraded by introducing increasingly dense chips for sequencing (314, 316 and 318 chip). The PGM 318 chip delivers 4-8 million reads over 100 bp in length, and over 1 Gb of output in 2 hours (142). There is now a 300 base sequencing kit for the PGM, and a 400 base kit is projected by the end of 2012, stated to be include a "long-read sequencing enzyme" (153). Read lengths are typically limited by problems of dephasing, so polymerase nucleotide incorporation efficiency is likely affecting the Ion Torrent read length – leading to the need of improved long-read enzymes. There are concerns regarding the linearity of the pH change for long stretches of homopolymers, just as for the 454 sequencing method. Homopolymer tracts above 8 bp have been reported to be problematic (154).

A new version of the instrument started shipping in September 2012, called Ion Proton (155). The Ion Proton aims to deliver a human genome for $1,000 within a few hours in 2013. The stated throughput at launch is 10 Gb with the Ion PI chip, and the Ion PII chip that is to be released in 2013 is aimed to deliver up to 60 Gb. Read lengths are up to 200 bp, with 60-80 million reads passing filter from a run of 2-4 hours (Table 5)(156).

Since of 2012, paired-end sequencing has been available for the PGM, and Life Technologies is working to increase its read lengths and insert sizes. For paired-end sequencing to work, the user must remove the chip and fully extend the forward primer to the bead surface (creating a full dsDNA template on the bead). Using undisclosed enzymatic reagents the covalently coupled template is cleaved and degraded, leaving a short stretch hybridized to the other strand that is used as a primer for the reverse read (157). It seems that this is mostly used to increase accuracy rather than to form two discrete reads. It may be that the emPCR imposes limits on the fragment size that can be efficiently amplified on the bead surface. In September 2012, Ion Torrent announced that the third version of the Ion Proton chip (PIII) will feature a 30-minute emPCR-free library preparation termed Avalanche, but further details of this process are not available at this time (155,158).

Chip	Sensors	Throughput	Release
314	1.3 million	10 Mb	2011
316	6.3 million	100 Mb	2011
318	11 million	1 Gb	2012
PI	165 million	10 Gb	2012
PII	660 million	60 Gb	2013
PIII	1200 million		

Table 5. The development of Ion Torrent throughput according to manufacturer's specifications (PII and PIII are currently unreleased). Chips 314-318 are physically smaller chips used for PGM, while chips PI-PIII are for the Ion Proton.

Complete Genomics

Complete Genomics is a company whose sequencing technology is only available as a service. The technology involves a sophisticated DNA preparation processes featuring four circularizations and two Type IIS restriction enzymes (see Chapter 5), and they focus exclusively on human re-sequencing. Their sequencing chemistry is similar to the polonator in that a ligase is used and the degenerated probes are washed away after each step. Read length is increased by having a degenerated anchor probe as well, and incorporating four adapters rather than three through a series of circularization and PCR amplifications steps. The read structure also becomes more complicated stemming from the multiple use of Type IIS enzymes in combination with ligation to incorporate the adapters. The basic read structure is 35+35 bp separated by 400 bp. Each 35 bp read consists of four sub-reads forming a 5+10+10+10 structure. The variation in length specificity of the Type IIS enzymes[i] creates variable gaps and overlaps between these reads, which needs to be accounted for in mapping (159). BGI and Complete Genomics have recently announced a merger (160).

Summary

Some of the key detection and amplification properties of the various sequencing technologies are summarized in the table below.

	Sequencing							Amplification		
	Ligase	Polymerase	Reversible terminators	Fluorophores	Byproduct detection	Patterned array	Single molecule	Planar surface amplification	emPCR	RCA
Polonator	+	-	-	4	-	-	-	-	+	-
454	-	+	-	-	+	+	-	-	+	-
Illumina	-	+	+	4	-	-	-	+	-	-
SOLiD	+	-	+	4	-	-	-	+	+	-
Helicos	-	+	+	1	-	-	+	-	-	-
PacBio	-	*phi29*	-	4	-	+	+	-	-	-
Ion Torrent	-	+	-	-	+	+	-	-	+	-
Complete Genomics	+	-	-	4	-	+	-	-	-	+

Table 6. Current detection and amplification techniques of each technology.

[i] Paper V describes an investigation into the length-specificity of Type IIS enzymes via massive sequencing

The sequencing platform's advantages, disadvantages and major applications are summarized in table 7.

Sequencing technology	Advantages	Disadvantages	Applications
454	Longer reads	High cost per base	De novo sequencing, amplicon sequencing, metagenomics
Illumina	High throughput, low cost per base	High instrument cost	Relevant for most applications
SOLiD	Highest accuracy	Short reads	Resequencing, transcript counting, rare variant discovery
Helicos	Large number of single molecule reads	Short reads, low accuracy	Resequencing, transcript counting
Pacific Biosciences	Longest reads, short runtime	Low accuracy, high instrument cost	De novo sequencing, amplicon sequencing, metagenomics
Ion Torrent	Short runtime	Biased coverage for AT-rich genomes	Targeted resequencing, amplicon sequencing (more applications with increased throughput)
Complete Genomics	Cost competitive	Less even coverage, special read structure	Human resequencing

Table 7. The advantages, disadvantages and applications of current sequencing technologies. The applications are intended to capture the major areas of use, and do not include all applications possible. (142,154,161)

Emerging technologies

There is a great number of active participants working to develop new sequencing technologies using a wide range of different technologies. These include electron microscopy (Electron optica and Halcyon Molecular), nanopores (Oxford Nanopore, IBM, Genia Technologies, NABsys, Noblegen), atomic force microscopy (Reveo), synthesis (LaserGen), and hybridization/microfluidics (GnuBio) among others. Independent or peer-reviewed results relating to these technologies are not available but nanopore sequencing is expected to be the next major development of the field.

Oxford Nanopore

Oxford Nanopore Technologies has been developing nanopore sequencing since 2005. Typically, a nanopore is a narrow channel deposited in a membrane that enables current shifts to be monitored as single molecules flow through the pore (162). The pores may be of biological origin (using proteins as channels) or solid-state pores that are manufactured from various materials (162). At the 2012 AGBT conference in Marco Island, FL (USA), Oxford Nanopore disclosed two sequencing platforms based on biological nanopores that are to be launched late in 2012: the GridION and the MinION. The GridION is a flexible rack-based instrument where each *node* takes a cassette containing the sequencing reagents and nanopores (2000 parallel at launch and 8000 in 2013). Each rack can be connected to other racks enabling flexible throughput. Although 2000 pores is arguably not massively parallel, the technology stands to be very fast, and most excitingly, stands to deliver tremendous read lengths with the potential to directly detect methylations and sequence RNA (163). Oxford Nanopore claim to have read through the lambda genome, ligated with a hairpin, in a single stretch (~100 kb) (164). A raw read error rate of 4% was reported, but is aimed to be reduced to 1% at launch (163). Several bases are detected simultaneously in the pore as *words*, e.g. if the word size would be 3 nt, there would be 64 (4^3) combinations to decipher. The speed through the pore is 100-1000 bp/s, and by projecting 8000 pores in parallel with 20 GridION nodes operating at a speed of 500 bp/s, a human genome could be read through in 15 minutes, at 24x (assuming constant and complete pore occupancy). At this stage no data is publically available to evaluate the potential advantages and disadvantages of this technology. To ensure pore occupancy and long reads, large amounts of DNA could be needed, and sequencing errors seem to be systematic (i.e. some words are harder to decipher than others). The long reads could potentially facilitate de-novo-like sequencing of large genomes more routinely than is currently possible. The MinION is a USB-based disposable sequencing instrument aimed to generate a limited amount of data from 500 pores in parallel, but with the same general qualities as the GridION (163).

Present limitations

Sequencing has undergone a tremendous development in terms of cost reduction during the last decade. The human reference genome (12,13) was completed at an estimated cost of $300 million, and Craig Venter's personal reference genome (165) at $70 million (166). Today a human genome costs less than $10,000. However, the sequencing of human genomes today relies on the availability of an already-complete high quality draft genome to align the reads against. A de novo assembly of human genomes costs considerably more to complete, and is greatly facilitated by the construction of cloned libraries, which is usually not executed (167). This section reviews present limitations of massive sequencing.

Data analysis

High-throughput means a tremendous amount of data. As sequencing technologies continued to increase throughput and decrease cost, computer development has lagged behind. Today, sending a hard-drive via express mail is often faster than sending it digitally over networks (142). Storage capacity on servers can quickly run out, and processing power limits speedy variant calling and assembly. There has been a re-evaluation of raw data, going from saving images captured from the sequencing instrument, to bases called from the images, to binary files of mapped and unmapped reads to the genome – but data is still easier to produce than to analyze it.

Although there is no immediate solution to this problem, much of the data deluge is related to the ease of sequencing short-read shotgun libraries. If libraries were constructed in ways designed to more efficiently answer the question being asked, the amount of data could be decreased. This could potentially be achieved by producing more long insert paired-end libraries, which consume more material and are time-consuming, expensive and difficult to produce, but carry information from a greater physical part of the genome per base sequenced. Another approach is to enrich for those parts that are most likely to be relevant to the question (see Chapter 5). Also, whole genomes are over-sequenced with the current technologies to compensate for bias, accuracy and short read lengths. Bias lead to uneven coverage, which to some extent can be compensated for by more data. If accuracy is low, the risk for errors can be decreased by more reads. Shorter reads are more difficult to map, which also contributes to uneven coverage. Bias and accuracy is in part due to library preparation choices, such as which polymerase is used for amplification, or the use of

PCR-free preparations. As read lengths in massive sequencing increase, it is likely that lesser amount of data will be needed to fully resolve whole genomes.

Run-Time

Although massive throughput is now readily achievable, the sequencing time per run is still quite high (168). Runtimes for the HiSeq 2000 instrument are about 11 days, and this does not include the time required for data transfer and data analysis. Sequencing thus takes multiple weeks, which is unacceptable in some cases such as certain clinical applications. Major efforts are currently being made to improve on this lag-time. The Ion Proton (Life Technologies) aims to deliver a human genome at 20x, including variant calling, in under one day, and the HiSeq 2500 features a rapid mode where two human genomes can be analyzed in 27 hours at 20x. Illumina has internally demonstrated complete workflows using 500 ng of DNA and a PCR-free preparation, to call variants in 50 hours (169). The flexibility of the GridION also seems to facilitate quick runtimes, although little is known about library preparations at this time.

When these improvements of library-to-variant runtime have been made, there will be great demands for fast, cost-effective and low-bias library preparation techniques. The Nextera technology (Illumina), features a transposase-based library preparation that demonstrate great potential for speedy library construction, albeit somewhat biased (80) (see chapter 5). Transposase libraries also have potential to reduce cost since they only involve a single reaction. The most costly part of a standard library is the ligation step, while fragmentation is the most time-consuming if many samples are to be processed. As library workflows are streamlined to reduce the amount of input DNA required, less ligase is needed for ligation thus enabling cost reduction. Enzymatic fragmentation has great throughput capabilities since it can be automated but may increase bias. These steps are further discussed in Chapter 5.

Bias, Accuracy & Read Length

Massive sequencing have since its conception never been able to compete with Sanger sequencing in terms of read length and accuracy. Accuracy has implications for variant calling, and read length is crucial for analyzing structural variation, repeat regions and de-novo sequencing. The recent de novo assemblies of a Han Chinese, and a Yoruban individual (170) were 16.2 % shorter than the reference genome and had 420 Mb missing (of which 2,377 coding exons), since the missing sequence must be artifacts of short-read de-novo assembly this suggests it would be worth focusing on quality sequencing rather than quantity sequencing (171). As stated above, another problem is sequence bias leading to uneven genome coverage. Uneven coverage poses tremendous problem to completely cover genomes at certain depth, creating a need to over-sequence. A recent comparison between Illumina and Complete Genomics revealed less uniform coverage of Complete Genomics' platform compared to Illumina sequencing. Complete Genomics was more accurate (i.e. produced less false positives) in finding single nucleotide variations, but less sensitive (i.e. missed more variants) most likely due to the differences in read length (161).

There is typically a tradeoff to be made between quality and read length. Pacific Biosciences can deliver read lengths of over 1 kb but with 85% accuracy, and SOLiD has a new chemistry for improving accuracy (Exact Call Chemistry) with 99.99% base-calling accuracy (172,173) but reads of only 75 bp. Another strategy is to error-correct long reads obtained by PacBio sequencing using short but accurate shotgun reads from another system. These corrected reads can have accuracy increased to 99.9% and substantially improve assemblies (174). This is an interesting way of combining both accuracy and read length through different libraries and different sequencing systems. The early reports from Oxford Nanopore represent an exciting development in the direction of extremely long reads (>10 kb) with acceptable accuracy (96-99%) but demands on DNA material is presently not disclosed. Single molecule sequencing platforms have an inherent potential for lower sequence bias, as template amplification is not necessary, but they also have less templates to detect signal from, which compromise accuracy. In this way, bias and accuracy are also trade-off choices.

Library preparations can improve on accuracy by elevating coverage with target enrichment, which will reduce the stochastic errors of the sequencing system. There have also been some reports on how to circumvent limitations in read length through directed DNA manipulation and tagging (175,176) thereby increasing the region of analysis. A new such method is described in paper IV. The sequence bias can be limited by simplifying the library preparation and eliminating PCR amplification (177). It is feasible to assume that the repeated steps of PCR amplification between circularizations in the Complete Genomics workflow contributes to a less even coverage, as well as the short read length affecting the ability to map in low complexity regions.

The following chapter provides a more detailed discussion of the opportunities and bottlenecks encountered when preparing DNA for sequencing.

As the biggest library if it is in disorder is not as useful as a small but well-arranged one, so you may accumulate a vast amount of knowledge but it will be of far less value than a much smaller amount if you have not thought it over for yourself.

– Arthur Schopenhauer

Chapter 5

PREPARING DNA FOR MASSIVE SEQUENCING

Sequencing DNA involves four steps: obtaining the DNA (*extraction*), modifying the DNA to suit the biological question and sequencing system at hand (*library preparation*), reading the sequence (*sequencing*), and *data analysis*. The focus of this thesis is library preparation, which can in turn be divided into three basic components: *Fragmentation*, *modification* and *quantification* (see Figure 2).

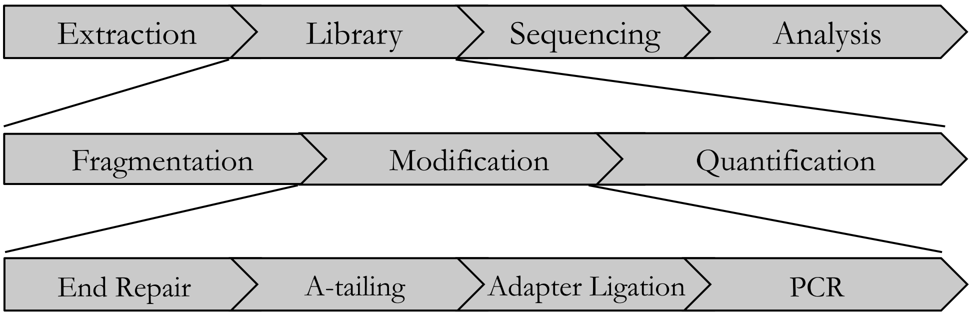

Figure 2. Schematic representation of an idealized sequencing workflow (top) a library preparation workflow (middle) and a typical modification procedure (bottom).

Adapter incorporation is the most central part of a library preparation, as this step involves the addition of synthetic DNA sequences used for subsequent amplification and sequence priming. Amplification is a common but not universal part of library construction and can be performed at various stages. If starting material is limiting, amplification may be performed before fragmentation. It may also be used to enrich for sequence content, either genomic or library introduced modifications such as adapters, or for quantification via quantitative PCR (qPCR).

The bottleneck in sequencing-based research has shifted position with the development of massive sequencing over the last decade. For the human genome project, sequencing was the main bottleneck in terms of cost, time and throughput. As massive sequencing developed, cost decreased and throughput increased. Today a human genome can be sequenced in 27 hours for under $10,000, and in 2013 we will probably see the first one-day $1,000-genome. Preparing the DNA library and analyzing the data have largely become the new bottlenecks. The throughput of library preparations has been significantly improved by efforts towards automating the procedures involved[i], and similar efforts towards standardizing data analysis are appearing. As discussed in Chapter 4, data analysis is often limited by technical bottlenecks such as storage capacity, network speed and computational power. Library preparation is partly limited by the intensive workload, as well as the cost of reagents per sample and the particular demands of the biological question under investigation.

[i] Papers I and II deal with increasing throughput in library preparation for 454 and Illumina sequencing, respectively.

This chapter will discuss standard methods of preparing DNA for massive sequencing, as well as some special approaches.

On libraries

A traditional definition of a biological library is: *a collection of DNA fragments that is stored and propagated in a population of micro-organisms through the process of molecular cloning (178)*. In massive sequencing protocols, nearly all libraries are clonal-free. That is, no live system is needed for selection, storage and propagation. It is therefore necessary to introduce a broader definition of the term *library*. In this work library is used to mean (a) a set of molecules prepared (b) for subsequent analysis using a specific technology (c) to be representative of a specific set of features (d) that are relevant to some biological question or questions.

a.) It is possible to envision a sequencing system that would not need any form of sample preparation. For example, it may be possible to analyze raw blood samples. There have been exciting recent developments relating to nanopores, which suggest that such a system could soon become reality (163). However, these library-free sequencing instruments would probably still benefit from library preparation. In the case of Oxford Nanopore, for instance, a hairpin ligated to the templates would enable consensus readout.

b.) Libraries are bound to specific technologies and are generally not compatible with multiple systems without introducing extra steps to convert the library.

c.) Libraries do not necessarily need to cover the entire genome. For instance, an exome library would aim to only cover exomes. They should however be as representative as possible, i.e. if entire exomes are to be sequenced they should preferably be covered evenly. There is however an inherent nature of library preparation protocols that the more steps they involve, the more bias they are likely to present.

d.) Libraries are intended to fulfill a purpose in a scientific context, e.g. giving a representative view of structural differences between tumor and normal cells. Different libraries have different likelihoods of fulfilling the intended purpose

By analogy to regular libraries, the nature of the books they hold determines the nature of the library, and ordering the books according to their relevant properties (subject, author, language etc.) may alleviate the use of them. Modifying DNA in libraries for sequencing can be viewed as a means to select DNA fragments with properties that are relevant to the question being asked, and order the molecules in such a way as to assist in readout and data analysis.

Amplification

Amplification is performed inside or outside of the sequencing instrument, or both as is typically the case in e.g. Illumina sequencing. There are PCR-free libraries for Illumina, but not amplification-free sequencing, which may be worth stressing. Single molecule sequencers (Pacific Biosciences and Helicos) are the only currently available amplification-free sequencing systems. Libraries should be amenable to storage and re-sequencing. For this thesis, amplifications that are not directly required for the sequencing chemistry will be regarded as being part of library preparation, while *sample preparation* will be used as a wider term also encompasses amplifications such as emPCR and bridge amplification, which are directly linked to the sequencing rather than used for storage. Sample preparation can also be used more loosely to include DNA extraction.

There is a wide panel of methods available for amplifying DNA, and many proprietary kits with reagents designed for whole genome amplifications. The most common amplification strategies are based on PCR and *ϕ*29 amplification. PCR is usually included to elevate the number of molecules in a library to compensate for a lack of starting material, subsequent losses or detection limits. PCR can also be used to quantify a library by means of quantitative PCR (qPCR), in which a fluorophore is linked to the accumulation of product and can be used to follow the reaction in real-time. Every PCR changes the composition of the library and accumulates errors, but different polymerases and buffers have different propensities for doing so. The recommendations from manufacturers of particular polymerases usually vary and are often related to the reagents they can provide. Early estimates of polymerase fidelity suggested error rates in the range of 0.5-1 errors every 10,000 bases (179), but this varies widely between different polymerases and many improvements have been made. The *Pyrococcus furiosus*-derived polymerase *Phusion* (Finnzymes/Thermo Scientific) has been the most popular for Illumina sequencing due to its low error rate (4.4 x 10e-7) (180) but recent studies on GC and AT dependence suggest that the KAPA polymerase (KAPA Biosystems) may be preferable (181) and also has a lower reported error frequency (2.8 x 10e-7)(67).

*ϕ*29-based amplification is mostly used for whole genome amplification (WGA) or amplifying circular constructs, for which rolling circle amplification (RCA) can be employed. The circle is primed and amplified isothermally to produce a long linear concatemer, which can be fragmented and prepared for sequencing. RCA is a linear amplification process. *ϕ*29 is typically used when the starting material is extremely scarce such as when working with single cells (182). In such cases, degenerated oligonucleotides, e.g. hexamers, are used to prime long fragments or circular fragments, which will produce a branched structure where each strand displaced by the polymerase is primed and extended to generate more branches. This exponential form of amplification is termed multiple displacement amplification (MDA) (183). MDA can be used to amplify whole genomes but can display problems of incompleteness and allelic dropout as well as causing the formation of chimeric artifacts (184). Recent developments in MDA and microfluidics have enabled whole genome sequencing of single sperm cells at up to 46% physical coverage (185), and some recent publications indicate that it is possible to reach over 90% (186,187).

Starting material

There is currently no general requirement regarding how much material is needed to prepare a library for massive sequencing. This quantity is much dependent on the library preparation protocol used and the target genome, as well as the sequencing system. In addition to this, DNA purity and quality will introduce variations in yield for the same procedure. In general, for minute samples, special low-binding plastics should be adopted to minimize DNA adsorption to the surface of the reaction vessels (188). Quantifying input material is best achieved by adding a fluorescent dye that specifically targets dsDNA, such as SYBR Green. This approach is better than absorbance-based methods such as Nanodrop, in which the presence of organic solvents, single nucleotides, RNA and ssDNA will lead to overestimattions of the amount of constructive genomic DNA. Both these methods give the amount of DNA (weight) but cannot directly measure the complexity of the material. The target genome has a profound influence on the amount of DNA required to make a sequencing library. One human haploid genome weighs about 3.5 pg, and as a rough simplification, 1 Gb can be taken to weigh 1 pg. A bacterial genome, such as that of *E. coli* will have a size of roughly 5 Mb and will thus weigh about 5 fg. The duplication levels (i.e. the number of redundant reads from the amplification that do not contain additional biological information) achieved by sequencing 1 ng of *E. coli* DNA or 1 ng of human is a poor comparison. Further, duplication levels are dependent on how deep you sequence. It is better to think of amount of DNA that goes in, and number of unique molecules in the final library (this number can be estimated from sequencing by plotting the number of reads and the number of duplicates, then extrapolating the size of the library). The number of unique molecules in the final library and the number of bases possible to sequence of each molecule can be said to give the *potential* of the library. A human library needs 700 times more potential than an *E. coli* library to achieve the same level of coverage. Considerable effort has been made to reduce the amounts needed for sequencing (i.e. increase the yield of library preparation), and reports of shotgun sequencing genomes starting from as little as 10-20 pg has been reported using Nextera (80,189), albeit with incomplete coverage. The Nextera library preparation is an interesting example of how new enzymatic reagents can be used to decrease the amount of starting material needed, and is based on a modified version of the Tn5 transposase (78,79), but instead of inserting a transposon, the enzyme carries the sequencing adapters needed for sequencing. In this way, both fragmentation and adapter ligation are achieved in a single step. It is difficult to directly compare the reports from Nextera sequencing from low starting material, but 2 million uniquely mapped reads where reported using 10 pg human DNA (80), and a 20 pg mouse sample DNA yielded 0.4x coverage (189). The potential of a Nextera library can be roughly estimated from this. If 10 pg yields 2 million unique fragments with an average of 200 bp genomics sequence, 2x100 bp reads give 400 Mb of sequence, or a yield of 40 Mb/pg genomic DNA. There is also a microfluidics-based library preparation instrument available from NuGen, called the Mondrian SP Workstation. The Mondrian workstation automates sample preparation and can make libraries from 1 ng of bacterial genomes. Standard library preparations usually require starting material in quantities ranging from 500 ng to 5 μg for the different systems. The first amplification free approach to Illumina sequencing used 0.5-5 μg of starting material and demonstrated more even genome coverage and improved GC-bias than standard methods involving PCR (177). Illumina have by reference to internal preparations, produced amplification-free libraries from as little as 100 ng (169).

Fragmenting DNA

Before a genome can be sequenced it must be broken down into manageable pieces. This process is called fragmentation or shearing. There are in principle three types of methods for fragmenting DNA: physical, enzymatic and chemical. The most common is physical, where the DNA is broken into pieces by some external force. The most widespread procedures for this type of fragmentation are nebulization (gas pressure), sonication (high frequency sound), or hydroshearing (fluidic pressure change). Sonication via Covaris instruments have been commonly adopted due to low sample loss and low risk of contamination (non-probe based shearing). Covaris sonication produces relatively focused fragment lengths up to 5 kb, while fluidic pressure change is usually used for longer fragments (5-20 kb). The term hydroshearing introduced to describe several existing methods for fragmentation, such as the Hydroshear (Digilab) and g-TUBE (Covaris), as well as regular pipetting. Enzymatic shearing has greater scalability in terms of multiplexing. NEB offers *Fragmentase* based on *V. vulnificus* nuclease that introduces random nicks and T7 endonuclease for cleaving the opposite strand. DNase I has also been used for fragmentation (188). The general worry of using enzymes is the risk of bias, i.e. that certain sequence combinations may be cleaved more extensively than others, as well as being sensitive to epigenetic modifications. Enzymes can also generally be more difficult to control in terms of fragment size, and often produce a wider range of different fragment lengths. One of the greatest advantages of using enzymatic cleavage is that all products should have biologically active ends, i.e. ends that can be manipulated with enzymes. Physical (and certain chemical) fragmentation risk cleaving C-C bonds, which will be resistant to subsequent enzymatic reactions. Chemical fragmentation is the least utilized shearing method but is potentially very cheap since chemicals can be obtained in bulk. One such example would be the Maxam Gilbert sequencing method although this has the obvious drawback of sequence dependence (99).

Modification

The modification is the most diverse part of library preparations, and typically what distinguishes the protocols from each other. The term modification regards any type of directed DNA manipulation used to modify the content of the library. The modification step often also includes steps of selection, enrichment and purification. Enrichment is the intentional favoring of certain kinds of chemical properties in the library, such as a biotin moiety, circular topology, or a particular sequence composition. Selection refers to size selection for specific fragment lengths, and purification is the removal of previous reagents or reaction byproducts. The broad nature of the modification step does not permit comprehensive characterization, and so the aim here is to provide a representative view of the different procedures that typically are used.

The most general modification of a library involves attaching sequencing specific adapters to the fragmented DNA. These adapters are used for priming amplification and/or sequencing, and potentially also multiplexing (see the section below). The simplest way to do this is to add T4 DNA polymerase, which will extend from recessed 3'-ends, and remove 3'-overhangs (*end repair*). Adding a polynucleotide kinase will phosphorylate the 5'-ends to produce fragments that are ready for blunt ligation to the adapters. It is also common to introduce an adenine 3'-overhang using a polymerase without proofreading activity (*A-tailing*). An adapter with a complementary thymine-overhang can then be ligated to the end. When the adapters have been ligated, a PCR step directed towards the introduced adapters can be applied to enrich for the ligated product and make enough material for quantification and sequencing.

Multiplexing and barcoding

In many cases, the sequencing throughput of the system significantly exceeds that required to address the biological question at hand. For such cases, various techniques of multiplexing analysis have been developed. In its simplest form, multiplexing is achieved through physical separation. For instance, the 454-sequencing system optionally enables gaskets of 4, 8 or 16 chambers, in which different libraries can be loaded. Similarly a flow cell on an Illumina instrument consists of 8 lanes. However, a more versatile way of achieving parallel analysis is to add molecular barcodes into the library itself, since this enables a continuum of demultiplexing features – anything from a few static barcodes to millions or billions of degenerated tags (Figure 3). A barcode (sometimes referred to as a tag or index) is a stretch of synthetic nucleotides added to the library as a unique internal identifier, that will subsequently be used to distinguish between the sequences of that library and those of other libraries sequenced simultaneously. An early barcoding strategy for the 454 system was described by Meyer et al in 2008 (190). In this paper, the authors added barcodes to the dsDNA library prior to 454-adapter ligation. Nowadays it is more common to include the barcodes in the adapters from the beginning, and to devise sets of different sizes depending on the scale needed. With error correcting hamming codes, barcodes can be designed to counteract the relatively high error rates of massive sequencing (191,192). Open-source programs for designing barcodes that present a minimal risk of misclassification have also been developed (193).

Dual barcoding (Figure 3) is the idea of attaching two independent barcodes to the library fragments, which has recently been adapted for Nextera and TruSeq library preparations of the Illumina system. In this way more secure demultiplexing can be achieved or more samples can be processed simultaneously.[i] Attention has also been focused on demultiplexing individual molecules, which is achieved by attaching unique (degenerated) molecular barcodes to each unique molecule of a library. This is useful for tracking individual clones within a complex mixture (194-197).[ii]

[i] Paper III is an early adaptation of dual barcoding to process thousands of samples simultaneously.

[ii] Paper IV further explores the use of degenerated barcodes.

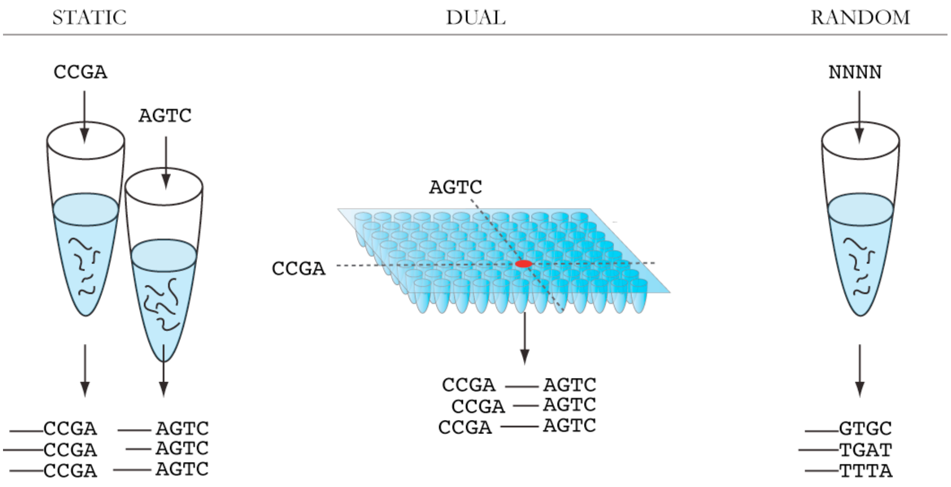

Figure 3. Barcoding libraries can involve the use of static barcodes that code for a specific library that enable the pooling of sequences, dual barcodes that can be combined to increase levels of multiplexity or accuracy when pooling libraries, or random barcodes that code for individual molecules within a complex sample.

Targeted enrichment

Focusing on a subsection of the genome based on its known sequence content is called targeted enrichment. The upside of enriching in this way is that it reduces the cost of sequencing since the experiment focuses on the regions of interest. The downside is increased library preparation cost, as well as an inherent reference bias (i.e. an assumption that the genome under study looks roughly the same as the reference genome that is used to design the probes or primers). One way to classify targeted enrichment strategies is: *Hybrid Capture*, *Selective Circularization* and *PCR amplification* (198). Hybrid capture can be performed either on solid phase (199) or in solution (200), stemming from technologies of highly dense in situ synthesis of nucleic acids on solid surfaces (201). The synthesized oligos are used as baits to hybridize to the regions of interest, thereby enriching the sample for target content. This technique offers the greatest target capacity, and reagents for targeting entire exomes are commercially available from different providers (e.g. Nimblegen, Agilent and Illumina) with different specifications in the quantity, length and density of the baits (202). Selective circularization includes molecular inversion probes (MIPs) (203) and selector probes (204,205) (HaloPlex). The MIP technology is based on fragmenting the genome and using the fragments as templates to bridge synthetic probes specifically designed for the targeted regions, thereby producing circular constructs. The single-stranded region between the hybridized regions of the synthetic probes is extended and ligated, and the circular constructs are either amplified by RCA or PCR (198,203) thereby enriching the targeted regions. Selector probes are deployed onto a genome fragmented by restriction enzymes, and are complementary to the restriction pattern. The selector probe bridges the restriction fragments, which enables extension, ligation and subsequent amplification (198). Multiplex PCR has traditionally not worked well for more than ten targets in parallel (204), primarily because of primer-dimer build up (206). One way of counteracting this is highly multiplexed simplex PCRs via microdroplet miniaturization (207). Raindance is a company that currently offers systems for droplet generation for up to 20,000 targets (208). Another microfluidics-based approach for multiple simplex PCRs are offered by Fluidigm designed for 48 targets and 48 samples per array and run. Multiplex PCR is also aided by improved primer design algorithms and library preparation protocols so that thousands of amplicons can be generated in parallel. Today, Ion Torrent and Illumina both offer custom options for designing primers for multiplex amplification 1,536 amplicons. Illumina uses a kit based on extension and ligation to introduce universal handles used for product amplification (209), while Ion Torrent specifically degrades primers to reduce background (210).

Selection and Purification

Purification steps are necessary during library preparation to ensure optimal enzymatic activity, limit byproduct formation, and to concentrate the sample. A common way to purify samples for Sanger sequencing is ethanol precipitation, but ethanol precipitations are time consuming and labor-intensive. Silica based spin column purification (211) is a quick and simple way of purifying single samples, and is consequently the standard method for sample purification when preparing libraries. When higher throughput is needed carboxylic acid coated paramagnetic beads in combination with PEG precipitation is preferred (212).[i]

Size selection is performed to ensure that the library is amplified and sequenced properly in the sequencing system, as well as to increase the informational content of a library used for paired-end sequencing directly in the sequencer. The traditional way of size selecting is to run an agarose slab gel, and manually excise the fragment lengths of interest and purify the DNA from the excision. This is time-consuming, labor-intensive, and tends to give variable results. Two automated systems to perform gel-excision are Caliper XT (213) and Pippin Prep (Sage Science), but their throughput is limited to 4 samples in parallel. [ii]

Quantification

Accurate quantification of the libraries is important for proper performance in subsequent sequencing reactions, as well as to ensure that barcoded libraries can be evenly pooled. Library quantification is based on three types of methods: fluorescence detection, amplification and sequencing. Absorbance can be used to measure DNA concentration but is more sensitive to contaminating substrates and is for this reason discouraged (214). Fluorescence combined with gel electrophoresis enables the length composition to be determined. Quant-iT assays (Invitrogen) in combination with automated capillary gel electrophoresis systems such as the Bioanalyzer (Agilent) is the most common approach to quantify a library. The amplification approaches are based on either qPCR or digital PCR (215), in which fluorophores are used to report the accumulation of amplified library product. Both these methods are more sensitive than direct quantification (216). Digital PCR separates the targets into reactions in which one or zero target molecules are present. If a signal is obtained the reaction contained a target molecule, and the fraction of signal-containing

[i] Generic automated PEG precipitations of DNA preparations for massive sequencing is further characterized in Paper I.

[ii] Paper II describes an automated PEG precipitation bead-based approach for size selection.

droplets can be used for quantification. This digital response further increase sensitivity compared to qPCR, and commercial platforms are available from BioRad and Fluidigm. Quantification by sequencing is the most expensive and time-consuming option, but also the most reliable. Typically, this option is relevant for highly multiplexed libraries that have been quantified by the other approaches and pooled accordingly. A sequencing test run is then performed to verify that the pooled libraries are represented evenly, or alternatively, the result is used to prepare a new pool that will be used for the full sequencing run.

Long Insert Paired-End Libraries

The first paired-end libraries were conceived (217) and created as part of the HGP (218). The purpose was to combine the process of mapping and sequencing into one step. By sequencing both ends of a molecule with an approximately known insert size, the map could be derived directly from the sequence data. The information of read spacing allowed the scaffolding of contigs. By typical terminology used today (e.g. by SOLiD and Illumina) *paired-end* sequencing means the sequencing of both ends of the immobilized library fragment in the sequencing system. This is usually a short molecule (below 1 kb). Mate-pair sequencing, refers to longer spacing of the reads that have been constructed during library preparation. The documentation for the 454 system refers to these libraries as paired-end libraries and so for clarity they are here described as long insert paired-end libraries, or *LIPE libraries.*

The basic idea of a LIPE library is to overcome the problems associated with short sequence reads. By knowing the approximate spacing between two shorter reads, you are effectively probing a larger part of the genome than you otherwise could. This increases the likelihood of discovering large changes (structural variation) and spanning low-complexity regions otherwise hard to cover. The JCV genome was completed using Sanger sequenced paired-ends from BACs, fosmids and plasmids, and detected over 900,000 structural variants (166) but the James Watson genome, completed by 454 re-sequencing (219), detected much less (166), illustrating the gap between traditional approaches for human de-novo sequencing and methods used for massive re-sequencing. Cloning-free LIPE libraries have been developed to bridge this gap, and the methods for their preparation are continuously being improved. A LIPE library is constructed by circularizing DNA fragments of known average size (see Figure 4). The two ends that are joined are sequenced, and their distance is determined by the library preparation. The approximately known distance can then be used to detect structural variation as deviations from the distance or read orientation.

The library preparation protocol for the polonator was the first cloning-free mate pair protocol. This library was constructed by the circularization of a 1 kb DNA molecule and a universal spacer containing two Type IIS enzyme (MmeI) restriction sites. By using MmeI to digest the circular molecules, 17-18 bp tags at each end of the 1 kb construct could be extracted and amplified on streptavidin beads (124). For SOLiD, this protocol was later extended to 26+26 bp reads by using EcoP15I instead of MmeI (139). Korbel et al. reported the first non-clonal long range (above 1 kb) mate pair protocol using 454 sequencing (220). This protocol has since been revised to include a Cre-lox recombinase, and can produce inserts of up to 20 kb. A common trend for clonal-free mate pair libraries is that the longer the fragments are, the lower the potential of the resulting library is going to be. To compensate for the circularization inefficiencies, more input DNA is needed, which in turn limits the application of the protocol. A general recommendation for 454 Cre-lox libraries with insert sizes of 20 kb is to start with 30 μg of DNA per half of a picotitre plate (0.5 million reads). This can be contrasted with a Nextera shotgun library producing 2 million unique reads from 10 pg.

A generalized LIPE library is prepared in three steps: circularization, fragmentation, and enrichment. Circularization achieved using either a recombinase (454), a ligase (Illumina) or by hybridization of ligated adapters (SOLiD). Fragmentation is achieved by physical (454, Illumina) or enzymatic (SOLiD, nick-translation+nuclease) means, and enrichment is based on biotin incorporation (454, Illumina, SOLiD) although inverse PCR can also be used (221).

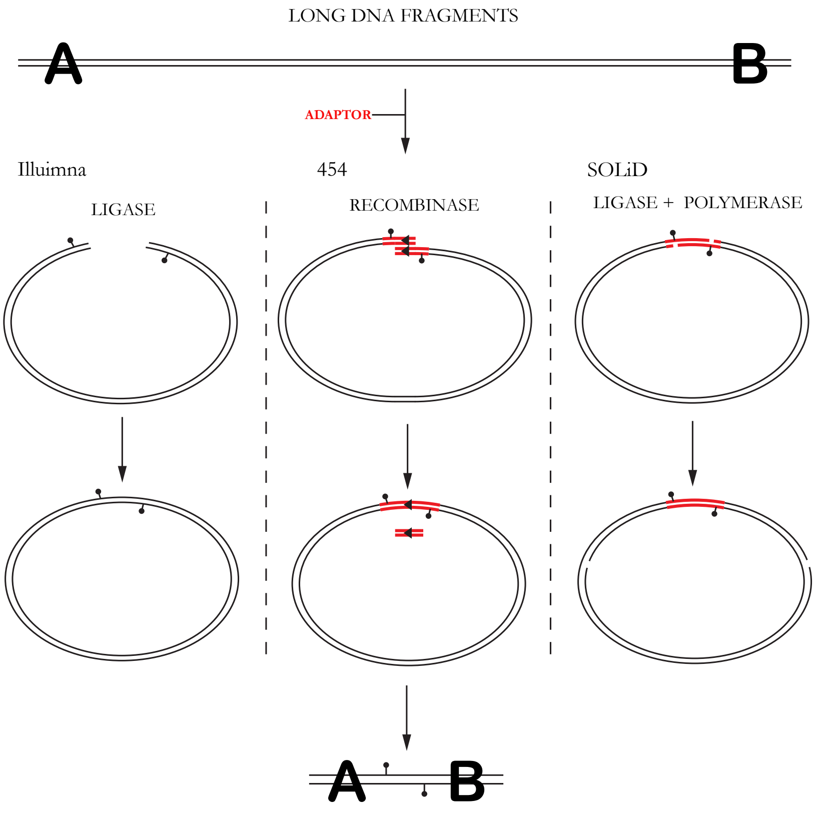

Figure 4. Schematic view of long insert paired-end libraries for Illumina, 454 and SOLiD. The gray box illustrates the synthetic adapters used for circularization; black circles indicate biotinylated nucleotides added either using adapters or by end-repair. A and B represent the two ends that are joined in the final construct.

The standard Illumina protocol for mate pairs is based on blunt end ligation, via end-repair using biotinylated nucleotides. The 454 Cre-lox system, however, has been revised and adapted for the Illumina system (221,222). SOLiD's protocol was originally based on the polonator protocol and involved the restriction enzyme EcoP15I (139). As read length increased, they adopted a different strategy based on an internal linker ligation and nick translation shearing, which is also used in the Ion Torrent mate pair sequencing protocol.

Specialized libraries

This section provides a brief overview of some variations in library preparation protocols. DNase-seq is commonly used to enrich the library for open chromatin structures. DNase I is an endonuclease that is more prone to cleave where the DNA in the nucleus is more exposed (not wrapped around histones), and can therefore reveal patterns of gene regulation and expression that are useful when examining the differences between different cell types (223,224). Bisulfite sequencing can be used to detect methylation patterns in the genome. Bisulfite deaminates unmethylated cytosine which can be converted to uracil by alkaline desulfonation (225). The uracils will be detected as thymines in sequencing, so any detected C->T substitution will correspond to an unmethylated C residue in the untreated genome. Bisulfite sequencing can be further focused by reduced representation bisulfite sequencing (RRBS), in which the genome is first cleaved using methylation-insensitive restriction enzymes that target CpG islands (226).

Complete genomics have a library preparation protocol that is specially designed for their sequencing platform and involves four circularizations of 400 bp sized fragments that are subjected to intermediate Type IIS enzymatic digestion, adapter incorporation and PCR to produce a circular construct carrying four incorporated adapters. The circular construct is then primed for RCA in order to produce DNA nanoballs that are used for sequencing (159). Complete Genomics have also developed a procedure for long range haplotyping in which fragments of 10-1,000 kb are diluted (using as little as 10 cells), MDA amplified, and tagged for spatial separation in 384 different pools. The dilution limits the likelihood of allelic co-occurrence in that particular pool, enabling long range clustering of physically connected reads. In this way they are able to phase SNPs over entire genomes (97% of heterozygous SNPs)(227).

A recent library preparation protocol what was specially adapted for ancient DNA uses CircLigase to ligate single stranded adapters to both strands of the highly degraded sample. This ssDNA library preparation increases efficiency and has been used to sequence the Denisovan hominin (228).

Cloning is generally not used to create libraries for massive sequencing, but is some times necessary. Depending on the type of project, the limits on read length of massive sequencing can greatly affect the outcome. De-novo sequencing of large genomes is tremendously difficult using only short reads, and although possible to achieve (229) there is a price to be paid in terms of completeness. Fosmid cloning has therefore been adapted for massive sequencing projects seeking to resolve difficult genomes (167).

Although the libraries discussed above represent only a fraction of all the published variants, they are presented here as examples of the widely different opportunities available for addressing specific technical limitations or biological questions. There will be an ongoing need for protocol development as new sequencing technologies emerge and mature, and new enzymatic tools and other reagents become available.

Investigation

Five papers constitute the foundation of this thesis. All of them strive to aid sample processing for massive sequencing by modifying library preparation protocols. Papers I and II aid the efficiency of massive sequencing by increasing sample throughput via automation and purification of DNA by PEG precipitation and magnetic beads. Papers III-V concern library protocol development, where paper III and IV are molecular barcoding protocols to increase levels of multiplexity and read length, respectively, and paper V presents a systematic characterization of enzymatic reagents (Type IIS restriction enzymes) used in library preparation for massive sequencing. The following section briefly describes these papers.

AUTOMATION (I-II)

Papers I and II

An important aspect of preparing DNA for massive sequencing is scalability. As the throughput of the sequencing instruments increased, so did the demands placed on sample processing procedures. This first work was based around a fully automated DNA purification protocol enabling robots to perform the sample processing prior to sequencing. We had systematically tested an automated protocol for DNA purification using carboxylic acid coated paramagnetic beads, which proved to be very useful for capturing DNA precipitations in solution. When DNA is precipitated using PEG/NaCl, fragments of particular lengths can be selectively purified. For 454-sequencing this is important as it will affect the read length, and therefore the utility of sequencing. At this time, very little had been published of how to automate the protocols for the massive sequencing instruments. We found that these CA-purifications using PEG/NaCl precipitations worked just as well as any columns that were used in the standard protocols, but with better yields and a flexible point of fragment length selection. It turned out that this chemistry was the basis of a commercial kit from Agencourt called Ampure beads. In this product, Ampure beads are predisposed in a PEG solution, and the user can adjust the lengths of the precipitated fragments by altering bead-to-sample volume ratio. The optimal bead-to-sample ratio may vary from batch to batch and therefore required calibration. Moreover, the amount of beads added cannot be controlled independently of PEG-concentration. In this paper we presented the first generalized description of the PEG concentrations required for controlling fragment size when capturing the precipitate on solid surface. With this protocol in place, the 454 library preparation was performed automatically and benchmarked to a manual procedure using Ampure beads. Building on the protocol of paper I, we also automated the protocol for Illumina sequencing presented in paper II. In this paper, we described the first automated protocol for gel-free size selection prior to Illumina sequencing. The general idea had been used previously with Ampure beads (230), but the generic nature of the protocol we describe and the separation of PEG and beads allows for a generic automated size selection protocol that tolerates many sample concentrations and fragment sizes in parallel.

LIBRARY DEVELOPMENT (III-V)

Paper III

One problem early projects faced as massive sequencing began to scale in throughput was to adapt experiment designs to match the increased sequence capacity. In this study, we wanted to take advantage of the massive throughput of the 454 FLX Titanium instrument for amplicon sequencing. The barcoding kit supplied by the manufacturers at this time contained 10 different barcodes, and was complemented by a set of 143 extended oligo designs that could be ordered from a third party. For a phylogenetic project in house we had a project seeking to sequence roughly 5,000 amplicons. The standard 454 set of 10 barcodes, and the 16 lane gasket for the picotitre plate would accommodate 160 different samples, and the extended set 1,728 samples. The manual problems of handling this many barcodes are significant. Even more problematic with a linear barcoding strategy is the expansion of cost. Every new sample requires a new set of primers, so for 5,000 samples this is a significant contribution to overall cost. A set of primers costs roughly 200-500 SEK depending on length, which would require around 150,000 SEK spent on primers alone (including the use of the 16 lane gasket). What we adopted was a two-dimensional strategy where one barcode was used to code for a specific position on a 96 well plate (96 unique barcodes, *position-tag*) and another barcode was added specific for the plate used for the amplification (in this case, 26 unique *plate-tags* were used). This enabled barcoding of 2,496 samples, and by also taking advantage of the physical regions on the picotitre plate the whole set of samples was accommodated. In this way, another set of 96 samples can be added by just adding one extra *plate-tag*. To avoid the risk of introducing chimeras during library preparation, the position-tag was added to both ends of the amplicons, but a slightly revised tag design could be adopted, using an 8+12 position-tag scheme whit 8 unique tags coding for the rows of a microtitre plate, and the 12 unique tags coding for the column. Using this scheme and the 16 lane gasket, only 24 tags would be needed for 5,000 samples. This *two-tagging* approach based on a position/plate-tag combination greatly increased throughput in this project, and by taking advantage of the automation of the 454 library preparation, the 52 PCR plates were readily be prepared in just 2 days. Of the 3,700 successfully amplified samples, 3,500 where covered with 20 reads or more, which illustrates the potential for genotyping a many samples within a pool.

Paper IV

In this project we focused on one of the great drawbacks of massive sequencing – read length. Sanger sequencing of a human genome requires approximately 6x genome coverage (231), whereas for Illumina sequencing, the standard is 30x. Much of the need to over-sequence is due to limitations in read length. The improvements in sequence throughput over the last decade has been exceptional, going far beyond e.g. the increases in transistor density described in Moore's law (which predicts doubling every 2 years). However, read lengths have matched Moore's law relatively closely. The possible exception is Ion Torrent, for which the read lengths at least initially have increased more rapidly (Figure 5).

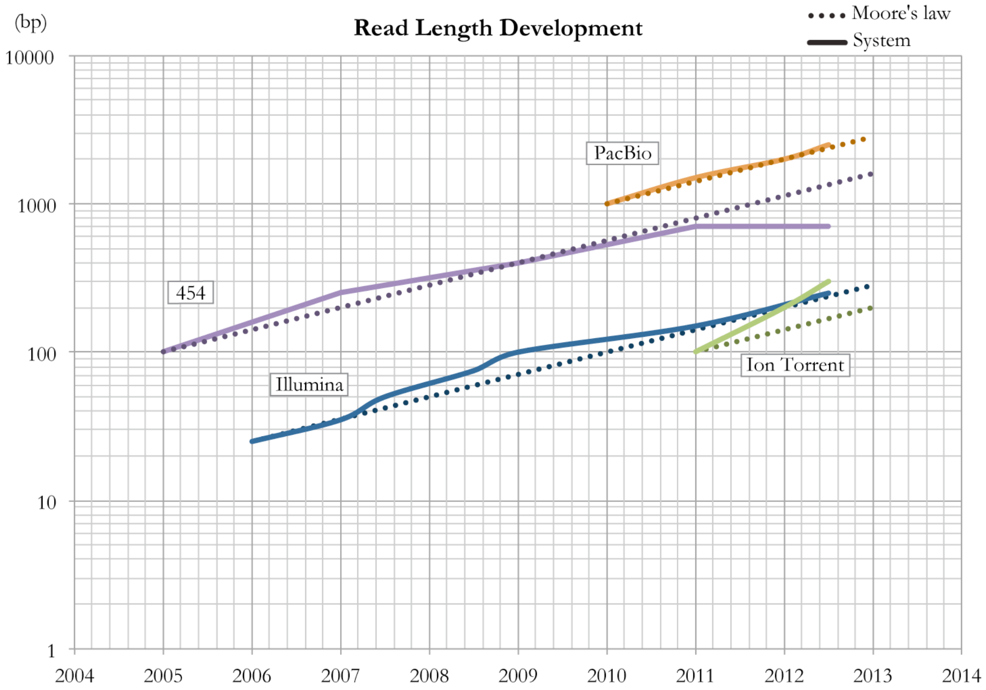

Figure 5. Read length development for four different sequencing systems (SOLiD was excluded due to their focus on accuracy rather than read length). Dotted lines show the increases predicted by Moore's law, which is commonly used as reference for sequence throughput development.

We wondered whether it might be possible to use a series of enzymatic steps to convert some of the throughput of massive sequencing into read length – i.e. to somehow link many short reads together. For this to work, we needed automation, barcoding and an exonuclease. Exonuclease III had been used quite extensively in the 80s and 90s in conjunction with cloning and Sanger sequencing (232-234), but had never been adapted for use with cloning-free sequencing technologies post-HGP. We developed several barcoding steps and introduced a three-level barcoding structure: one barcode encoding a specific fragment, one coding for a general target or sample, and one coding for position. on the fragment To achieve this, each molecule of the original pool received a unique barcode from a degenerated pool via PCR, as well as a static barcode coding for the target. A second PCR is applied to amplify each tagged molecule (so that several copies of each tagged molecules are created). By protecting the end containing the *target-tag* and the *fragment-tag*, the amplicon can be unidirectionally degraded. By sub-sampling and interrupting this degradation in an automated protocol, variable degradation lengths can be achieved for each indexed starting molecule. The third barcode (the *time point-tag*) is added to each subsample to give the approximate length at which the corresponding pool was degraded. These degraded ends of the fragments later enable multiple short-read starting positions spread over a larger distance. The time point-tag gives the approximate position, the *target-tag* indicates which sample or amplicon the reads belong to, and the *fragment-tag* specifies which clone it belongs to. We illustrated this approach by sequencing the lambda genome, p53 and mtDNA with eight position-tags allowing for over 3,000 bp to be continuously covered.

Paper V

Many DNA protocols involve Type IIS enzymes – enzymes that recognize a specific sequence and cleave unknown sequence at a specified distance from the recognition sequence. The use of such enzymes are the basis of many libraries, such as LIPE libraries and Complete Genomics' library preparation. Some of the recent reports had indicated unspecific behavior in the cleavage reaction of these enzymes (124,159). We wanted to characterize these enzymes in terms of their length specificity and sequence dependence to facilitate the development of future protocols. In this first large scale characterization of Type IIS enzymes, we demonstrate an automatable workflow for studying the phenomenon of unspecific cleavage (termed slippage) and sequence dependence using massive sequencing. Illumina sequencing provided very sensitive and accurate insights into these events and allowed us to demonstrate a wide diversity of properties of the studied enzymes. We argue that these properties are highly relevant for other researchers in the field, both developing methods and doing functional or biological studies of these enzymes. These results point to a need of more extensive standardized studies and availability of resources. We analyzed 15 Type IIS enzymes in parallel, using roughly 11 million reads, and observed varying levels of slippage even among isoschizomers. The enzymes also differed in terms of sequence dependence. We believe this new way of separate classification of sequence cleavage is an important tool for restriction enzyme characterization and will provide valuable insights into their mechanisms of operation and future protocol development.

Outlook

The imminent release of nanopore sequencing stands to transform biology research once more. A sequencing system without intrinsic limitations in read length will, regardless of its sample preparation requirements or systematic error rate, radically alter the options available for resolving genomes. The high-throughput technologies that have accumulated over the last decade are likely to mature and be further standardized to deliver cheap and rapid re-sequencing of genomes. It is likely the one-day $1,000-dollar genome will be achieved already in 2013. Long reads however, will permit de-novo-like sequencing on a great scale, which should generate less bias and increase the resolution of low complexity regions and structural events. Paired with accurate short-read ensemble sequencing, this new paradigm of single molecule sequencing is likely to instate a new era of sample preparation research in order to explore all the potential windows to biology this new technology might bring. It is with great humility to the collective effort of the method developing scientists in academia and industry I conclude this thesis. The transformations brought about by Galileo and others, enabling Hooke to discover cells in 1665, may in this century be surpassed by the scope into the nuclei.

Acknowledgements

Tack **Joakim** för möjligheten att doktorera i din grupp. Och tack **Stefan** och **Elin** för de varma orden på Öland den sommaren som fick Joakim intresserad till att börja med! Det har varit en fantastisk resa där 454 i pyrorummet på Albanova känns som en svunnen tid långt långt bort. När jag satt på ditt kontor och du ritade upp DNA-nedbrytning som en funktion av tid anade jag inte riktigt vilket explosivt fält jag gav mig in i. Jag tycker du alltid varit intresserad av mina idéer oavsett om de varit vetenskapliga eller sociala och det har verkligen fått mig att trivas!

Stefan och Caroline, ska ha tack för exjobbet också! För att ha fått mig intresserad av forskning på ett praktiskt plan. På första mötet med dig Stefan pratade du om att man var tvungen att ha gröna fingrar i det här fältet och jag var tvungen att dölja min panik eftersom jag upplevde det som att jag alltid misslyckats på alla labbar hittills. Det gick ju inte mycket bättre på exjobbet egentligen, men jag lärde mig i alla fall otroligt mycket, och på något märkligt sätt var det ändå roligt.

Tove och Elin som handledde mig under Affibody-tiden. Ni är fantastiska på att skapa ett mysigt och kreativt klimat! Det var då det började gå bra att labba. Vilket lärde mig att inte klandra mig själv så mycket för resultaten. Det vanligaste är ju faktiskt att det är proteinernas fel när saker inte fungerar. Kanske var det därför rätt att fokusera på DNA istället.

Nåväl, tillbaka till KTH och doktorera. **Max!** Vad hade jag gjort utan dig. För det här med veckomöten med Joakim var man inte bortskämd med i början. Utan dig hade det här aldrig funkat! Framför allt uppskattade jag alltid hur du muntrade upp när resultaten tröt. ”Det är ju då ett möte behövs som mest!”

Magnus, du var ju aldrig nån officiell handledare men likväl diskuterade vi saker länge och väl, högt och lågt. Det hade aldrig gått lika bra att problemsöka eller prova helt nya vägar utan din input. Att du ville prova nya vägar när jag väl hade bestämt mig lärde mig också att stå på mig.

Henrik, ingen officiell handledare du heller men sannerligen min go-to-guy. Du visade mig labbet på Albanova, MBS:en, protokoll, kurser man kunde läsa, ja i stort sett allt. Jag fick anstränga mig hårt för att hitta på något eget.

Afshin för din oändliga expertis och intresse för detaljer och för att dörren alltid är öppen. **Peter** för att du delar med dig av dina hundar! **Anders A.** för att du är så inspirerande att prata med och lätt att hitta på projekt tillsammans med. **Lasse** för turerna med TES även om det aldrig riktigt kom dit vi tänkt. **Per-Åke** för att du alltid svarar på mina frågor och villigt går in i diskussioner angående forsningen på skolan. Plus alla andra PIs som korsat min väg, såväl vetenskapligt som vid kaffemaskinen**.**

Tack också till alla fellow-doktorand/postdoc-kollegor som kommit och gått genom åren. **Julia** för att du är så grym på allt och alltid gör mig glad! AGBT-resan är en av de absoluta höjdpunkterna! **EB** för att du alltid är så nära tillhands för att svara på frågor. Jag har ju även börjat nästla mig in i era projekt för att visa min uppskattning, som ni märkt. **Mårten**, som lämnat men ändå är kvar. Du fick mig nästan att börja med klättring vilket ju är en enorm bedrift, och det finns alltid utrymme för en ordvits eller en sär skrivning. **Daniel** för att du alltid har ett svar. Jag tycker det är en helt sjuk förmåga, och en bra matchning eftersom jag alltid har frågor. Supertack för att du korrade avhandlingen! **Patrik** för att vi alltid kunnat lösa våra meningsskiljaktigheter (eller i alla fall leva med dem) tack vare din raka stil, och för att du alltid är villig att diskutera forskning och titel på en avhandling. **EP** för vår tid med MBS:en och mitt första PEK! **Kimi** för sociala klubben och framförallt alla roliga diskussioner. **Anders** min partner-in-beer! **Anna K** för att startat upp sociala klubben och för det goda lab-sällskapet. **Basia** för att du är så bra på att sätta saker i perspektiv! **Håkan** för alla turer med (tår)dropparna! **Pawel**, **Johan, Fredrik**, **Anna**, **Johanna**, **Beata**, **Pelin**, **Lilia**, **Paul**, **Mahya**, **Sanja**, **Yajing**, **Peter** – utan er hade inte gruppen varit så bra som den varit! **Ulrika**, **Karin**, **Erik**, **Sebastian**, **Joel** och andra roomies. **Andreas**, **Danne, Helena och Joel** från Affibody-tiden och alla glada fester. **Maja**, du hjälpte mig när det var som tuffast i doktorandrådet och jag förlåter dig ständigt för att du är en sån rojalist. **Anna W** och **Simon** för att ni är störtsköna på lab. **Rockis, John, Anna, Camilla, Nilve, Magnus, Paul, Inger, Kristina, Lotta, Martina, Björn, Holger, Marica, Karin, Mattias** ... det är för många att tacka, så ni övriga på Albanova och Scilife får nöja er med ett grupptack – ni vet vilka ni är – tack!

Alla ovärderliga personer som hjälpt mig sekvensa: **Bahram**, **Kicki**, **Christian**, **Anna, Emma**, **Jimmie**, **Joel**, **Tiina**, **Elsie**, **Miriam, Desirée**, **Sofie** och **Mattias.**

Alla data-folk som gjort en massa vettiga saker med sekvenserna innan, efter eller bredvid som de kommit till mig! **Per**, **Roman**, **Mikael**, **Patrik, Ellen, Pontus, Björn** – som även handledde dataanalysen på TES-peket och alltid kommer med allmänt briljant input på det mesta! **Fredrik och Eric** också för all IT support! Ni är alla grymma och utan er stannar stället.

LingVitae folksen: **Valtteri**, **Andreas**, **Jonas**, **Terje, Preben** för kontakten med verkligheten (alltså det där abstrakta utanför akademien, med många fina middagar och roliga diskussioner).

Tack projektarbetare jag handlett: **Joel**, **Hugo** och **David**. Jag var varit lyckligt unnad att får arbeta med er. Jag tittar fortfarande i din exemplariska labbok då och då Joel, ITS-sekvensering verkar bli en hit tack vare dig Hugo, och snart har David nog löst det här problemet med polyG.

Tack även familj och vänner som gjort resan möjlig! **Mamma**, **Pappa**, **Farmor** och **Susanna**. Tack **Lina, Erik, Matilda, Totte, Magnus, Maria, Oscar, Janie, Jonas, Karin, Arthur, Gustav, Cicci, Victor, Malin, Kina, Therese, Henning** och alla andra som jag (oftast) inte pratar forskning med. Och såklart – tack **Evelina** och **Jimmy**!

Fred, kärlek och vegetarisk mat till er allihop!

Sverker
Stockholm, 2012

References

1. Aristotle. (Retrieved 29 Sep 2012) The History of Animals. *University of Adelaide*, http://ebooks.adelaide.edu.au/a/aristotle/history/index.html.
2. Drew, G.A. (1911) Sexual activities of the squid, Loligo pealii (Les.) I. Copulation, egg-laying and fertilization. *Journal of Morphology*, **22**, 327-359.
3. Iep.utm.edu. (Retrieved 29 Sep 2012) Aristotle: Biology. *Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/aris-bio/*.
4. Aristotle. (Retrieved 29 Sep 2012) On the Generation of Animals. *University of Adelaide*, http://ebooks.adelaide.edu.au/a/aristotle/generation/index.html.
5. Hooke, R. (1665) *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses.*
6. Yadav, S.P. (2007) The wholeness in suffix -omics, -omes, and the word om. *J Biomol Tech*, **18**, 277.
7. Oed.com. (Retrieved 2 Oct 2012) gene, n.2. *OED Online. Oxford University Press. http://www.oed.com/view/Entry/63127?redirectedFrom=enzyme#eid.*
8. Mendel, G. (1865) Versuche über Pflanzenhybriden.
9. Waldeyer, W. (1889) Karyokinesis and its Relation to the Process of Fertilization. *Quarterly Journal of Microscopical Science*, **s2-s30**, 215-281.
10. Choudhuri, S. (2003) The Path from Nuclein to Human Genome: A Brief History of DNA with a Note on Human Genome Sequencing and Its Impact on Future Research in Biology. *bull sci technol soc*, **23**, 360-367.
11. WATSON, J.D. and CRICK, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737-738.
12. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
13. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.
14. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physiology or Medicine 1910. *http://www.nobelprize.org/nobel_prizes/medicine/laureates/1910/*.
15. Levene, P.A. (1919) THE STRUCTURE OF YEAST NUCLEIC ACID: IV. AMMONIA HYDROLYSIS. *J. Biol. Chem.* , **40**, 415-424.
16. Griffith, F. (1928) The Significance of Pneumococcal Types. *J Hyg (Lond)*, **27**, 113-159.
17. Avery, O.T., Macleod, C.M. and McCarty, M. (1944) Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med*, **79**, 137-158.
18. Chargaff, E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **6**, 201-209.
19. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Chemistry 1954. *http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1954/*.
20. Nobelprize.org. (29 Sep 2012) The Nobel Peace Prize 1962. *http://www.nobelprize.org/nobel_prizes/peace/laureates/1962/*.
21. Pauling, L. (1929) THE PRINCIPLES DETERMINING THE STRUCTURE OF COMPLEX IONIC CRYSTALS. *Journal of the American Chemical Society*, **51**, 1010-1026.
22. Pauling, L., Corey, R.B. and Branson, H.R. (1951) The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, **37**, 205-211.
23. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physics 1915. *http://www.nobelprize.org/nobel_prizes/physics/laureates/1915/*.
24. Pauling, L. and Corey, R.B. (1953) A Proposed Structure For The Nucleic Acids. *Proc Natl Acad Sci U S A*, **39**, 84-97.
25. osulibrary.orst.edu. (Retrieved 29 Sep 2012) Linus Pauling and the Race for DNA – A document in history. *http://osulibrary.orst.edu/specialcollections/coll/pauling/dna/index.html.*
26. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physiology or Medicine 1962. *http://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/*.
27. Nirenberg, M.W. and Matthaei, J.H. (1961) The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proc Natl Acad Sci U S A*, **47**, 1588-1602.
28. Crick, F.H., Barnett, L., Brenner, S. and Watts-Tobin, R.J. (1961) General nature of the genetic code for proteins. *Nature*, **192**, 1227-1232.
29. Nirenberg, M. and Leder, P. (1964) Rna Codewords and Protein Synthesis. The Effect of Trinucleotides Upon the Binding of Srna to Ribosomes. *Science*, **145**, 1399-1407.
30. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physiology or Medicine 1968. *http://www.nobelprize.org/nobel_prizes/medicine/laureates/1968/*.
31. Holley, R.W., Apgar, J., Everett, G.A., Madison, J.T., Marquisee, M., Merrill, S.H., Penswick, J.R. and Zamir, A. (1965) Structure of a Ribonucleic Acid. *Science*, **147**, 1462-1465.
32. Ralph, R.K., Smith, R.A. and Khorana, H.G. (1962) Studies on polynucleotides. XV. Enzymic degradation. The mode of action of pancreatic deoxyribonuclease on thymidine, deoxycytidine, and deoxyadenosine polynucleotides. *Biochemistry*, **1**, 131-137.
33. Schrödinger, E. (1944) What is Life? *http://whatislife.stanford.edu/LoCo_files/What-is-Life.pdf*, Retrieved 29 Sep 2012.
34. Oed.com. (Retrieved 2 Oct 2012) enzyme, n. *OED Online. Oxford University Press. http://www.oed.com/view/Entry/63127?redirectedFrom=enzyme#eid.*
35. Buchner, E. (1907) Cell-free fermentation. *http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1907/buchner-lecture.pdf*, Retrieved 13 Sep 2012.
36. Gellert, M. (1967) Formation of covalent circles of lambda DNA by E. coli extracts. *Proc Natl Acad Sci U S A*, **57**, 148-155.

37. Weiss, B. and Richardson, C.C. (1967) Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proc Natl Acad Sci U S A*, **57**, 1021-1028.

38. Olivera, B.M. and Lehman, I.R. (1967) Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. *Proc Natl Acad Sci U S A*, **57**, 1426-1433.

39. Little, J.W., Zimmerman, S.B., Oshinsky, C.K. and Gellert, M. (1967) Enzymatic joining of DNA strands, II. An enzyme-adenylate intermediate in the dpn-dependent DNA ligase reaction. *Proc Natl Acad Sci U S A*, **58**, 2004-2011.

40. Gefter, M.L., Becker, A. and Hurwitz, J. (1967) The enzymatic repair of DNA. I. Formation of circular lambda-DNA. *Proc Natl Acad Sci U S A*, **58**, 240-247.

41. Wilkinson, A., Day, J. and Bowater, R. (2001) Bacterial DNA ligases. *Mol Microbiol*, **40**, 1241-1248.

42. Neb.com. (Retrieved 29 Sep 2012) New England Biolabs. *www.neb.com.*

43. Epibio.com. (Retrieved 29 Sep 2012) Epicentre. *http://www.epibio.com/.*

44. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physiology or Medicine 1978. *http://www.nobelprize.org/nobel_prizes/medicine/laureates/1978/.*

45. Arber, W. (1965) Host-controlled modification of bacteriophage. *Annu Rev Microbiol*, **19**, 365-378.

46. Gupta, R., Capalash, N. and Sharma, P. (2012) Restriction endonucleases: natural and directed evolution. *Appl Microbiol Biotechnol*, **94**, 583-599.

47. Smith, H.O. and Wilcox, K.W. (1970) A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J Mol Biol*, **51**, 379-391.

48. Kelly, T.J., Jr. and Smith, H.O. (1970) A restriction enzyme from Hemophilus influenzae. II. *J Mol Biol*, **51**, 393-409.

49. Danna, K. and Nathans, D. (1971) Specific cleavage of simian virus 40 DNA by restriction endonuclease of Hemophilus influenzae. *Proc Natl Acad Sci U S A*, **68**, 2913-2917.

50. Roberts, R.J. (2005) How restriction enzymes became the workhorses of molecular biology. *Proc Natl Acad Sci U S A*, **102**, 5905-5908.

51. Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. (2005) REBASE--restriction enzymes and DNA methyltransferases. *Nucleic Acids Res*, **33**, D230-232.

52. Roberts, R.J., Vincze, T., Posfai, J. and Macelis, D. (2010) REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res*, **38**, D234-236.

53. Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S., Dryden, D.T., Dybvig, K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res*, **31**, 1805-1812.

54. Kornberg, A., Lehman, I.R., Bessman, M.J. and Simms, E.S. (1956) Enzymic synthesis of deoxyribonucleic acid. *Biochim Biophys Acta*, **21**, 197-198.

55. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Physiology or Medicine 1959. *http://www.nobelprize.org/nobel_prizes/medicine/laureates/1959/.*

56. Grunberg-Manago, M., Oritz, P.J. and Ochoa, S. (1955) Enzymatic synthesis of nucleic acidlike polynucleotides. *Science*, **122**, 907-910.

57. Lehman, I.R. (2003) Discovery of DNA polymerase. *J Biol Chem*, **278**, 34733-34738.

58. Friedberg, E.C. (2006) The eureka enzyme: the discovery of DNA polymerase. *Nat Rev Mol Cell Biol*, **7**, 143-147.

59. Blanco, L. and Salas, M. (1984) Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proc Natl Acad Sci U S A*, **81**, 5325-5329.

60. Blanco, L., Bernad, A., Lazaro, J.M., Martin, G., Garmendia, C. and Salas, M. (1989) Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem*, **264**, 8935-8940.

61. Binga, E.K., Lasken, R.S. and Neufeld, J.D. (2008) Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *ISME J*, **2**, 233-241.

62. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. and Erlich, H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487-491.

63. Cadwell, R.C. and Joyce, G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl*, **2**, 28-33.

64. Holland, P.M., Abramson, R.D., Watson, R. and Gelfand, D.H. (1991) Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase. *Proc Natl Acad Sci U S A*, **88**, 7276-7280.

65. Neb.com. (Retrieved 29 Sep 2012) Properties of DNA Polymerases. *http://www.neb.com/nebecomm/tech_reference/polymerases/properties_dna_polymerases.asp#.UHbe8BzEUZY.*

66. Langhorst, B.W., Jack, W.E., Reha-Krantz, L. and Nichols, N.M. (2012) Polbase: a repository of biochemical, genetic and structural information about DNA polymerases. *Nucleic Acids Res*, **40**, D381-387.

67. Kapabiosystems.com. (Retrieved 29 Sep 2012) High Fidelity PCR. *http://www.kapabiosystems.com/products/category/high-fidelity-pcr).* .

68. Clarke, J., Wu, H.C., Jayasinghe, L., Patel, A., Reid, S. and Bayley, H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*, **4**, 265-270.

69. Lehman, I.R. (1960) The deoxyribonucleases of Escherichia coli. I. Purification and properties of a phosphodiesterase. *J Biol Chem*, **235**, 1479-1487.

70. Lehman, I.R. and Nussbaum, A.L. (1964) The Deoxyribonucleases of Escherichia Coli. V. On the Specificity of Exonuclease I (Phosphodiesterase). *J Biol Chem*, **239**, 2628-2636.

71. Richardson, C.C. and Kornberg, A. (1964) A Deoxyribonucleic Acid Phosphatase-Exonuclease from Escherichia Coli. I. Purification of the Enzyme and Characterization of the Phosphatase Activity. *J Biol Chem*, **239**, 242-250.

72. Neb.com. (Retrieved 29 Sep 2012) Properties of Exonucleases and Endonucleases. *http://www.neb.com/nebecomm/tech_reference/modifying_enzymes/prop_exonucleases_endonucleases.asp#.UGcHqhyOlS M.*

73. Baltimore, D. (1970) RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, **226**, 1209-1211.

74. Temin, H.M. and Mizutani, S. (1970) RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, **226**, 1211-1213.

75. Verma, I.M., Temple, G.F., Fan, H. and Baltimore, D. (1972) In vitro synthesis of DNA complementary to rabbit reticulocyte 10S RNA. *Nat New Biol*, **235**, 163-167.

76. Kacian, D.L., Spiegelman, S., Bank, A., Terada, M., Metafora, S., Dow, L. and Marks, P.A. (1972) In vitro synthesis of DNA components of human genes for globins. *Nat New Biol*, **235**, 167-169.

77. Sternberg, N. and Hamilton, D. (1981) Bacteriophage P1 site-specific recombination. I. Recombination between loxP sites. *J Mol Biol*, **150**, 467-486.

78. Goryshin, I.Y. and Reznikoff, W.S. (1998) Tn5 in vitro transposition. *J Biol Chem*, **273**, 7367-7374.

79. Goryshin, I.Y., Miller, J.A., Kil, Y.V., Lanzov, V.A. and Reznikoff, W.S. (1998) Tn5/IS50 target recognition. *Proc Natl Acad Sci USA*, **95**, 10716.

80. Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X. *et al.* (2010) Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol*, **11**, R119.

81. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Chemistry 1948. *http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1948/*.

82. Tiselius, A. (1937) A new apparatus for electrophoretic analysis of colloidal mixtures. *Trans. Faraday Soc.*, **33**, 524.

83. Markham, R. and Smith, J.D. (1952) The structure of ribonucleic acid. I. Cyclic nucleotides produced by ribonuclease and by alkaline hydrolysis. *Biochem J*, **52**, 552-557.

84. Smithies, O. (1955) Zone electrophoresis in starch gels: group variations in the serum proteins of normal human adults. *Biochem J*, **61**, 629-641.

85. Aaij, C. and Borst, P. (1972) The gel electrophoresis of DNA. *Biochim Biophys Acta*, **269**, 192-200.

86. Hayward, G.S. and Smith, M.G. (1972) The chromosome of bacteriophage T5. I. Analysis of the single-stranded DNA fragments by agarose gel electrophoresis. *J Mol Biol*, **63**, 383-395.

87. Schwartz, D.C. and Cantor, C.R. (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell*, **37**, 67-75.

88. Hjerten, S. (1967) Free zone electrophoresis. *Chromatogr Rev*, **9**, 122-219.

89. Jorgenson, J.W. and Lukacs, K.D. (1981) Free-zone electrophoresis in glass capillaries. *Clin Chem*, **27**, 1551-1553.

90. Kasper, T.J., Melera, M., Gozel, P. and Brownlee, R.G. (1988) Separation and detection of DNA by capillary electrophoresis. *J Chromatogr*, **458**, 303-312.

91. Swerdlow, H. and Gesteland, R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res*, **18**, 1415-1419.

92. Swerdlow, H., Wu, S.L., Harke, H. and Dovichi, N.J. (1990) Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr*, **516**, 61-67.

93. Ryle, A.P., Sanger, F., Smith, L.F. and Kitai, R. (1955) The disulphide bonds of insulin. *Biochem J*, **60**, 541-556.

94. Sanger, F. (1988) Sequences, sequences, and sequences. *Annu Rev Biochem*, **57**, 1-28.

95. Wu, R. and Taylor, E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J Mol Biol*, **57**, 491-511.

96. Wu, R. and Kaiser, A.D. (1968) Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J Mol Biol*, **35**, 523-537.

97. Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, **94**, 441-448.

98. Hutchison, C.A., 3rd. (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*, **35**, 6227-6237.

99. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, **74**, 560-564.

100. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, 5463-5467.

101. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F. and Petersen, G.B. (1982) Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol*, **162**, 729-773.

102. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496-512.

103. Sinsheimer, R.L. (2006) To reveal the genomes. *Am J Hum Genet*, **79**, 194-196.

104. Lander, E.S. (2011) Initial impact of the sequencing of the human genome. *Nature*, **470**, 187-197.

105. Kleppe, K., Ohtsuka, E., Kleppe, R., Molineux, I. and Khorana, H.G. (1971) Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. *J Mol Biol*, **56**, 341-361.

106. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G. and Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol*, **51 Pt 1**, 263-273.

107. Nobelprize.org. (Retrieved 29 Sep 2012) The Nobel Prize in Chemistry 1993. *http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1993/*.

108. Smith, L.M., Fung, S., Hunkapiller, M.W., Hunkapiller, T.J. and Hood, L.E. (1985) The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res*, **13**, 2399-2412.

109. Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B. and Hood, L.E. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature*, **321**, 674-679.

110. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457-465.

111. Cohen, S.N., Chang, A.C. and Hsu, L. (1972) Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. *Proc Natl Acad Sci U S A*, **69**, 2110-2114.

112. Cohen, S.N., Chang, A.C., Boyer, H.W. and Helling, R.B. (1973) Construction of biologically functional bacterial plasmids in vitro. *Proc Natl Acad Sci U S A*, **70**, 3240-3244.

113. Collins, J. and Hohn, B. (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci U S A*, **75**, 4242-4246.

114. Kim, U.J., Shizuya, H., de Jong, P.J., Birren, B. and Simon, M.I. (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res*, **20**, 1083-1085.

115. Murray, A.W. and Szostak, J.W. (1983) Construction of artificial chromosomes in yeast. *Nature*, **305**, 189-193.

116. Burke, D.T., Carle, G.F. and Olson, M.V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science*, **236**, 806-812.

117. O'Connor, M., Peifer, M. and Bender, W. (1989) Construction of large DNA segments in Escherichia coli. *Science*, **244**, 1307-1312.

118. Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. *Proc Natl Acad Sci U S A*, **89**, 8794-8797.

119. genomics.energy.gov. (Retrieved 29 Sep 2012) Major Events in the U.S. Human Genome Project and Related Projects. *http://www.ornl.gov/sci/techresources/Human_Genome/project/timeline.shtml.*

120. Venter, J.C., Adams, M.D., Sutton, G.G., Kerlavage, A.R., Smith, H.O. and Hunkapiller, M. (1998) Shotgun Sequencing of the Human Genome. *Science*, **280**, 1540-1542.

121. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol*, **26**, 1135-1145.

122. Ronaghi, M., Uhl√©n, M. and Nyr√©n, P.l. (1998) A Sequencing Method Based on Real-Time Pyrophosphate. *Science*, **281**, 363-365.

123. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

124. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D. and Church, G.M. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, **309**, 1728-1732.

125. Mitra, R.D. and Church, G.M. (1999) In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Res*, **27**, e34.

126. Mitra, R.D., Shendure, J., Olejnik, J., Edyta Krzymanska, O. and Church, G.M. (2003) Fluorescent in situ sequencing on polymerase colonies. *Anal Biochem*, **320**, 55-65.

127. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*, **242**, 84-89.

128. Tawfik, D.S. and Griffiths, A.D. (1998) Man-made cell-like compartments for molecular evolution. *Nat Biotechnol*, **16**, 652-656.

129. Dressman, D., Yan, H., Traverso, G., Kinzler, K.W. and Vogelstein, B. (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A*, **100**, 8817-8822.

130. Ghadessy, F.J., Ong, J.L. and Holliger, P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc Natl Acad Sci U S A*, **98**, 4552-4557.

131. Leamon, J.H., Lee, W.L., Tartaro, K.R., Lanza, J.R., Sarkis, G.J., deWinter, A.D., Berka, J., Weiner, M., Rothberg, J.M. and Lohman, K.L. (2003) A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, **24**, 3769-3777.

132. 454.com. (Retrieved 29 Sep 2012) GS FLX+ System – Now delivering sequencing reads up to 1,000 bp in length! *http://454.com/products/gs-flx-system/index.asp.*

133. Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53-59.

134. Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P. and Kawashima, E. (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res*, **28**, E87.

135. Ju, J., Kim, D.H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M.S., Shi, S., Wu, J. *et al.* (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proc Natl Acad Sci U S A*, **103**, 19635-19640.

136. Illumina.com. (Retrieved 29 Sep 2012) HiSeq 2500 Sequencing System. *http://www.illumina.com/Documents/%5Cproducts%5Cappnotes%5Cappnote_hiseq2500.pdf.*

137. Illumina.com. (Retrieved 29 Sep 2012) MiSeq System. *http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf.*

138. Cloonan, N., Forrest, A.R., Kolle, G., Gardiner, B.B., Faulkner, G.J., Brown, M.K., Taylor, D.F., Steptoe, A.L., Wani, S., Bethel, G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, **5**, 613-619.

139. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, **19**, 1527-1541.

140. Lifetechnologies.com. (Retrieved 29 Sep 2012) A Theoretical Understanding of 2 Base Color Codes and Its Application to Annotation, Error Detection, and Error Correction *http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_058265.pdf.*

141. Lifetechnologies.com. (Retrieved 29 Sep 2012) 5500 W Series Genetic Analyzers – Wildfire technology provides radical improvements in simplicity and higher throughput. *https://tools.invitrogen.com/content/sfs/brochures/CO111759_5500_W_prelim_spec_sheet_FLR.pdf.*

142. Glenn, T.C. (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, **11**, 759-769.

143. Harris, T.D., Buzby, P.R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J.W. *et al.* (2008) Single-molecule DNA sequencing of a viral genome. *Science*, **320**, 106-109.
144. Orlando, L., Ginolhac, A., Raghavan, M., Vilstrup, J., Rasmussen, M., Magnussen, K., Steinmann, K.E., Kapranov, P., Thompson, J.F., Zazula, G. *et al.* (2011) True single-molecule DNA sequencing of a pleistocene horse bone. *Genome Res*, **21**, 1705-1719.
145. Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M. and Milos, P.M. (2009) Direct RNA sequencing. *Nature*, **461**, 814-818.
146. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138.
147. Pacificbiosciences.com. (Retrieved 29 Sep 2012) SMRT cells. *http://www.pacificbiosciences.com/products/consumables/SMRT-cells/*.
148. Pacificbiosciences.com. (Retrieved 29 Sep 2012) Pacbio RS. *http://www.pacificbiosciences.com/brochure*.
149. Song, C.X., Clark, T.A., Lu, X.Y., Kislyuk, A., Dai, Q., Turner, S.W., He, C. and Korlach, J. (2012) Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine. *Nat Methods*, **9**, 75-77.
150. Karow, J. (25 Sep 2012) New Sample Loading Device for PacBio Improves Results, Users Say. *http://www.genomeweb.com//node/1130726?hq_e=el&hq_m=1356034&hq_l=4&hq_v=34d6766888*, Retrieved 29 Sep 2012.
151. Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M. *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348-352.
152. Pourmand, N., Karhanek, M., Persson, H.H., Webb, C.D., Lee, T.H., Zahradnikova, A. and Davis, R.W. (2006) Direct electrical detection of DNA synthesis. *Proc Natl Acad Sci U S A*, **103**, 6466-6470.
153. Lifetechnologies.com. (6 Sep 2012) Longer Reads Better Assemblies. *http://ir.lifetechnologies.com/releasedetail.cfm?ReleaseID=704985*, Retrieved 29 Sep 2012.
154. Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.
155. Lifetechnologies.com. (13 Sep 2012) Life Technologies Begins Shipping Ion Proton System. *http://ir.lifetechnologies.com/releasedetail.cfm?releaseid=706924*, Retrieved 29 Sep 2012.
156. Lifetechnologies.com. (Retrieved 29 Sep 2012) The Ion Proton System. *http://tools.invitrogen.com/content/sfs/brochures/CO111809_Specification%20Sheet_Ion%20Proton%20System_0712.pdf*.
157. Heger, M. (20 Dec 2012) New Paired-End Protocol for Ion PGM Increases Accuracy when Analyzed with SoftGenetics' Algorithm. *http://www.genomeweb.com/sequencing/new-paired-end-protocol-ion-pgm-increases-accuracy-when-analyzed-softgenetics-al*, Retrieved 29 Sep 2012.
158. Heger, M. (18 Sep 2012) Life Tech Begins Proton Shipments; Announces Proton 3 and emPCR-Free Sample Prep. *http://www.genomeweb.com/sequencing/life-tech-begins-proton-shipments-announces-proton-3-and-empcr-free-sample-prep*, Retrieved 29 Sep 2012.
159. Drmanac, R., Sparks, A., Callow, M., Halpern, A., Burns, N., Kermani, B., Carnevali, P., Nazarenko, I., Nilsen, G., Yeung, G. *et al.* (2009) Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*.
160. Genomeweb.com. (17 Sep 2012) Complete Genomics, BGI Agree to $117.6M Merger. *http://www.genomeweb.com/sequencing/complete-genomics-bgi-agree-1176m-merger*, Retrieved 29 Sep 2012.
161. Lam, H.Y., Clark, M.J., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., Butte, A.J. *et al.* (2012) Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol*, **30**, 78-82.
162. Venkatesan, B.M. and Bashir, R. (2011) Nanopore sensors for nucleic acid analysis. *Nat Nanotechnol*, **6**, 615-624.
163. Karow, J. (21 Feb 2012) AGBT: Oxford Nanopore to Begin Selling Two Low-Cost DNA Strand Sequencing Instruments this Year. *http://www.genomeweb.com/sequencing/agbt-oxford-nanopore-begin-selling-two-low-cost-dna-strand-sequencing-instrument*, Retrieved 29 Sep 2012.
164. Nanoporetech.com. (Retrieved 29 Sep 2012) Oxford Nanopore introduces DNA 'strand sequencing' on the high-throughput GridION platform and presents MinION, a sequencer the size of a USB memory stick. *http://www.nanoporetech.com/news/press-releases/view/39*.
165. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.
166. Metzker, M.L. (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31-46.
167. Williams, L.J., Tabbaa, D.G., Li, N., Berlin, A.M., Shea, T.P., Maccallum, I., Lawrence, M.S., Drier, Y., Getz, G., Young, S.K. *et al.* (2012) Paired-end sequencing of Fosmid libraries by Illumina. *Genome Res*.
168. Stahl, P.L. and Lundeberg, J. (2012) Toward the single-hour high-quality genome. *Annu Rev Biochem*, **81**, 359-378.
169. Karow, J. (21 Feb 2012) At AGBT, Illumina Shows Data for HiSeq 2500 ‘Genome in a Day,’ Outlines 400-Base PE Reads for MiSeq. *http://www.genomeweb.com/sequencing/agbt-illumina-shows-data-hiseq-2500-‘genome-day’-outlines-400-base-pe-reads-mise*, Retrieved 29 Sep 2012.
170. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, **20**, 265-272.
171. Alkan, C., Sajjadian, S. and Eichler, E.E. (2011) Limitations of next-generation genome sequence assembly. *Nat Methods*, **8**, 61-65.
172. Lifetechnologies.com. (Retrieved 29 Sep 2012) SOLiD System accuracy with the Exact Call Chemistry module *http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf*.
173. Massingham, T. and Goldman, N. (2012) Error-correcting properties of the SOLiD Exact Call Chemistry. *BMC Bioinformatics*, **13**, 145.
174. Koren, S., Schatz, M.C., Walenz, B.P., Martin, J., Howard, J.T., Ganapathy, G., Wang, Z., Rasko, D.A., McCombie, W.R., Jarvis, E.D. *et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, **30**, 693-700.

175. Hiatt, J.B., Patwardhan, R.P., Turner, E.H., Lee, C. and Shendure, J. (2010) Parallel, tag-directed assembly of locally derived short sequence reads. *Nat Methods*, **7**, 119-122.

176. Sorber, K., Chiu, C., Webster, D., Dimon, M., Ruby, J.G., Hekele, A. and DeRisi, J.L. (2008) The long march: a sample preparation technique that enhances contig length and coverage by high-throughput short-read sequencing. *PLoS ONE*, **3**, e3495.

177. Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M. and Turner, D.J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, **6**, 291-295.

178. En.wikipedia.org. (Retrieved 2 Oct 2012) Library (biology). *http://en.wikipedia.org/wiki/Library_(biology).*

179. Barnes, W.M. (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. *Gene*, **112**, 29-35.

180. Neb.com. (Retrieved 29 Sep 2012) Thermophilic DNA Polymerases. *http://www.neb.com/nebecomm/tech_reference/polymerases/thermophilic_dna_polymerase.asp#.UGAaJRwv134),* .

181. Quail, M.A., Otto, T.D., Gu, Y., Harris, S.R., Skelly, T.F., McQuillan, J.A., Swerdlow, H.P. and Oyola, S.O. (2012) Optimal enzymes for amplifying sequencing libraries. *Nat Methods*, **9**, 10-11.

182. Raghunathan, A., Ferguson, H.R., Jr., Bornarth, C.J., Song, W., Driscoll, M. and Lasken, R.S. (2005) Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol*, **71**, 3342-3347.

183. Dean, F.B., Nelson, J.R., Giesler, T.L. and Lasken, R.S. (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res*, **11**, 1095-1099.

184. Lasken, R.S. (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. *Curr Opin Microbiol*, **10**, 510-516.

185. Wang, J., Fan, H.C., Behr, B. and Quake, S.R. (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, **150**, 402-412.

186. Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**, 886-895.

187. Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D. *et al.* (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, **148**, 873-885.

188. Linnarsson, S. (2010) Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res*, **316**, 1339-1343.

189. Parkinson, N.J., Maslau, S., Ferneyhough, B., Zhang, G., Gregory, L., Buck, D., Ragoussis, J., Ponting, C.P. and Fischer, M.D. (2012) Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res*, **22**, 125-133.

190. Meyer, M., Stenzel, U. and Hofreiter, M. (2008) Parallel tagged sequencing on the 454 platform. *Nat Protoc*, **3**, 267-278.

191. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J. and Knight, R. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods*, **5**, 235-237.

192. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. *Nat Methods*, **7**, 111-118.

193. Faircloth, B.C. and Glenn, T.C. (2012) Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*, **7**, e42543.

194. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W. and Vogelstein, B. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A*, **108**, 9530-9535.

195. Casbon, J.A., Osborne, R.J., Brenner, S. and Lichtenstein, C.P. (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res*, **39**, e81.

196. Fu, G.K., Hu, J., Wang, P.H. and Fodor, S.P. (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci U S A*, **108**, 9026-9031.

197. Kivioja, T., Vaharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, **9**, 72-74.

198. Mertes, F., Elsharawy, A., Sauer, S., van Helvoort, J.M., van der Zaag, P.J., Franke, A., Nilsson, M., Lehrach, H. and Brookes, A.J. (2011) Targeted enrichment of genomic DNA regions for next-generation sequencing. *Brief Funct Genomics*, **10**, 374-386.

199. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods*, **4**, 903-905.

200. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, **27**, 182-189.

201. Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. and Davis, R.W. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, **16**, 301-306.

202. Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Euskirchen, G., Butte, A.J. and Snyder, M. (2011) Performance comparison of exome DNA sequencing technologies. *Nat Biotechnol*, **29**, 908-914.

203. Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F. *et al.* (2007) Multiplex amplification of large sets of human exons. *Nat Methods*, **4**, 931-936.

204. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. and Nilsson, M. (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res*, **33**, e71.

205. Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W.F., Davis, R.W. and Ji, H. (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc Natl Acad Sci U S A*, **104**, 9387-9392.

206. Broude, N.E., Zhang, L., Woodward, K., Englert, D. and Cantor, C.R. (2001) Multiplex allele-specific target amplification based on PCR suppression. *Proc Natl Acad Sci U S A*, **98**, 206-211.

207. Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., David, P.H., Kotsopoulos, S.K., Samuels, M.L., Hutchison, J.B., Larson, J.W. *et al.* (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol*, **27**, 1025-1031.

208. Raindancetech.com. (Retrieved 29 Sep 2012) ThunderStorm System. *http://raindancetech.com/targeted-dna-sequencing/thunderstorm/*.

209. Illumina.com. (Retrieved 29 Sep 2012) TruSeq Custom Amplicon. *http://www.illumina.com/products/truseq_custom_amplicon.ilmn.*

210. Lifetechnologies.com. (Retrieved 29 Sep 2012) Ion AmpliSeq Designer and Ion AmpliSeq Custom Panels *http://tools.invitrogen.com/content/sfs/brochures/Ion_AmpliSeq_Designer_CustomPanels_ProductBulletin_CO24461_March2012.pdf.*

211. Melzak, K.A., Sherwood, C.S., Turner, R.F.B. and Haynes, C.A. (1996) Driving Forces for DNA Adsorption to Silica in Perchlorate Solutions. *Journal of Colloid and Interface Science*, **181**, 635-644.

212. DeAngelis, M.M., Wang, D.G. and Hawkins, T.L. (1995) Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Res*, **23**, 4742-4743.

213. Mellmann, A., Harmsen, D., Cummings, C.A., Zentz, E.B., Leopold, S.R., Rico, A., Prior, K., Szczepanowski, R., Ji, Y., Zhang, W. *et al.* (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One*, **6**, e22751.

214. Quail, M.A., Kozarewa, I., Smith, F., Scally, A., Stephens, P.J., Durbin, R., Swerdlow, H. and Turner, D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods*, **5**, 1005-1010.

215. Vogelstein, B. and Kinzler, K.W. (1999) Digital PCR. *Proc Natl Acad Sci U S A*, **96**, 9236-9241.

216. White, R.A., 3rd, Blainey, P.C., Fan, H.C. and Quake, S.R. (2009) Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC Genomics*, **10**, 116.

217. Edwards, A. and Caskey, C.T. (1991) Closure strategies for random DNA sequencing. *Methods*, **3**, 41-47.

218. Roach, J.C., Boysen, C., Wang, K. and Hood, L. (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics*, **26**, 345-353.

219. Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872-876.

220. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420-426.

221. Peng, Z., Zhao, Z., Nath, N., Froula, J.L., Clum, A., Zhang, T., Cheng, J.F., Copeland, A.C., Pennacchio, L.A. and Chen, F. (2012) Generation of long insert pairs using a Cre-LoxP Inverse PCR approach. *PLoS One*, **7**, e29437.

222. Van Nieuwerburgh, F., Thompson, R.C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P. and Head, S.R. (2012) Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Res*, **40**, e24.

223. Song, L. and Crawford, G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, **2010**, pdb prot5384.

224. Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S. and Crawford, G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311-322.

225. Darst, R.P., Pardo, C.E., Ai, L., Brown, K.D. and Kladde, M.P. (2010) Bisulfite sequencing of DNA. *Curr Protoc Mol Biol*, **Chapter 7**, Unit 7 9 1-17.

226. Gu, H., Smith, Z.D., Bock, C., Boyle, P., Gnirke, A. and Meissner, A. (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat. Protocols*, **6**, 468-481.

227. Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J. *et al.* (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, **487**, 190-195.

228. Meyer, M., Kircher, M., Gansauge, M.T., Li, H., Racimo, F., Mallick, S., Schraiber, J.G., Jay, F., Prufer, K., de Filippo, C. *et al.* (2012) A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science*.

229. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y. *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311-317.

230. Lennon, N.J., Lintner, R.E., Anderson, S., Alvarez, P., Barry, A., Brockman, W., Daza, R., Erlich, R.L., Giannoukos, G., Green, L. *et al.* (2010) A scalable, fully automated process for construction of sequence-ready barcoded libraries for 454. *Genome Biol*, **11**, R15.

231. Genome.gov. (Retrieved 29 Sep 2012) DNA sequencing costs. *http://www.genome.gov/sequencingcosts/*.

232. Henikoff, S. (1984) Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. *Gene*, **28**, 351-359.

233. Hoheisel, J. and Pohl, F.M. (1986) Simplified preparation of unidirectional deletion clones. *Nucleic Acids Res*, **14**, 3605.

234. Hoheisel, J.D. (1993) On the activities of Escherichia coli exonuclease III. *Anal Biochem*, **209**, 238-246.