Identification of new molecular species using second-generation sequencing

Joachim De Schrijver

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

“Too few people in computer science are aware of some of the informational challenges in biology and their implications for the world.” – Sergey Brin

Promoter:	**Prof. dr. ir. Wim Van Criekinge** Lab. of Bioinformatics and Computational Genomics Dep. of Mathematical Modelling, Statistics and Bioinformatics Faculty of Bioscience Engineering Ghent University, Belgium
Co-promoter:	**Prof. dr. ir. Sofie Bekaert** Bimetra Clinical Research Center Ghent, Belgium
Dean:	**Prof. dr. ir. Guido Van Huylenbroeck**
Rector:	**Prof. dr. Paul Van Cauwenberge**

Identification of new molecular species using second-generation sequencing

ir. Joachim De Schrijver

Thesis submitted in fulfilment of the requirements
for the degree of Doctor (PhD) of Applied Biological Sciences

Dutch translation of the title
Identificatie van nieuwe moleculaire soorten gebruikmakend van tweede generatie sequeneringstechnieken

Cover illustration
The cover illustration shows a scanning electron micrograph of a portion of a GS-FLX PicoTiterPlate after loading it with beads. Each DNA-bound bead is placed into a well of the PicoTiterPlate together with a mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase which are bound on smaller beads (which are visible in the image). The PicoTiterPlate is then placed into the 454 GS-FLX system for sequencing.

The scanning electron micrograph was converted to a so-called *ascii-art* using the IMG2TXT application available on www.DeGraeve.com (by Steve De Graeve). The original image is property of 454 Life Sciences/Roche.

Reference
De Schrijver Joachim (2012), Identification of new molecular species using second-generation sequencing. PhD thesis. Ghent University.

Printing
University Press, Zelzate

ISBN
978-90-5989-575-1

Joachim De Schrijver Wim Van Criekinge

Members of the examination committee:

Prof. dr. ir. Peter Bossier
Chairman
Department of Animal Production
Faculty of Bioscience Engineering, Ghent University

Prof. dr. Danny Geelen
Secretary
Department of Plant Production
Faculty of Bioscience Engineering, Ghent University

Prof. dr. ir. Wim Van Criekinge
Promoter
Department of Mathematical Modelling, Statistics and Bioinformatics
Faculty of Bioscience Engineering, Ghent University

Prof. dr. ir. Sofie Bekaert
Co-promoter
Bimetra
Clinical Research Center Ghent

Prof. dr. ir. Olivier Thas
Department of Mathematical Modelling, Statistics and Bioinformatics
Faculty of Bioscience Engineering, Ghent University

Prof. dr. Els Van Damme
Department of Molecular Biotechnology
Faculty of Bioscience Engineering, Ghent University

Prof. dr. Peter Dawyndt
Department of Applied Mathematics and Computer Science
Faculty of Sciences, Ghent University

Prof. dr. Manon Van Engeland
GROW - School for Oncology and Developmental Biology
Maastricht University Medical Centre

Table of contents

Abbreviations

A	adenine
αHL	alpha hemolysin
5hmC	5-hydroxy-methylcytosine
5mC	5-methylcytosine
APS	adenosine-5′-phosphosulfate
aTIS	alternative translation initiation site
ATP	adenosine-5′-triphophate
BAM	binary alignment/map
BLAST	basic local alignment search tool
BLAT	BLAST-like alignment tool
bp	base pair
BWT	Burrows-Wheeler transform
C	cytosine
CAGE	cap analysis of Gene expression
cDNA	complementary DNA
CF	Cystic Fibrosis
ChIP	chromatine immunoprecipitation
CNV	copy number variation
CRC	colorectal cancer
DAC	5′-aza-2′-deoxycytidine
dATP	2′-deoxy ATP
ddATP	2′,3′-dideoxy ATP
DNMT	DNA methyltransferase
DRS	direct RNA sequencing
dsDNA	double strand DNA
DSN	duplex-specific thermostable nuclease
emPCR	emulsion PCR
EST	expressed sequence tag
G	guanine

Gb	gigabase
GRCh	Genome Reference Consortium – human
GSEA	Gene set enrichment analysis
HAT	histone acetyltransferase
HDAC	histone deacetylases
I	inosine
IHGSC	International Human Genome Sequencing Consortium
IP	immunoprecipitation
IRES	internal ribosome entry site
ISS	ion semiconductor sequencing
LDS	Loeys-Dietz Syndrome
lncRNA	long non-coding RNA
MBD	methyl-CpG-binding domain
MeDIP	methylated DNA immunoprecipitation
MFS	Marfan Syndrome
MiGS	MBD-isolated genome sequencing
miRNA	microRNA
MSP	methylation specific PCR
MLPA	multiplex ligation-dependent probe amplification
MPSS	massively parallel signature sequencing
mRNA	messenger RNA
NAT	natural antisense transcription
NCBI	National Center for Biotechnology Information
NIH	National Institute of Health
ncRNA	non-coding RNA
NGS	next-generation sequencing
ORF	open reading frame
PCR	polymerase chain reaction
PPi	pyrophosphate

PGM personal genome machine

RPKM reads per kilobase of exon model per million mapped reads

rRNA ribosomal RNA

SFF standard flowgram format

SAGE serial analysis of gene expression

SAH S-adenosyl-L-homocysteine

SAM sequence alignment/map (next-generation sequencing data type)

SAM S-adenosyl-L-methionine (methyl group donor in DNA methylation)

SMRT single molecule real time [sequencing]

SNP single nucleotide polymorphism

SNV single nucleotide variant

ssDNA single strand DNA

T thymine

tRNA transfer RNA

TSA trichostatin A

TSS transcription start site

tSMS true single molecule sequencing

U uracil

UTR untranslated region

ZMW zero-mode waveguide

Research goals & outline

Research goals

The objective of my PhD was to develop methods and use the developed methods to identify **new molecular species** in the human genome **using second-generation sequencing**. New molecular species are biological molecules (or variants thereoff) inside a cell not seen before. These molecular species can be diverse, ranging from DNA over DNA modifications and RNA to proteins. These newly identified molecular species can act as a biological marker or, more broadly, can link phenotype to genotype

Two very distinct forms of **molecular species** have been investigated and will be described extensively in this dissertation. Firstly, genomic DNA variation potentially leading to alterations of the shape and structure (and thus the function) of the expressed proteins, which can eventually result in a disease phenotype. Secondly, DNA methylation which can suppress the expression of certain genes and corresponding protein products, also potentially resulting in a disease phenotype. Both genomic variation and methylation of certain genes can cause similar pathogenic effects, whilst the genetic background is clearly different [1]. The identification of these two types of molecular species by using **second-generation sequencing** was the goal of my PhD research.

When I started my PhD in 2008, **second-generation sequencing** instruments had just become commercially available and were still considered rather experimental. The first instrument available, 454/Roche's GS-FLX machine, offered a sequencing throughput of some gigabases, at the time considered high, but now considered low. Since then, several other next-generation sequencing manufacturers have entered the market and new techniques have been developed which cannot only read DNA sequences, but also DNA methylation, RNA expression and even ribosomal activity [2]. Clearly there was, and still is, a need for tools to analyse these new types of datasets.

Using a combination of Perl, MySQL, and some R, I developed the Variant Identification Pipeline, Variant Interpretation Pipeline and SNP-CUB3. These three pipelines were developed to identify and annotate genomic variants in second-generation sequencing datasets. Subsequently, patient samples were analysed for collaborators using these pipelines.

In a second stage, I set up a broad methylation detection experiment using a HCT116/DKO model system to detect DNA methylation. Using only second-generation sequencing (and no array based methodologies) to both identify transcription and methylation, I could identify known genes and new genes under the control of methylation in colorectal cancer. Furthermore, using this genome-wide approach I was able to form a broader picture of the interplay between transcription and methylation in colorectal cancer cell lines.

Outline

This thesis consists of four different parts. **Part 1** describes the background needed to fully comprehend the following parts. Part 1 starts with a short introduction on the structure of DNA. I am fully aware that this introduction could be considered known background information by anyone familiar with genetics and/or biochemistry, however, a thorough understanding of the structure of DNA will help you understand the finesses of the different next-generation sequencing technologies which are explained afterwards. Furthermore, part 1 gives an overview of the possible biological applications of next-generation sequencing and more specifically the identification of new molecular species such as mutations and methylation patterns. Part 1 closes with an overview of the computational methods that are frequently used to analyze next-generation sequencing data.

Part 2 is a summary of different papers which describe several Perl/MySQL pipelines I developed to identify and annotate genomic variants by means of second-generation sequencing. Part 2 starts with a pipeline I developed to analyze 454 sequencing data. Recent pyrosequencing techniques - such as Ion Torrent sequencing – yield the exact same data format and can benefit from the pipeline as well. After I initially designed the pipeline using a small *BRCA1/2* sample, these pipelines were successfully applied in the analysis of several sequencing experiments to confirm known and identify new genetic variants in different diseases or disorders such as breast cancer, genetic deafness, and Marfan and Loeys-Dietz syndrome. The final section of this part describes SNP-CUB3, a pipeline I developed to identify SNPs in 3D-pooling experiments using very high throughput instruments such as the Illumina GAIIx.

Part 3 describes an extensive study wherein I linked DNA methylation to RNA expression in a colorectal cell line model. HCT116, a colorectal cancer cell line and a double knock-out (DKO) cell line lacking methylation functionalities were profiled using next-generation sequencing. Combining expression and methylation data from both HCT116 and DKO, I described the interplay between methylation and expression in detail and identified genes under control of methylation.

Part 4 is the concluding chapter and offers a general discussion and perspectives on future research.

Part 1: Discovering new molecular species

1 Introduction to genetics, epigenetics, and (epi)genetic variation

1.1 Genetics and genomics

Genetics (from the Greek *γένεσι*; genesis, meaning origin) is a discipline of biology and studies genes (and their function), heredity and variation in living organisms [3]. The human genome is composed of approximately three billion bases. Hence, human genetics is in a broad sense the study of the function, heredity and variation of those three billion bases. The field which studies complete genomes is often referred to as *genomics*, and can be considered a subfield of genetics.

1.1.1 The structure of DNA

A single strand DNA (ssDNA) molecule is a long polymer composed of four different repeating nucleotide residues bound on a phosphate-deoxyribose backbone. The four nucleotides can be divided into purines (fused five- and six-membered heterocyclic compounds) and pyrimidines (six-membered heterocyclic compounds). Adenine (A) and guanine (G) are purines; cytosine (C) and thymine (T) are pyrimidines. The phosphodiester bond between the sugar and phosphate residue is asymmetric, giving a single ssDNA molecule a specific orientation. The *head* of a DNA molecule is generally referred to as the 5′-end and the *tail* as the 3′-end. Two complementary ssDNA molecules pair together in an anti-parallel fashion (i.e. the strands run in opposite directions) and the double-helix (as described by Watson and Crick in 1953 [4]) is stabilized by hydrogen bonds between purines and pyrimidines (A-T and C-G) (Figure 1.1 A).

The double-helix DNA-strand is wound around specific proteins to form a condensed chromatin structure. This chromatin structure is then further condensed by several levels of packing to eventually form chromosomes (Figure 1.1 B). The diploid human genome is stored on twenty-two autosomal chromosome pairs and one pair sex chromosomes, which reside in the cell nucleus [5].

The specific order of the three billion nucleotides of the haploid human genome (i.e. 23 chromosomes) is referred to as the *human genome*. It is this sequence and the variation between different individuals that will be studied in this dissertation.

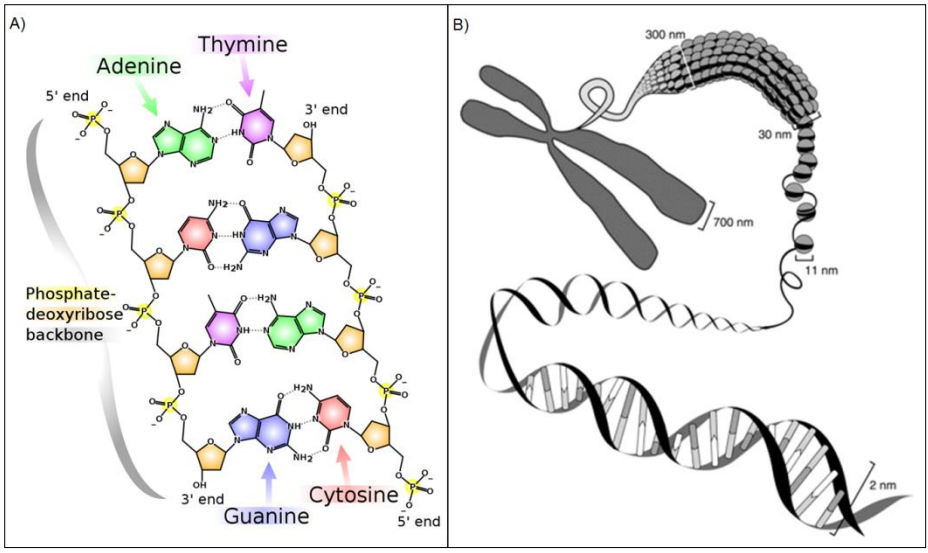


Figure 1.1: *A) Overview of the biochemical structure of the double-helix DNA-strand. The four different nucleotides (adenine, thymine, guanine, and cytosine) are bound on two anti-parallel phosphate-deoxyribose backbones. Hydrogen bonds stabilize the double-helix structure. B) Packing of DNA into chromatin and chromosomes. Several levels of packing of DNA can be discriminated. Two-nanometer-wide DNA molecules are folded into 11-nm chromatin segments which, when tightly packed with nucleosomes, appear as 30-nm chromatin fibers. Further packing leads to chromosomal sections with a diameter of about 300 nm and the 700-nm-thick chromatids of metaphase chromosomes. Source: Madeleine Price Ball (personal communication) and [5].*

1.1.2 Central dogma of molecular biology

The central dogma of molecular biology states there are, in living organisms, three classes of information-carrying molecules with a sequential nature. These three classes of molecules are DNA, RNA, and proteins. As a consequence, in principle, nine transfers could occur within and between those levels (3 x 3). However, the dogma states only three transfers occur under normal conditions [6] (Figure 1.2).

1) Information stored in a DNA nucleotide sequence can be transferred to another DNA nucleotide sequence (DNA replication)
2) Information stored in a DNA nucleotide sequence can be transferred to an RNA nucleotide sequence (transcription)

3) Eventually, the information stored in the RNA molecule can be transferred by the ribosome to a sequence of amino acids to form a protein (translation)

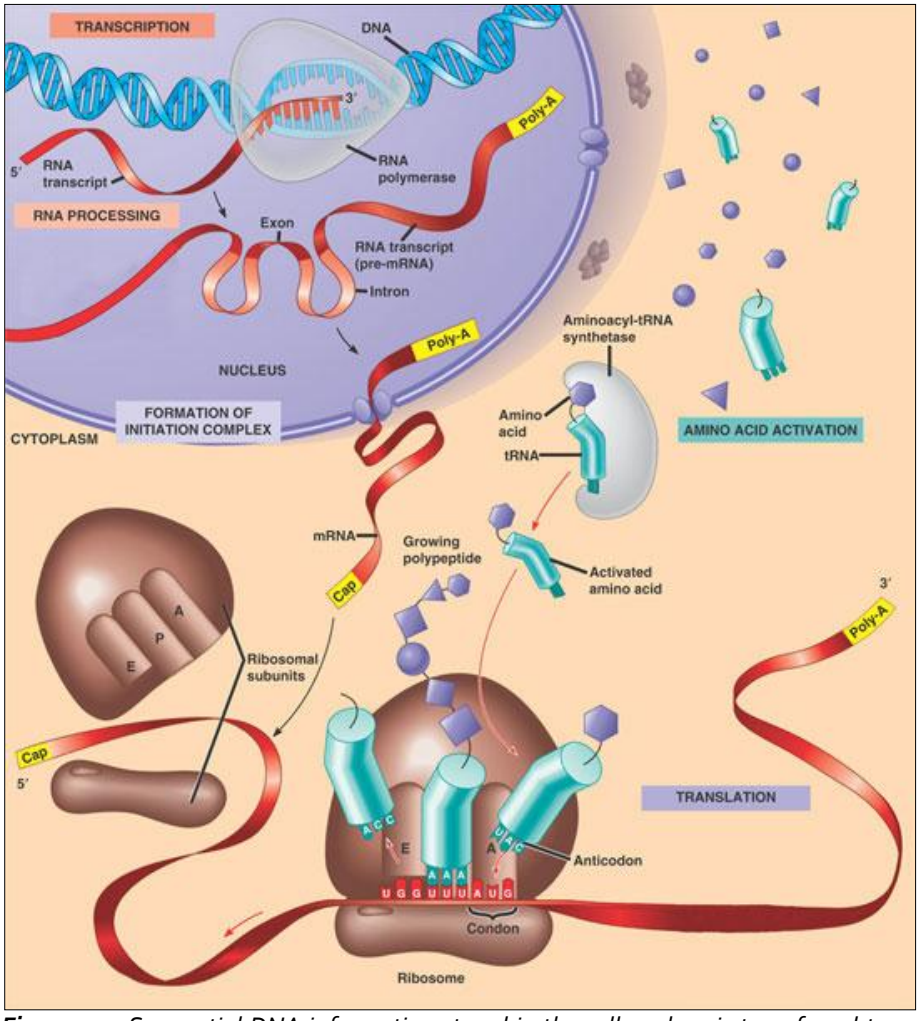


Figure 1.2: *Sequential DNA information stored in the cell nucleus is transferred to an RNA transcript by the RNA polymerase during the transcription process. This RNA molecule is further processed to become a messenger RNA (mRNA) molecule and is eventually exported to the cell cytosol. In the cytosol, mRNA is processed by the ribosome which translates the RNA molecule into a protein. The protein is the final effecter of the original information which was stored in the nucleotide sequence. Source: [7].*

1.2 Epigenetics

Epigenetics (from the Greek *επί*; above) is the study of changes in gene expression or, more broadly, the cellular phenotype, caused by modifications of the genome that do not involve a change in the actual nucleotide sequence [8]. It can be considered a level of information which is present on top of the information stored in the nucleotide sequence itself; hence the name epi-genetics, on top of genetics. The genome can be epigenetically altered by addition of certain chemical molecules to specific nucleotide residues or to the proteins around which the double-helix DNA-strand is wound, eventually altering the regulation of the gene expression without altering the underlying nucleotide sequence. A schematic overview of the two epigenetic mechanisms is given in Figure 1.3.

It is generally accepted that epigenetic information is heritably passed on to the next generation, however considerable discussion exists in the scientific literature to which extent heritability is a key feature of epigenetics [9].

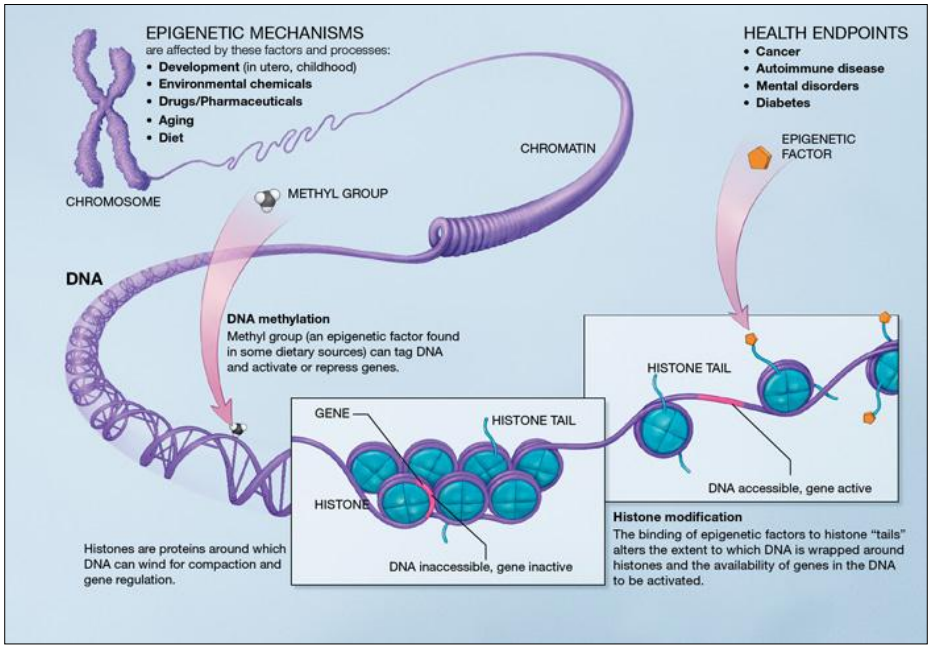


Figure 1.3: *Schematic overview of the two different epigenetic mechanisms. Both DNA methylation and histone modifications can alter the expression of biologically relevant genes. Source: The National Institute of Health (NIH) Common Fund.*

1.2.1 DNA methylation

DNA methylation is the most studied epigenetic process and involves the addition of a methyl group to a nucleotide residue, typically a cytosine. The reaction is catalyzed by a DNA methyltransferase (DNMT) enzyme which converts S-adenosyl-L-methionine (SAM) to S-adenosyl-L-homocysteine (SAH) whilst transferring a methyl group to the nucleotide residue [10]. This process is illustrated in Figure 1.4.

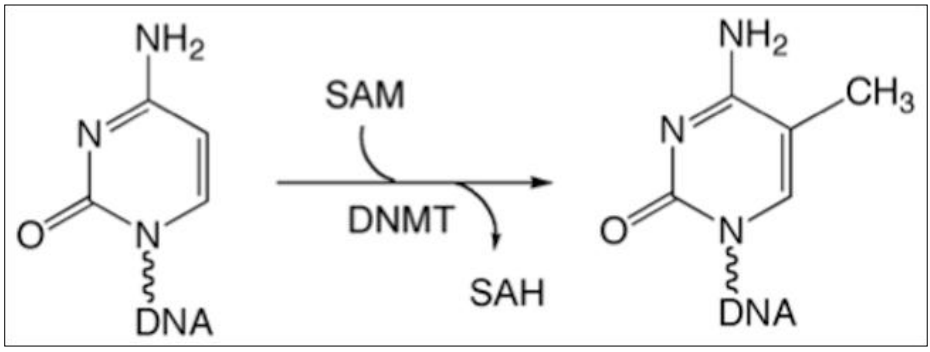


Figure 1.4: *Methylation of a cytosine residue by DNMT using SAM as a substrate.*

DNMT1 is mainly involved in maintaining DNA methylation after DNA replication and thus reacts preferably with hemi-methylated DNA. However, both DNMT3a and DNMT3b mainly catalyze *de novo* methylation and thus transfer methyl groups from SAM to previously unmethylated DNA [11].

Hydroxy-methylation, which involves the addition of a hydroxy-methyl group to a nucleotide residue, has been described as well, although its biological background and functioning is still less well understood compared to conventional DNA methylation [12].

1.2.2 Histone modifications

The double-helix DNA-strand is tightly wrapped around nucleosomes to form highly condensed chromatin. The density of this packing influences the accessibility of the transcription machinery to the DNA. The packing density determines whether the information stored in the nucleotide sequence is 'readable' or 'unreadable'.

The nucleosome is composed of an octamer of four different histones (H2A, H2B, H3, and H4), referred to as the core nucleosome, around which 147 base pairs DNA are wrapped (Figure 1.5 A). Histones undergo post-translational modifications that change their interaction with the double-helix DNA-strand and other proteins. The histones have long tails protruding from the nucleosome, which can covalently be modified at several places. These modifications include predominantly acetylation, deacetylation, methylation, demethylation, and phosphorylation. Acetylation and deacetylation are catalyzed by histone acetyltransferases (HATs) and histone deacetylases (HDACs) respectively (Figure 1.5 B) [13]. Other modification which include ubiquitination, sumoylation, ADP-ribosylation, deimination, proline isomerization, and others have been described recently [14, 15].

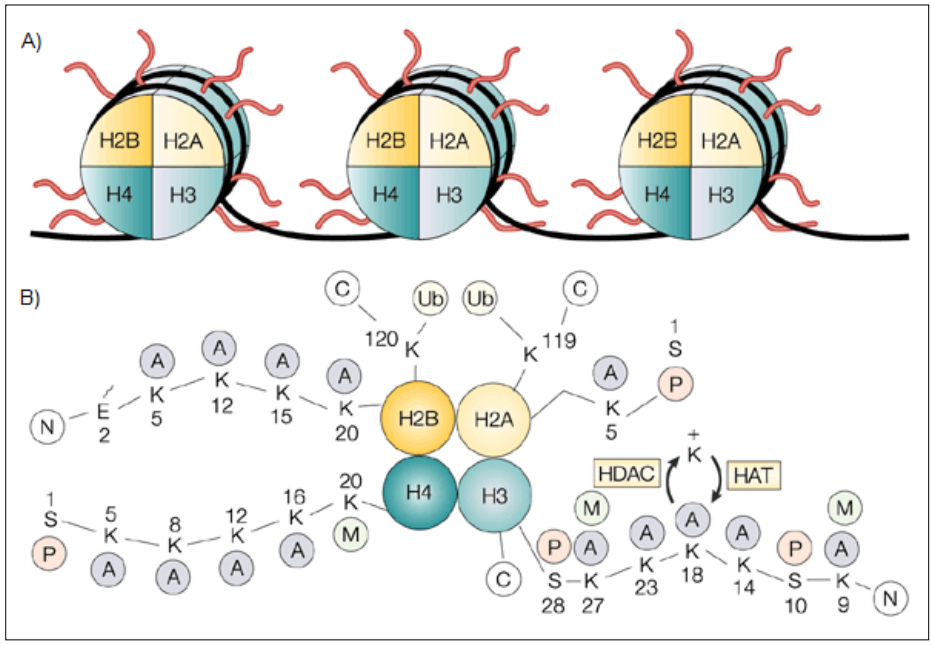


Figure 1.5: *A) The nucleosome is composed of an octamer of four different core histones (H2A, H2B, H3, and H4). Each histone is present in two copies, so the DNA (black) wraps around an octamer of histones; the core nucleosome. The amino-terminal tails of the core histones protrude from the nucleosome (red). B) The amino-terminal tails of core histones in detail. Lysines (K) in the amino-terminal tails of histones H2A, H2B, H3, and H4 are potential acetylation/deacetylation sites for histone acetyltransferases (HATs) and histone deacetylases (HDACs). Acetylation neutralizes the charge on lysines. A, acetyl; C, carboxyl terminus; E, glutamic acid; M, methyl; N, amino terminus; P, phosphate; S, serine; Ub, ubiquitin. Source: [13].*

Given that different modifications exist and multiple sites can be modified, the cell has a plethora of possibilities to modify the core nucleosome. Different modifications can be present simultaneously and this ensemble of modifications is referred to as the *histone code* [16]. Depending on the set of modifications present, the DNA wrapped around the modified nucleosome is accessible or inaccessible (i.e. respectively activated or repressed). For example histone H_3 lysine 4 tri-methylation ($H_3K_4me_3$) is an activating modification [17].

1.3 On the origin of variation

Apparent physical differences and non-physical differences (such as predisposition for certain diseases) within and between populations are mainly caused by different proteins actively being expressed. These differences, the total set of variation between individuals, are relative and dynamic.

Relative because the reference, to which a certain individual is compared, is frequently chosen arbitrarily. For example, the first complete human genome ever sequenced, and consequently the first reference genome ever, originated largely from a single anonymous donor (named RPCI-11) from Buffalo, NY, USA [18]. Dynamic because the reference can change over time. Recently, the National Center for Biotechnology Information (NCBI) released the Genome Reference Consortium - human (GRCh) version 37 of the human genome, indicating that it has changed frequently over the recent years. Assuming an appropriate reference is available, individual variation can be described as compared to this reference.

Bearing the central dogma of molecular biology in mind, there are three levels (i.e. DNA, RNA, and protein) of biological information and two transitions (i.e. transcription and translation) between these levels. Each of those five levels or transitions is a possible source of variation.

1.3.1 Genomic variation

Differences in the nucleotide sequences are the most obvious source of phenotypical differences. These differences can be inherited from the parents during the reproductive process or arise *de novo* during DNA replication despite the *proofreading* functionalities present in the DNA polymerase enzyme. These differences of the nucleotide sequence are referred to using different terms such as mutations, poly-

morphisms, structural variation etc. Although these terms carry different nuances, the underlying concept is identical, i.e. changes in the nucleotide sequence.

1.3.2 Transcriptional variation

Although an individual can be perfectly identical to a reference on the level of the genomic nucleotide sequence, the individuals can differ phenotypically due to different proteins being expressed. During the transcription process, RNA polymerase can introduce *de novo* mutations in the mRNA in a random fashion. However, as each gene is transcribed from DNA to multiple mRNA copies, the downstream effect will be limited. The cell will have sufficient correctly transcribed mRNA molecules and thus functional translated proteins.

Gene expression levels can differ by alteration of the transcription rate; the human cell has a plentitude of mechanisms at its disposal to regulate these transcription rates. By means of an initiating signal, gene expression is initiated, after which a transcription factor, a molecule binding to both DNA and RNA polymerase, is activated which recruits other members of the transcription machinery. Along this transcription activating pathway, different inducing and repressive mechanisms exist. For example, the tryptophan repressor (trp repressor) is a very well described negative feedback regulator of transcription in *Escherichia coli*. When the amino acid tryptophan is in plentiful supply in the cell, trpR binds two molecules of tryptophan, which alters its structure and dynamics so that it becomes able to bind to DNA. When the DNA is bound by this molecule, transcription of the DNA is prevented, suppressing tryptophan producing gene expression. When the cellular levels of tryptophan decline, the tryptophan molecules on the repressor fall off, allowing the repressor to return to its inactive form [19].

Compared to the transient nature of the transcription regulation mechanism described in the previous paragraph, epigenetic modifications can prevent gene transcription and shut down certain genes for longer periods of time or even permanently. This suppressed gene activity can be desired (e.g. suppression of embryonic genes in adults) or disease causing (e.g. suppression of tumor suppressor genes) [20] (Figure 1.6).

After the initial transcription process has been carried out, the pre-mRNA is spliced into the mature mRNA. Alternative splicing is a process by which the exons of the pre-mRNA are reconnected in another than the default way during RNA splicing.

The resulting mRNAs will be translated into different protein isoforms, greatly increasing the biodiversity of proteins that can be encoded by the genome. There are numerous modes of alternative splicing observed, of which the most common are exon skipping and intron retention. A particular exon or intron may be included in mRNAs under some conditions or in particular tissues, and omitted from the mRNA in others [21]. Alternative splicing is a widespread phenomenon in humans with ~95% of the multi-exonic genes showing alternative splicing [22].

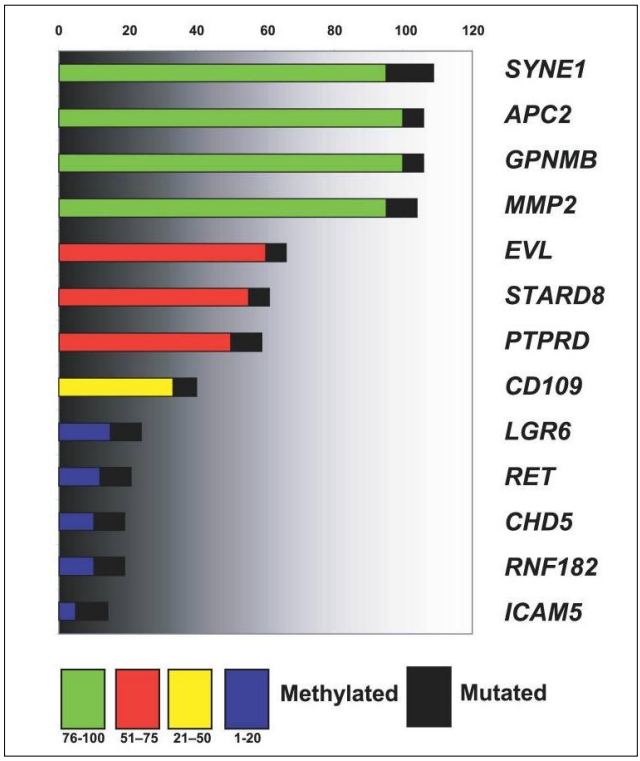


Figure 1.6: *Relationship between methylation status and mutation for 13 genes involved in colorectal cancer. The same genes which are often methylated in patients are also often mutated. Source: [1].*

1.3.3 Post-transcriptional variation

RNA editing describes a set of processes in which transfer RNA (tRNA), ribosomal RNA (rRNA), and mRNA is edited in the nucleus or cytosol after it has been transcribed. RNA editing in mRNAs (and *de novo* transcription errors) effectively alter

the amino acid sequence of the encoded protein so that it differs from that predicted by the genomic DNA sequence [23].

RNA editing through the addition or deletion of one or more uracils (U) has been described and is sometimes referred to as *pan-editing*. During the pan-editing process, one or more Us are inserted or deleted at one or more locations, resulting in a frameshift and a translated protein which differs from the native gene product [24].

Topical RNA editing refers to the process in which one or a few sites are modified, typically of an mRNA molecule. Topical RNA editing involves the deamination of a cytosine to a uracil residue (C-to-U editing) or deamination of an adenine to a hypoxanthine. Hypoxanthine together with the ribose group is referred to as inosine, hence the name adenosine-to-inosine editing (A-to-I editing) for the latter type of editing. The modified RNA will yield a protein with an amino acid sequence altered at the modification site [25] or can be targeted for degradation by the modification [26].

1.3.4 Translational variation

Over the years, evidence has amassed which indicates the human genome is transcriptionally very active. Many of the transcribed RNA molecules show no open reading frame (ORF) and thus do not form the basis for a protein product [27, 28]. As these RNA molecules have a function, despite the lack of an ORF, and thus lack of protein coding functionality, they are referred to as non-coding RNAs (ncRNAs).

MicroRNAs are short (~22 bp) ncRNA molecules found in cells of almost each eukaryotic species. MiRNAs are processed from longer precursor molecules, in several distinct steps, to yield the mature 22 bp molecule. These mature miRNAs then bind to complementary sequences on target mRNA transcripts. The binding usually results in translational repression or target degradation and subsequent gene silencing [29].

Long non-coding RNA (lncRNA) were recently described and comprise the RNA molecules which have no protein coding function and are longer than 200 bp. Many functionalities are attributed to these RNAs, most notably the translational repression functionality of the dendritic BC1 RNA in neurons [30]. lncRNAs are also involved in splicing [31] and DNA methylation [32].

According to the scanning model, 40S ribosomal subunits are recruited to the 5′-terminal cap structure, scan mRNA in the from 5′ to 3′ and can initiate translation at the first AUG encountered, eventually yielding a methionine as the first amino acid of the translated protein [33]. Recently, however, this standard model has been challenged and has been shown to be prone to considerable variation. For example, if the sequence context is suboptimal, some 40S ribosomal subunits recognize the AUG as a translation initiation site, but others may skip it, continue to scan in the 3′-direction, and initiate translation at a downstream AUG, thus initiating translation at an alternative translation initiation site (aTIS) in a process known as leaky scanning [34]. Furthermore, translation initiation has been observed at non-AUG codons as well [35].

The concept of alternative translation initiation can be taken to an extreme when the ribosome is capable of initiating translation in the middle of an mRNA transcript without initially being recruited to the 5′-cap. Usually, in eukaryotes, translation can be initiated only at the 5'-end of the mRNA molecule since 5'-cap recognition is required for the assembly of the initiation complex. An internal ribosome entry site (IRES) is a nucleotide sequence that allows for translation initiation in the middle of an mRNA. Although internal ribosome entry sites are frequently observed in viruses, they are relatively rare in humans. However, they are occasionally observed in humans as well [36].

1.3.5 Post-translational variation

Proteins can be modified after their translation in a process called post-translational modification. Several different post-translational modification processes are described such as addition or removal of amino acids, addition of disulfide bridges, and addition of small chemical groups.

For example, insulin is synthesized from the pro-insulin precursor molecule by the action of two proteolytic enzymes and an exoprotease. These different enzymes remove the center portion – the C-chain – of the pro-insulin molecule. The remaining polypeptides (51 amino acids in total), the A- and B-chain, are bound together by disulfide bonds, eventually resulting in a protein where both the initial N-terminal portion and C-terminal portion of pro-insulin are bound together [37].

Post-translational modifications often involve the addition of biochemical functional groups to one or more amino acids after the translation has been completed.

These modifications do not necessarily change the amino acid sequence of the protein, but can alter the biochemical function, chemical nature, or the structure of the protein. These modifications include the addition of small chemical groups such as acetyl or methyl groups, addition of sugar groups in a process known as glycosylation, addition of other peptides or proteins (e.g. ubiquitination), and many more [38].

As these modifications are studied in the field of proteomics and peptidomics using a combination of complex spectrometry methodologies [39], these will not be further discussed in this dissertation.

2 DNA sequencing

2.1 Sanger sequencing

Early DNA sequencing efforts were focused on unraveling the genomic sequence of micro-organisms such as viruses, bacteria, and fungi. The chain-terminator method, named Sanger sequencing after its inventor Frederick Sanger, is a technique wherein an ssDNA template is amplified using a DNA primer, nucleotide triphosphates (both dNTPs and ddNTPs), and DNA polymerase. 2′-Deoxynucleoside triphosphates (dNTPs) are normal nucleotides used in cells for DNA synthesis. 2′,3′-dideoxynucleoside triphosphates (ddNTPs), the chain-terminating nucleotides, lack a 3′-OH group and render the molecule incapable of forming a phosphodiester bond between two nucleotides, effectively terminating the DNA strand (Figure 1.1 A). These terminator ddNTPs are mixed together with the regular dNTPs in a certain concentration and are used in a classical polymerase chain reaction (PCR) to amplify the DNA-template. The DNA polymerase will at random pick a regular or terminator nucleotide when incorporating the nucleotides during the amplification process. Eventually, this will result in the original template being amplified in a set of molecules with different sizes each ending with a terminator nucleotide [40].

In the early days of Sanger sequencing the whole process was repeated four different times (using ddATPs, ddCTPs, ddTTPs, and ddGTPs) with the terminator molecule being radioactively labeled. The terminated DNA-fragments with different sizes were then loaded onto a gel and separated using electrophoresis. The faster the fragment eluted on the gel, the smaller the DNA-fragment and the closer to the 5′-end of the DNA-fragment the terminator nucleotide was incorporated. By analyzing the four different gels together, it is possible to reconstruct the sequence of the original DNA-template. This technique allowed the nucleotide sequence of bacteriophage ϕ X174 to be determined [41]. A schematic overview of the Sanger sequencing technique is shown in Figure 2.1.

This technique was improved for more than ten years before another breakthrough was achieved. The introduction of fluorescent labels permitted sequencing in one single reaction. Rather than using four different reactions for each of the four different nucleotides, each terminator nucleotide could be labeled with a specific fluorescent dye. From then on, all four labeled terminator nucleotides were used in a single reaction and the obtained terminated DNA-fragments were separated on one single gel using (capillary) electrophoresis. The original DNA-sequence was then obtained

by exciting the fluorescent dyes attached to the terminator nucleotides and 'reading' the order of the fluorescent colors [42]. Improvements in electrophoresis techniques and improved sample preparation led to the first automated DNA-sequencing instrument that could sequence up to 384 DNA samples in a single run [43].

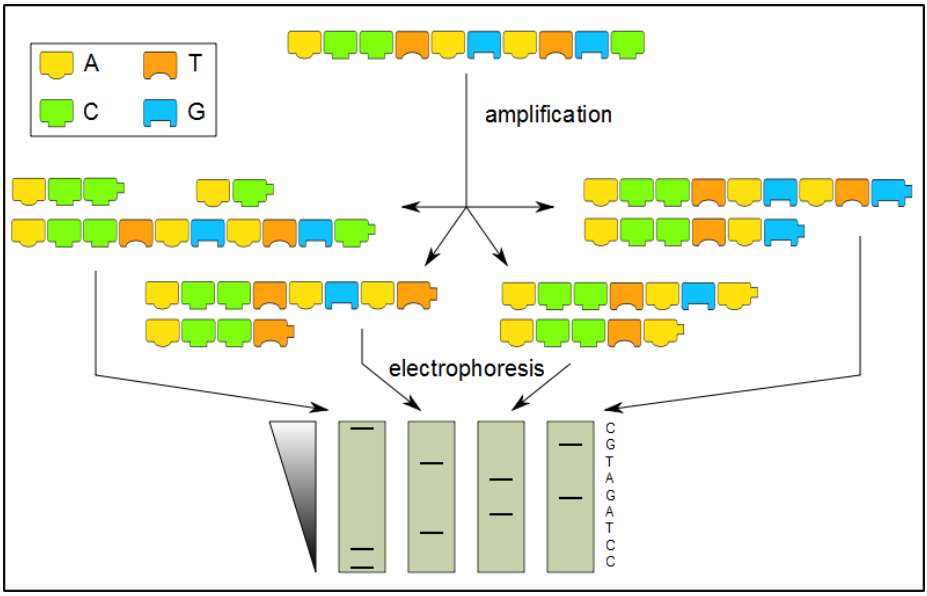


Figure 2.1: *Overview of the Sanger sequencing technique. The four different terminator nucleotides are added to four separate reactions. Each reaction contains a mix of fragments with different lengths. Once these fragments are separated on a gel using electrophoresis, the original sequence can be derived.*

Shortly after the development of the first automated sequencing instruments, both the public International Human Genome Sequencing Consortium (IHGSC) and the private company Celera Genomics started an effort to determine the sequence of the three billion base pairs that make up the human genome, and identify all the genes in the human genome. Both completed a draft version of the human genome at the beginning of the 21st century [44, 45].

The human genome project pushed the Sanger sequencing technique to its limits, and it was clear there was no room left for improvement. Sanger published its groundbreaking paper in 1977 and yet, approximately 25 years later, the technique was in essence still the same.

2.2 Next-generation sequencing

A combination of factors led to the development of a set of different sequencing technologies which hit the market commercially around the year 2007. These new technologies were truly disruptive technologies, and were collectively given the name *next-generation sequencing* (NGS) technologies (also known as second-generation sequencing technologies) to emphasize the breakaway from Sanger sequencing. The newly developed techniques were disruptive in many aspects. Sequencing became easier, faster, and cheaper than ever before (Figure 2.2).

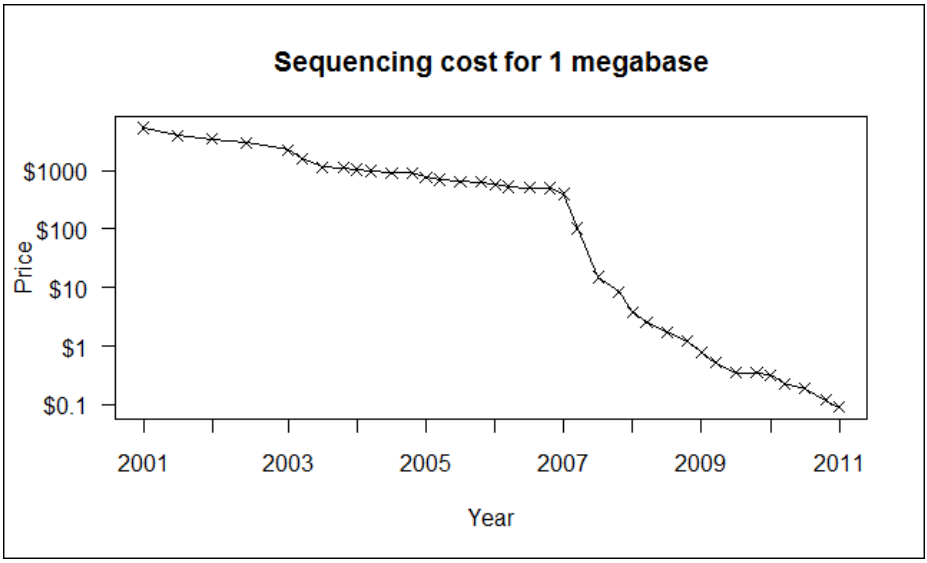


***Figure 2.2**: Evolution of the sequencing cost for 1 megabase. The real start of next-generation sequencing came with the introduction of the 454/Roche Genome Sequencer FLX in 2007. Source: [46].*

2.2.1 Early NGS technologies

Although earlier technologies, developed in the 1990s and early 2000s, could be considered NGS technologies, the first real NGS technology to have had broad market access is generally considered to be the GS20 developed by 454 Life Sciences in 2005. 454 Life Sciences was acquired by Roche in 2007 and experienced rapid growth with the backing of the pharmaceutical conglomerate. It was the release of the follow-up sequencer, the Genome Sequencer FLX (GS-FLX) in 2007, which marked the real start of next-generation sequencing.

One of the early NGS technologies that never made it commercially was *massively parallel signature sequencing* (MPSS), developed by Lynx Therapeutics in the 1990s [47]. The technology was complex and susceptible to sequence-specific bias and, as a consequence, was never commercialized. Lynx Therapeutics merged with Solexa in 2004. Together they developed the much more simple *sequencing-by-synthesis* methodology which turned MPSS obsolete and eventually sealed the fate of MPSS. Solexa was acquired by Illumina in the beginning of 2007.

Polony sequencing (a contraction of polymerase and colony) was a technology with high potential which never saw the market as a commercial device on itself. The technology combined an in vitro paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing chemistry to sequence DNA sequences [48]. Using this technique, an *Escherichia coli* genome was sequenced in 2005 at an accuracy of > 99.9999% and a cost approximately 1/10 that of Sanger sequencing [49]. The technology was never commercialized by the original developers, but licensed to Agencourt Biosciences, which was subsequently spun out into Agencourt Personal Genomics. The technique was ultimately incorporated into the Applied Biosystems/Life Technologies SOLiD platform which hit the market in 2008.

These three NGS technologies (454, Illumina, and SOLiD) enable researchers to read a number of DNA sequences that is several orders of magnitudes bigger and at a cost that is several orders of magnitude smaller than the conventional DNA sequencing technologies (i.e. Sanger sequencing and its variations). The cost of determining the human genome was estimated at $2.7 billion for the IHGSC genome and $300 million for the Celera genome. In 2008 several human genomes were already sequenced in approximately 1.5 months and at a cost of approximately $1.5 million [50, 51]. Since then, the cost and turn-around time has come down even more and the $1,000 genome is not far away any more [46].

The combined direct and indirect cost of sequencing a megabase using Sanger sequencing was over $5,000 at the end of 2001. The cost per megabase gradually declined to reach approximately $400 at the end of 2007. With the introduction of the GS-FLX in 2007, the sequencing cost nosedived. Over the recent years, the throughput of the different NGS platforms kept increasing and consequently the cost per sequenced base kept decreasing to reach an astonishing $0.09 per megabase at the end of 2011, again emphasizing the disruptive nature of the next-generation sequencing technologies [46] (Figure 2.2).

2.2.2 454 sequencing

Large-scale parallel *pyrosequencing* from 454 Life Sciences/Roche (formerly Branford, CT, USA; now Basel, Switzerland), commonly named 454 sequencing, analyzes hundreds of thousands DNA-templates in approximately half a day [52]. The latest version of the sequencing technology, Titanium, enables a throughput of around 1 million sequences with an average length between 400 and 600 bp resulting in a 0.4 to 0.6 gigabase yield per 10h run [53].

454 sequencing uses a rather complex library preparation procedure followed by an emulsion PCR (emPCR) to amplify the DNA-fragments to assure a sufficiently strong signal in further processing steps. Parallel pyrosequencing is then used to read the actual DNA sequences.

DNA is fractionated into smaller fragments of 300 to 800 bp. The fragments are made blunt at each end and short adaptors are then ligated onto the two ends of the fragments. These adaptors provide priming sequences for both amplification and sequencing of the DNA-templates. One adaptor (Adaptor B) contains a 5'-biotin tag for immobilization of the DNA library onto streptavidin-coated beads. After a cleanup step and a quality control the immobilized template DNA-fragment is amplified on the bead using an emPCR [54].

An emPCR is a compacted PCR with the template DNA-fragment being attached to a bead and the bead, together with all the necessary reagents, being embedded in an aqueous droplet within an oil phase. The template is amplified in the droplet eventually resulting in a bead carrying each up to ten million copies of the template DNA-fragment. After the emPCR, the emulsion is broken and the DNA strands are denatured. The beads carrying single-strand molecules are then enriched and deposited into wells of a fiber-optic slide (PicoTiterPlate) containing millions of such wells. Beads containing the necessary reagents are also distributed over the wells before the actual pyrosequencing starts. In effect, each well in the fiber-optic slide can be considered a micro-reactor wherein a single DNA-template will be sequenced [55].

The four different nucleotides are added sequentially in a fixed and known order over the fiber-optic slide. When the added nucleotide is complementary to the single-strand DNA-template present in a well, the nucleotide is incorporated by the DNA polymerase. Incorporation of one (or more) nucleotide(s) by the DNA poly-

merase releases pyrophosphate (PPi) stoichiometrically. Adenosine-5′-triphophate (ATP) sulphurylase quantitatively converts the PPi to ATP in the presence of adenosine-5′-phosphosulfate (APS). This ATP acts as fuel to the luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light in amounts that are proportional to the amount of ATP [55, 56] (Figure 2.4). This light signal is eventually captured and processed by the CCD camera present in the sequencer. The camera and downstream hardware and software components can capture and process the signals generated in the million wells simultaneously, hence the name *large-scale parallel pyrosequencing*.

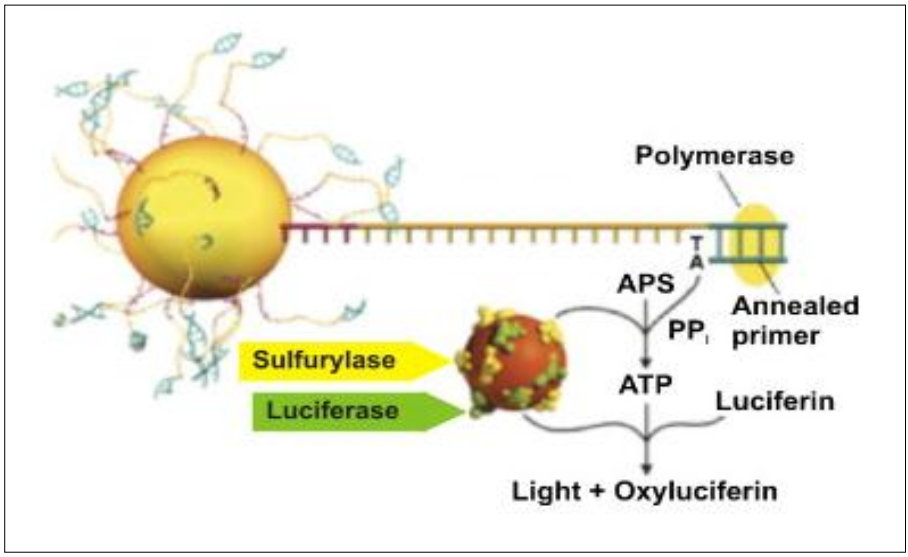


Figure 2.4: *Schematic illustration of the pyrosequencing reaction used in the 454 sequencing technology. The emitted light intensity is proportional to the number of incorporated nucleotides. Source: [55].*

The light signal intensity is proportional to the number of nucleotides incorporated. For example, when a stretch of four As (called a 4 A homopolymer) is present in the template DNA-fragment and Ts are added over the fiber-optic slide, the detected light signal upon incorporation will be approximately four times as strong as compared to a signal obtained when only a single base were incorporated. However, the signal intensity is not completely proportional to the bases incorporated, resulting in a global *under-calling* of homopolymer stretches [51]. For homopolymers that are longer than ~10bp, the signal tends to completely flatten out, making it completely impossible to discriminate between for example a 13bp and 14bp homopolymer.

As this global under-calling of homopolymers is specific for the pyrosequencing technique, techniques have been developed which try to alleviate this problem. The first software package specifically designed to fix this homopolymer problem was PyroBayes [57]. However, despite the development of PyroBayes, homopolymers remain problematic [58]. Other more specialized packages, which are only applicable in certain experimental designs, have been released as well. PyroNoise/AmpliconNoise is a set of packages specifically designed to improve the raw intensity signal (and thus improve the accuracy of homopolymers) in amplicon sequencing experiments in bacterial species [59]. Recently, SCOPE++ was developed to accurately call homopolymers in poly-A tails when sequencing cDNAs [60].

Eventually a sequence of raw light intensities is obtained for each of the wells present in the fiber-optic slide. These raw light intensities are processed by the proprietary image processing pipeline and converted to normalized light intensities which are proportional to the number of bases incorporated in each cycle of nucleotide addition. These intensities are stored in a Standard Flowgram Format (SFF) file (Figure 2.5). Using the known order in which the nucleotides were added sequentially, it is relatively easy to convert this *flowgram data* to actual DNA sequences. A global overview of the 454 sequencing methodology is given in Figure 2.6.

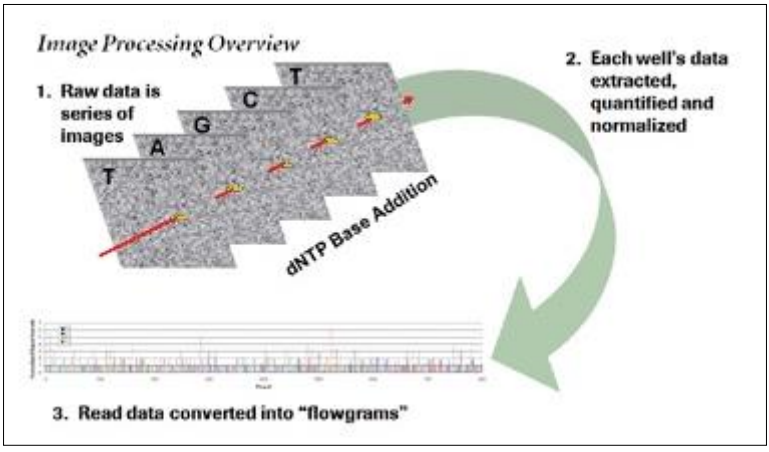


Figure 2.5: *Schematic overview of the conversion of light intensities to flowgrams. 1) The images are generated by the CCD camera. 2) Data from different flows in each well is extracted, quantified, and normalized. 3) All data from a single well is grouped and converted into a single flowgram per well. In a final step, flowgram data is easily converted to a nucleotide sequence. Source: 454 Life Sciences/Roche.*

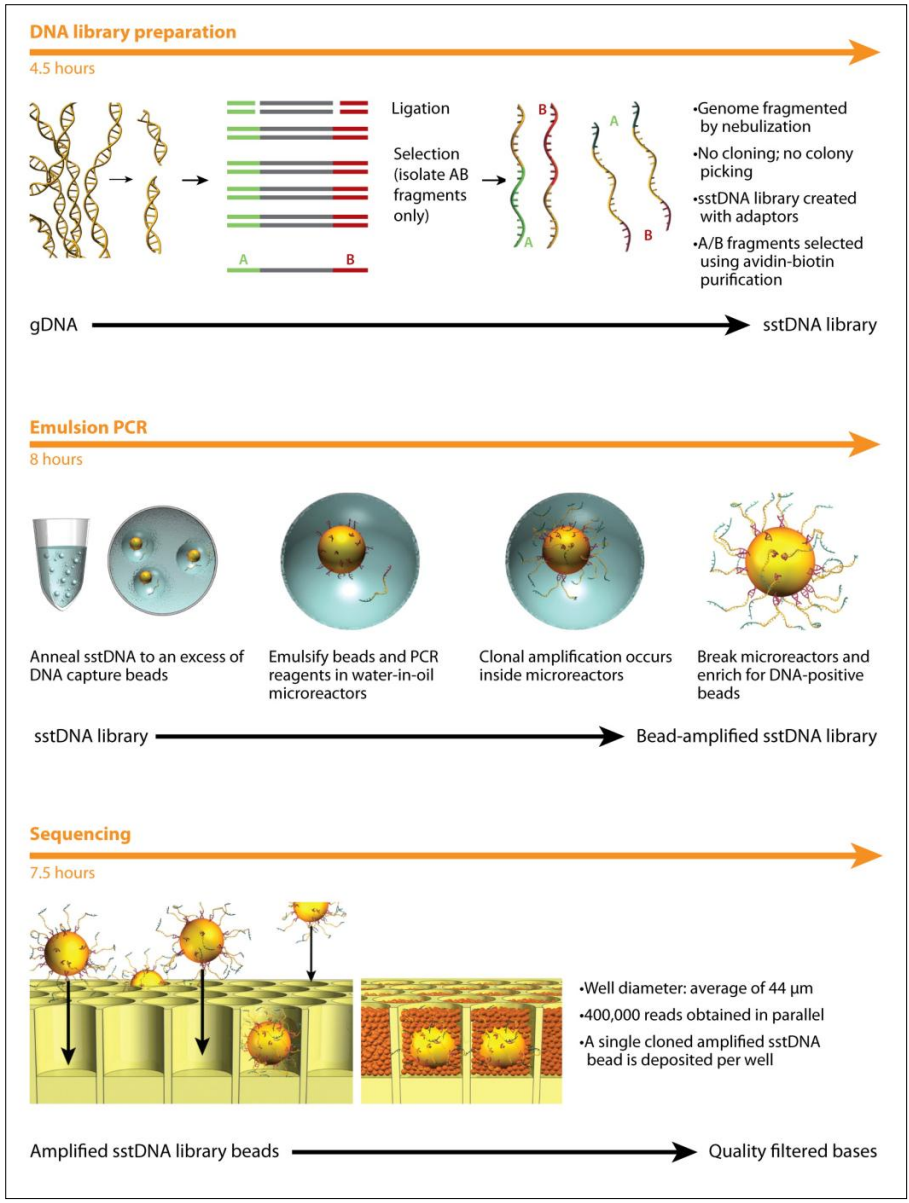


Figure 2.6: *Overview of the 454 sequencing methodology. DNA is fractionated and processed to yield an appropriate library. DNA sequences are then bound onto beads and amplified using an emulsion PCR. After the emulsion PCR, beads with amplified DNA are deposited into wells of a fiber-optic slide together with necessary reagents. Parallel pyrosequencing eventually yields the sequence of the original DNA molecules. Source: [61].*

2.2.3 Illumina sequencing

Large-scale parallel *sequencing-by-synthesis* from Solexa/Illumina (formerly Hayward, CA, USA; now San Diego, CA, USA), commonly named Illumina sequencing, analyzes billions of reads in approximately a week. The Illumina sequencing technology uses a user-defined sequencing length (between 36 bp and 150 bp) and the actual sequencing time depends on the chosen sequencing length [55]. Currently, Illumina has several different sequencers on the market such as the Genome Analyzer IIx (the first model developed), the HiSeq (the latest branch of models) and the MiSeq (a small benchtop machine allowing faster processing using few samples). At the moment, the HiSeq 2000 has a throughput of 3 billion reads with a maximum length of 2 x 100bp, eventually resulting in a 600 gigabase yield in approximately 11 days [62].

Illumina sequencing uses a relatively simple library preparation followed by a bridge amplification process to amplify DNA-fragments. A reversible dye-terminator sequence-by-synthesis process is used to read the actual sequences [61].

DNA is first fractionated into smaller fragments of 100 to 300 bp. After this fragmentation, the DNA-fragments are end-repaired and prepared for further processing. The fragments are then size-selected and purified. Cluster generation is performed on the Illumina cBot station using a process called *bridge amplification*. Cluster generation is necessary to concentrate molecules together to be able to obtain a sufficiently strong signal in further processing, just as an emPCR is needed in 454 sequencing.

DNA-fragments are attached to the flow cell by hybridizing to the lawn of primers which are present on the silica flow cell. The lawn of primers consist of a set of oligos, complementary to the adapters ligated to the ends of the fragments during the library preparation process, bound covalently to the flow cell's surface. The complementary strands of the hybridized DNA-molecules are synthesized starting from the primer. The original strand is denatured and washed away, resulting in a single strand DNA-molecule being covalently bound to the flow cell. The bound molecules hybridize with a neighboring complementary primer on the flow cell, effectively forming a U-bridge. The U-bridge is copied from the primer onwards to create a dsDNA. The resulting double-stranded molecule is denatured, hybridized to a neighboring primer to form a new bridge, and extended once again. This bridge amplification process is repeated 35 times to create a dense cluster of approximate-

ly 2000 molecules. Finally, the reverse strands are cleaved off and washed away and the sequencing primer is hybridized to the DNA-templates [63].

The generated clusters (up to 3 billion) are then sequenced simultaneously. The DNA-templates are copied base by base using the four nucleotides (A, T, C, and G) which are fluorescently labeled (each nucleotide has its own fluorescent dye) and reversibly terminated. Unlike pyrosequencing, the four different nucleotides are added simultaneously and the DNA-molecule can only be extended one nucleotide at a time. After each synthesis step, unbound nucleotides are washed away and the clusters are excited by a laser which causes fluorescence of the last incorporated base. After that, the fluorescence label and the blocking group are removed allowing the addition of the next base [64, 65] (Figure 2.6).

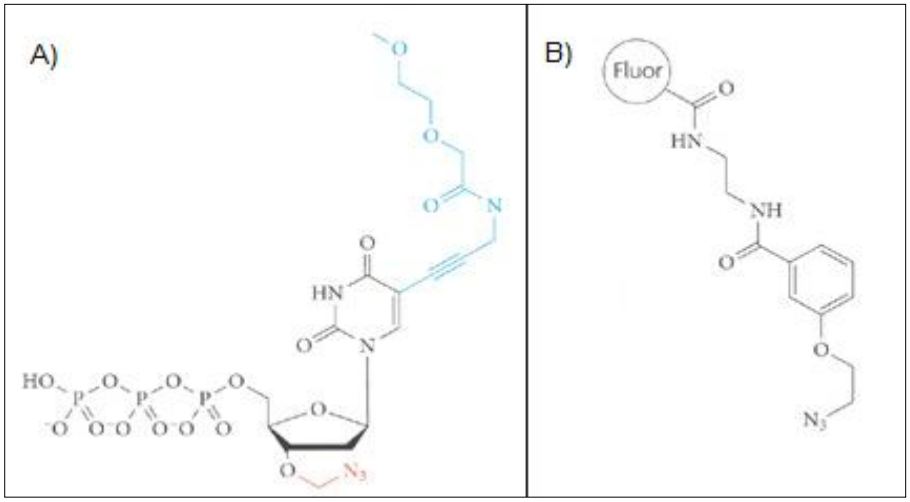


Figure 2.6: *Reversible dye-terminator molecule used in Illumina sequencing. A) Modified thymine nucleotide with a 3'-O-azidomethyl blocking group in red. The blue group indicates the residual linker structures from which the fluorescent group was cleaved off. B) The fluorescent group which is cleaved off in Illumina sequencing. Source: [65].*

The fluorescent signal after each incorporation step is captured by a built-in CCD camera. Each of the clusters on the silica flow cell will produce a certain colored signal depending on the base that was incorporated. Further downstream analyses finally yield the actual DNA-sequences [61, 64]. An overview of the Illumina sequencing-by-synthesis methodology is given in Figure 2.7.

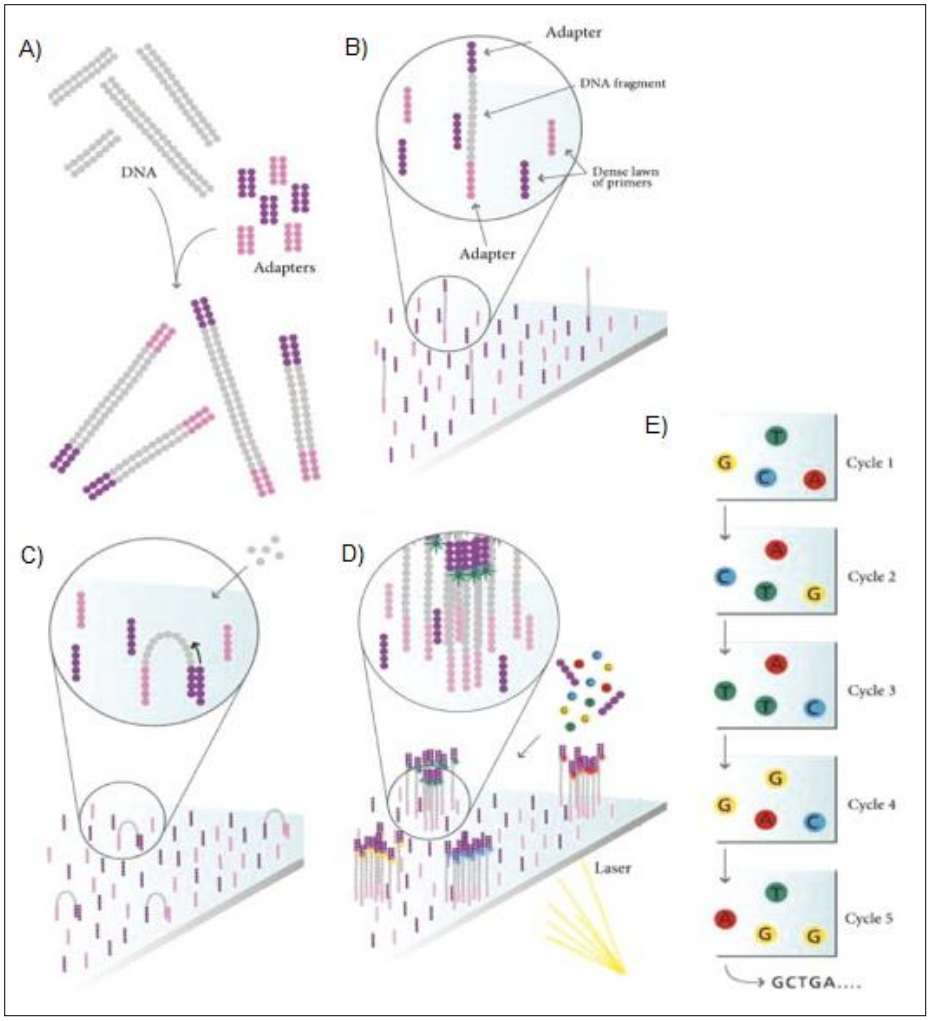


Figure 2.7: *Overview of the Illumina sequencing methodology. A) Genomic DNA is randomly fragmented and adapters are added to the fragments. B) Single-stranded fragments are attached randomly onto flow cells. C) Bridge amplification is performed to generate double-stranded fragments. D) Following repeated cycles of denaturation and bridge amplification, millions of DNA copies are present on each flow cell. E) The sequence is determined using four labeled reversible terminator nucleotides and DNA polymerase. Unincorporated terminators are washed away, and the sequence is determined in each flow cell following laser excitation. The blocked 3′-terminus and fluorophore are removed enzymatically from the incorporated base and the cycle is repeated. Source: [64].*

2.2.4 SOLiD sequencing

Large-scale parallel *sequencing-by-ligation* from Applied Biosystem (ABI)/Life Technologies (formerly Foster City, CA, USA; now Carlsbad, CA, USA), commonly named SOLiD sequencing (Sequencing by Oligonucleotide Ligation and Detection), analyzes hundreds of millions reads in approximately a week. The sequencing technology uses a user-defined sequencing length (between 35 bp and 75 bp) and the actual sequencing time, just as is the case with Illumina sequencing, depends on the chosen sequencing length [66].

Life Technologies has several sequencers on the market with different specifications. The *5500 System* has a throughput of approximately 50 gigabases in a 7 day run, the *5500xl System* has a throughput of approximately 75 gigabases in a 7 day run and the advanced *5500xl System* (using nanobeads rather than microbeads) has a throughput of approximately 150 gigabases in a 7 day run [67].

SOLiD sequencing uses a relatively simple library preparation followed by an emPCR process to amplify DNA-fragments. A sequencing-by-ligation process is used to read the actual sequences [61].

After DNA fractionating and size-selection, adaptors are ligated to the 3'-end (P1 adaptor) and the 5'-end (P2 adaptor). The DNA-fragments are then hybridized to microscopic beads which are coated with an adaptor complementary to the P1 adaptor. An emPCR is carried out, similar to the emPCR which is present in 454 sequencing, to ensure a sufficiently strong signal in subsequent steps. After the emPCR, the emulsion is broken and the beads are deposited on, and covalently bound to, a glass slide [68].

In the first sequencing-by-ligation cycle, a universal primer is bound to the P1 adaptor. Then, a mixture of semi-degraded nucleotide octamers is added, together with a ligation mix. These semi-degraded nucleotide octamers are characterized by one of four fluorescent labels which correspond with a certain bi-nucleotide on the fourth and fifth position of the octamer. When a matching octamer hybridizes to both the universal primer at the 3'-end and the DNA-template bound on the bead, DNA ligase seals the phosphate backbone. After the ligation step, a fluorescent readout identifies the fourth and fifth base of the octamer. The ligated octamer nucleotides are cleaved off after the fifth base, removing the fluorescent label. Hybridization, ligation, and fluorescence readout steps are now repeated several

times, determining base 9 and 10 of the sequence in the second cycle, base 14 and 15 in the third cycle, and so forth. This whole process is repeated five times in total but with a universal primer being one base shorter each time, allowing all the bi-nucleotides of the DNA-template to be identified (Figure 2.8) [61, 68].

The fluorescent signal in each cycle is captured and processed by a built-in CCD camera in a way similar to the process used in Illumina sequencing. The biggest difference, however, is that the DNA-template is not read from start to end, but rather in five different steps which identify bases each time five nucleotides away from each other. At first, this might seem unnecessary complex, but this system makes SOLiD sequencing a very accurate method. Effectively, each base is identified two times. For example base 4 is identified by the base 4-5 di-nucleotide (using universal primer 1) and the base 3-4 di-nucleotide (using universal primer 2). The overlap of both signals will finally identify the fifth base of the DNA-template. In a similar fashion, all the bases of the DNA-template are identified.

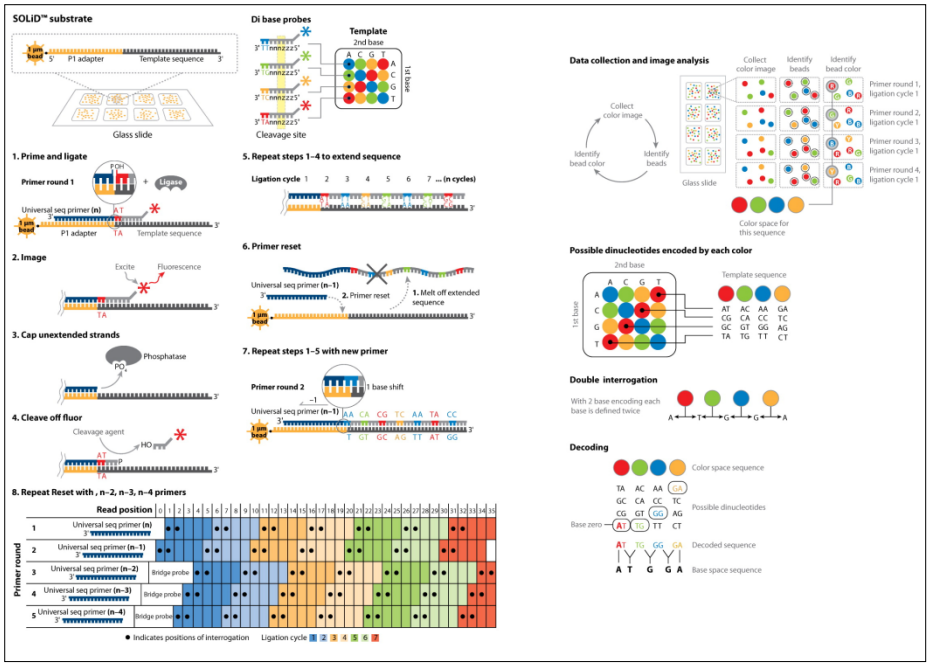


Figure 2.8: *Overview of the SOLiD sequencing methodology (left) and the principles of data analysis to decode two-base encoding (right). Source: [61].*

2.3 Next-next-generation sequencing

With the development of the novel sequencing technologies seen around the year 2008 came the term *next-generation sequencing technologies*. Early large-scale sequencing efforts (using NGS technology) were carried out almost exclusively by large genomics institutes such as the Wellcome Trust Sanger Institute and the Broad Institute. When the technology matured and sequencing prices dropped rapidly, next-generation sequencing became available to, and was used by, a broader public. As a consequence, next-generation sequencing was affordable for a plethora of companies, research institutes, and laboratories.

This increased adaptation of NGS technology and accompanying market potential spurred investments in even newer sequencing technologies that would never have been profitable before. It was this same market potential that led to a big consolidation phase in the sequencing market. As mentioned before, Solexa was acquired by Illumina in 2007, 454 Life Sciences was acquired by Roche Diagnostics in 2008 and Applied Biosystems and Invitrogen merged to form Life Technologies in 2008. The future potential of next-generation sequencing was confirmed in the beginning of 2012 when Roche launched a hostile $5.7 billion takeover bid to acquire Illumina. However, the bid was eventually withdrawn by Roche, resulting in a status-quo with three major companies in the market.

Since 2008, even newer sequencing technologies have been developed. These new technologies are collectively named third-generation sequencing or next-next-generation sequencing. Some of these new technologies have already been acquired by one of the big three NGS companies (e.g. Ion Torrent), others are on the market as an individual company (e.g. Pacific Biosciences) and others might be not as promising as thought in the beginning and may disappear in the very near future (e.g. Helicos).

2.3.1 Helicos BioSciences

One of the first third generation sequencing instruments, the HeliScope, was developed by Helicos BioSciences (Cambridge, MA, USA). The HeliScope distinguishes itself from former NGS instruments because it uses a technology called true single molecule sequencing (tSMS). This technology uses a single DNA molecule or RNA molecule as a template in comparison to an amplified template obtained by emPCR or bridge amplification [69].

Sequencing is carried out using a sequencing-by-synthesis methodology comparable to the technology used in both 454 and Illumina sequencing. Unlike Illumina's reversible dye-terminator technology though, *Helicos Virtual Terminators technology* is used. The four terminator nucleotides are labeled with the same dye and dispensed individually in a predetermined order, analogous to a single-nucleotide addition method used in 454 sequencing. A DNA polymerase copies or reverse transcribes the template molecule, temporary fixating an excitable dye. Signal is measured, the dye is cleaved and washed away, and the signal is processed in a fashion similar to the Illumina technology [65].

This tSMS technology offers tremendous advantages as both bridge amplification and emPCR processes are known to introduce biases. As tSMS technology uses a single molecule as a template, amplification biases are eliminated and real quantitative DNA or RNA data can be obtained. This is of special importance in transcription experiments where quantitative RNA data might get distorted by the amplification process [70, 71].

Although the technology appeared very promising in the beginning and was commercialized around 2009, only a limited amount of sequencers were sold in the past few years. Helicos BioSciences barely managed to survive 2011 and the future of the company remains uncertain.

2.3.2 Pacific Biosciences

Pacific Biosciences (Menlo Park, CA, USA) developed a technology called Single-Molecule Real-Time (SMRT) sequencing. A single DNA or RNA molecule is used as a template, just as is the case in the Helicos technology, eliminating the need for an amplification process. Real time sequencing uses a sequencing-by-synthesis methodology based on the properties of the zero-mode waveguide technology [72]. Pacific Biosciences' first platform, PacBio RS, was commercially released in 2011.

Unlike reversible terminator nucleotides, real-time nucleotides do not halt the process of DNA synthesis (as is the case in for example Illumina sequencing). In the Pacific Biosciences platform, single DNA polymerase molecules are attached to the bottom surface of individual zero-mode waveguide detectors (ZMW detectors). A ZMW detector is 100 nm in diameter, which is smaller than the 532 nm and 643 nm laser wavelengths used in the Pacific Biosciences platform. Light cannot propagate through these small waveguides, hence the term zero-mode. These aluminum-clad

waveguides are designed to produce an evanescent wave that substantially reduces the observation volume at the surface of the polymerase reaction down to the zeptoliter range (10^{-21} l) [65]. This level of confinement enables single-fluorophore detection despite the relatively high concentrations (between 0.1 and 10 µM) required by DNA polymerase for fast and accurate synthesis [73].

The method of real-time sequencing involves imaging the continuous incorporation of dye-labeled nucleotides during DNA synthesis using these ZMW detectors. The residence time of phospholinked dNTPs in the active ZMW site is determined by the rate of catalysis and is in the order of some milliseconds, effectively generating a low background signal. The actual DNA sequence is determined by detecting fluorescence from phospholinked dNTPs trapped in the active site of the DNA polymerase. A fluorescence pulse is produced by the polymerase which is retaining the dNTP with its color-coded fluorophore in the detection region of the ZMW. The duration of the fluorophore retention is much longer than the time scales associated with the background noise [73] (Figure 2.9).

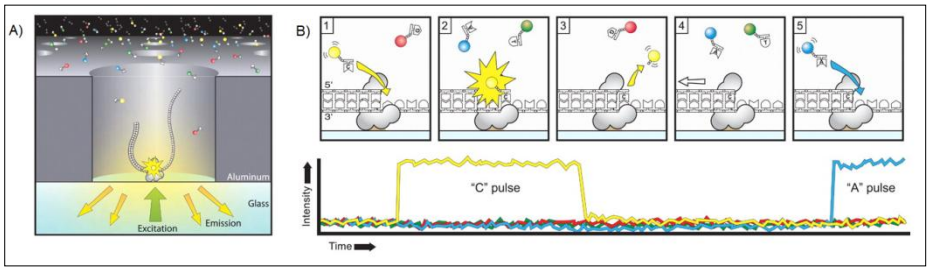


Figure 2.9: *Principle of single-molecule real-time (SMRT) sequencing. A) A single molecule of DNA polymerase is immobilized at the bottom of a ZMW, which is illuminated from below by laser light. The ZMW nanostructure provides excitation confinement in the zeptoliter regime. B) Schematic event sequence of the phospholinked dNTP incorporation cycle, with a corresponding expected time trace of detected fluorescence intensity from the ZMW. (1) A phospholinked dNTP associates with the template in the polymerase active site, (2) causing an elevation of the fluorescence output on the corresponding color channel. (3) Phosphodiester bond formation liberates the dye-linker-pyrophosphate product, which diffuses out of the ZMW, thus ending the fluorescence pulse. (4) The polymerase translocates to the next position, and (5) the next cognate nucleotide binds the active site beginning the subsequent pulse. Source: [73].*

The signal kinetics upon incorporation of a modified nucleotide differs from the signal kinetics upon incorporation of a native nucleotide. The Pacific Biosciences

platform is capable of distinguishing between native nucleotides and modified nucleotides using these signal kinetics. Recently, it was shown that the PacBio RS can sequence the four native nucleotides, methylated cytosines, and hydroxymethylated cytosines [74, 75].

2.3.3 Ion Torrent

Ion Torrent (formerly Guilford, CT, USA; now Carlsbad, CA, USA), acquired by Life Technologies in 2010, developed the revolutionary *ion semiconductor sequencing* (ISS) technology, commonly called Ion Torrent sequencing. Ion Torrent's first platform, the Personal Genome Machine (PGM), was commercially released in 2011. The PGM is a truly revolutionary machine as it offers a throughput similar to the existing NGS platforms, but has lower upfront costs, lower operational costs, and has a much shorter turn-over time per experiment.

In essence, the Ion Torrent technology is very comparable to the pyrosequencing technology used in 454 sequencing. However, unlike 454 sequencing, where a light signal is measured (hence *pyro*; from the Greek πυρ meaning fire), a pH is measured; hence the technology is sometimes called pH-mediated sequencing.

The addition of a nucleotide at the 3'-end of an existing DNA strand involves the formation of a covalent bond, the release of pyrophosphate, and the release of a hydrogen ion which has a positive charge [76, 77]. Ion Torrent sequencing exploits these features of the DNA polymerization process by determining whether hydrogen ions are released upon addition of native dNTPs to an existing single strand reference DNA-template. Native dNTPs are used (compared to dye-nucleotides or terminator-nucleotides) and no expensive optical equipment is needed, making Ion Torrent sequencers very cheap.

Wells on a semiconductor chip, each containing one single-stranded DNA-template to be sequenced and one DNA polymerase, are sequentially flooded with A, T, C or G dNTPs. When an introduced dNTP is complementary to the nucleotide on the template strand, it is incorporated into the growing complementary strand by the DNA polymerase. When it is not complementary, no incorporation occurs. The hydrogen ion released upon incorporation changes the pH of the solution, which can be measured by a pH-electrode. The unattached dNTP molecules are washed away before the next cycle is started [78] (Figure 2.10). Further downstream analyses are

very similar to the process used in 454 sequencing and the final output data is in the Standard Flowgram Format (SFF).

The number of released hydrogen ions in a single cycle is proportional to the length of the homopolymer stretch in the single-stranded reference DNA-template. Hence, the change in pH is proportional to the length of the homopolymer stretch. Just as was the case in 454 sequencing, the proportionality of the signal is not completely linear and there is a general *under-calling* of homopolymers (i.e. the signal for longer homopolymers is too small) [79].

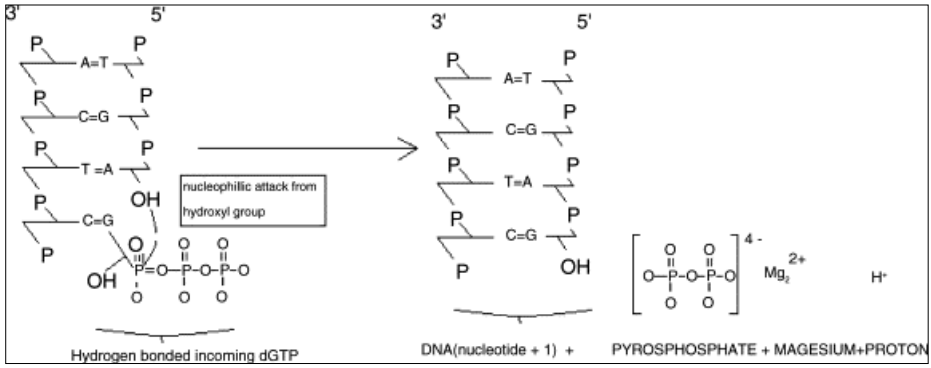


Figure 2.10: *Schematic overview of the pH mediated sequencing used by Ion Torrent. Source: [76].*

2.3.4 Oxford Nanopore Technologies

Oxford Nanopore Technologies (Oxford, United Kingdom), further developed the relatively old *nanopore sequencing* technology, which has been under development since the 1990s [80, 81].

Nanopore sequencing technology usually involves the application of the alpha hemolysin (αHL) protein, a nanopore from a bacteria, which has a diameter of a few nanometers (hence nanopore) and is immobilized in an impenetrable surface. A current, which is constantly measured, is applied in the interior of the nanopore. When a molecule passes through the interior of the nanopore, the applied current is disrupted. Using the profile of the disrupted current, the passing molecule can be identified [82]. Oxford Nanopore Technologies uses both biological nanopores (αHL) as artificial solid state nanopores (e.g. graphene) [83].

Multiple nanopores are fabricated in an array chip, enabling multiple DNA-templates to be analyzed simultaneously, making the technology parallel and thus high-throughout. Oxford Nanopore Technologies is developing different nanopore sequencing approaches: strand sequencing and exonuclease sequencing.

Nanopore strand sequencing is a technology wherein a single-strand DNA-template is forced through the nanopore. The disruption of the applied current in the nanopore is specific for the set of passing nucleotides. Using the current disruption profile, it is possible to identify the individual passing nucleotides (Figure 2.11) [84]. It is important to note there is no deterioration of the quality of the identified bases towards the end of the passing molecule, which indeed is the case on the second-generation sequencing instruments, theoretically enabling molecules with an indefinite length to be sequenced. Recently is has also been shown that this technology can identify modified nucleotides such as methylated cytosines and hydroxymethylated cytosines [12, 85].

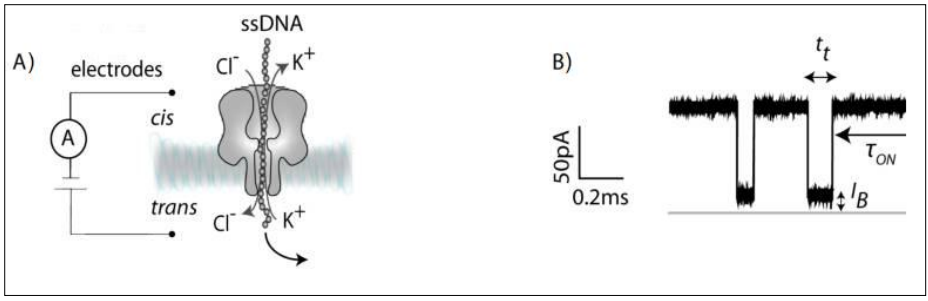


Figure 2.11: *Schematic overview of the strand sequencing technology using a nanopore. A) Electrodes are used to apply a potential to drive DNA through a nanopore and to measure the current. B) Typical current profile of a DNA molecule passing a nanopore showing the translocation times (t_t), the residual current value (I_B), and the inter-event intervals (t_{ON}) for individual DNA translocation events. The actual DNA sequence can be determined using this profile. Source: [84].*

Nanopore exonuclease sequencing is a technology wherein a single-strand DNA-template is temporarily captured on the top side of the nanopore near an exonuclease. The exonuclease enzyme consecutively cleaves off nucleotides from the DNA-strand and forces them to pass through the nanopore. As they pass through the nanopore, they transiently bound with a cyclodextrin adaptor molecule inside the nanopore. The bounding process causes a current disruption which is specific for each passing nucleotide, enabling the complete DNA-template to be identified [86].

Towards the end of 2012, both the GridION and MinION system will be commercially launched. The GridION system consists of scalable instruments (nodes) with consumable cartridges that contain proprietary array chips for multi-nanopore sensing. Each GridION node and cartridge is initially designed to deliver tens of gigabases (Gb) of sequence data per 24 hour period. Oxford Nanopore Technologies has also miniaturized these devices to develop the MinION; a disposable DNA sequencing device the size of a USB memory stick whose low cost, portability, and ease of use are designed to make DNA sequencing universally accessible. A single MinION is expected to retail at less than $900 [87].

3 Identifying new molecular species using second-generation sequencing

3.1 Sequencing libraries

Each source of sequence to be analyzed (dsDNA, ssDNA or RNA) needs to be converted to an appropriate molecule prior to the actual sequencing in a process known as library preparation. Three distinct methodologies exist to generate sequencing libraries, each having a specific structural and contextual organization of the final sequence product.

3.1.1 Single-end sequencing

Single-end sequencing libraries are considered the default sequencing library. After a possible conversion to DNA, in the case of an RNA sample, fragmentation, and size selection, linear DNA fragments with a given length are obtained. These DNA-templates are then sequenced, using one of the available sequencing technologies, from the 5′-end towards the 3-end once, eventually yielding only the 5′-end of the sequenced fragment.

3.1.2 Paired-end sequencing

In paired-end sequencing (available on the Illumina and SOLiD platforms), the initial library preparation is similar to the single-end sequencing library preparation. However, after sequencing from the 5′-end has finished, the DNA-fragment is 'flipped' and sequenced from the 3′-end. Eventually, two sequence tags with a known distance in between are obtained. The approximate distance between the two obtained sequence chunks is typically 200-600 bp.

The Illumina system generates sequence reads of the same length from both ends, whereas the second read from SOLiD is shorter. The 454 system currently does not support paired-end sequencing of fragment libraries. However, the 454 system uses a mate-pair sequencing library approach which yields *de facto* a paired-end library [88] (Figure 3.1).

3.1.3 Mate-pair sequencing

In mate-pair sequencing, a large DNA-fragment, typically ~5000 bp, is circularized so that the 5'-end and the 3'-end of the initial DNA-fragment are next to each other. The circularized DNA is then fragmented and the linked ends are isolated and sequenced. Both the sequence of the 5'- and the 3'-end is known and the distance between the two tags should be approximately the predefined distance [88, 89].

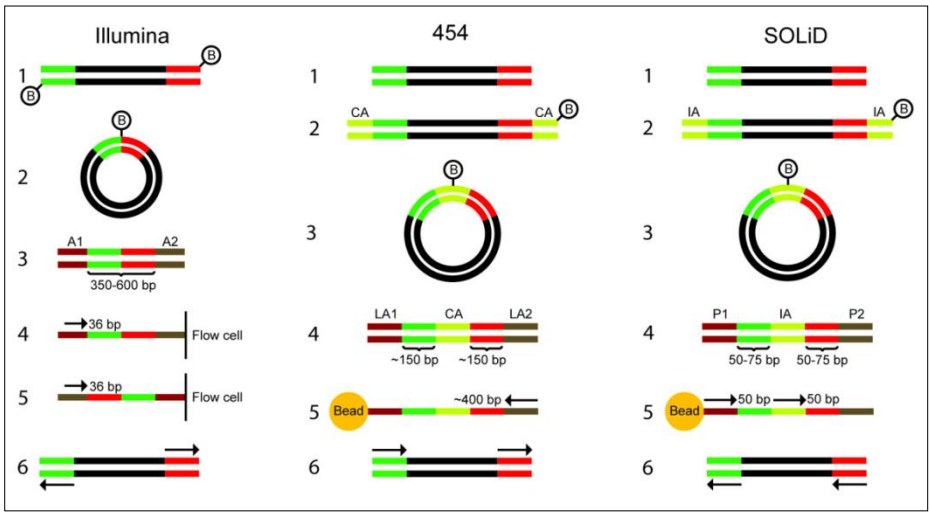


Figure 3.1: *Preparation of Illumina, 454 and SOLiD mate-pair libraries. Fragments are end-repaired using either unlabeled or biotinylated nucleotides. Depending on the technology, certain adapters are added. After circularization, the two fragment ends (green and red) become located adjacent to each other. The circularized DNA is then fragmented, and biotinylated fragments are purified by affinity capture. The isolated fragments are subsequently prepared for sequencing and eventually sequenced. B: biotinylated nucleotides, CA: circulization adapters, LA: library adapters, IA: internal adapter. Source: [88].*

3.2 Identifying genomic variation

3.2.1 Characterizing monogenetic diseases

Many hereditary characteristics are encoded in a single locus on a genetic level. Such hereditary characteristics are called monogenetic characteristics, Mendelian characteristics, single locus characteristics or one gene characteristics. Disease

phenotypes associated with monogenetic characteristics are commonly called monogenetic diseases.

Sickle-cell disease (or drepanocytosis) is a typical example of an autosomal recessive monogenetic disease caused by a single mutation in the *β-globin* gene causing a glutamic acid being substituted by a valine at position 6 [90]. Sickle-cell disease is characterized by red blood cells having an abnormal sickle shape (hence the name). In the Western world, life expectancies of persons having the disease are shortened. Life expectancies of persons being a heterozygous carrier is increased in malaria prone regions, due to the decreased survival of the malaria parasite in the sickle-shaped red blood cells, which is lowering the severity of the malaria symptoms, eventually increasing the chance of survival [91].

Certain monogenetic diseases were genetically characterized not that long after the discovery of the existence of the double-strand DNA-helix itself [92], indicating it is relatively easy to genetically characterize monogenetic diseases. The *CFTR* gene, for example, was identified as the disease causing gene in Cystic Fibrosis (CF) already back in the 1960s [93-95]. With the progression made in next-generation sequencing, identifying disease related variation became significantly easier over the recent years. Not only has there been significant progression in the identification of mutations in a research setting [96], but also in a diagnostic setting [97]. For example, women belonging to a high risk group for breast cancer (e.g. presence of breast cancer family history) are often genetically screened. Two important breast cancer genes, *BRCA1* and *BRCA2*, are typically sequenced using Sanger sequencing to verify whether known breast cancer associated mutations are present [98].

3.2.2 Characterizing polygenic diseases

Polygenic diseases (sometimes called complex or multi-factorial diseases) are caused by mutations in several genes; meaning these diseases are associated with different effects of different mutated genes each carrying one or more mutations. Polygenic diseases are often caused by a combination of genetics and environment where the genetic component increases the disease risk. Many inflammatory [99], cardiovascular [100], and mental diseases [101] are nowadays considered to be polygenetic. Asthma for example has been shown to be associated with many different genes. Latest research indicates at least 25 different genes might be involved [102].

With the advent of Sanger sequencing and especially with the transition from Sanger sequencing towards next-generation sequencing, it has become possible to identify genetic variation genome-wide. Genome-wide variation profiling has allowed complex polygenic diseases, such as cardiovascular diseases [103], to be genetically characterized. However, characterizing polygenic diseases remains a challenging endeavor, even with the tremendous sequencing outputs available present-day.

3.2.3 Characterizing DNA polymorphisms

A polymorphism is a genetic variation (of a certain locus) occurring between two or more individuals of the same species. In theory, a disease causing mutation can be considered a polymorphism as well, however, the term polymorphism is often reserved for cases where the variation is not disease causing.

A SNP (Single-Nucleotide Polymorphism) is a polymorphism with only one nucleotide differing between two or more individuals. Within a population, there is often a predominant allele and one or more minor alleles. SNPs usually occur more frequently in non-coding regions than in coding regions (where natural selection is not fixating the allele) [104].

Groups of different ethnicities vary in their SNP profile. A SNP that is common in one ethnic group may be much rarer in another. These characteristics are often exploited to study evolution and population genetics [105]. Furthermore, SNPs can affect how humans develop diseases [106] and respond to treatments [107, 108], concepts both of importance in personalized medicine.

In 2009, only a limited number of human genomes had been sequenced so far. However, next-generation sequencing allowed researchers to quickly identify SNPs on a genome-wide level. In one experiment, researchers identified approximately 3 million SNPs of which 13.6% were unknown. The experiment demonstrated the potential usefulness of next-generation sequencing technologies for personal genomics [50].

To capture, once and forever, the genetic variation present within the human population, the *1000 Genomes Project* was conceived. The 1000 Genomes Project aims to provide a deep characterization of human genome sequence variation as a foundation for investigating the relationship between genotype and phenotype. The con-

sortium aims to reach this goal by completely sequencing 1000 different human genomes and charting all the variation present across the different genomes [109].

3.2.4 Characterizing structural variation

However, other forms of variation are often present and are of significant importance as well. Besides SNPs, there are small insertion-deletion polymorphisms (indels), variable numbers of repetitive sequences, copy number variations (CNVs), structural rearrangements (e.g. inversions of certain regions), and complex combinations of the previous listed variations [110, 111]. SNPs and small indels can be detected by using single-end sequencing and using standard analysis methods.

Structural variation of the genome involving kilobase- to megabase-sized deletions, duplications, insertions, inversions, and complex combinations of rearrangements cannot be detected using single-end sequencing. To characterize such complex rearrangements, paired-end and/or mate-pair sequencing needs to be performed. A large deletion that occurred between the two sequenced ends will cause the two ends to map further from each other (e.g. 10 kb) on a reference genome than the intended distance between the two ends (e.g. 5 kb), making it possible to characterize the deletion. Insertions can be detected in a similar fashion. Large inversions for example can be detected by assessing the mapping orientation of both ends of the paired-end fragment. When both ends map on a different strand of the reference genome, an inversion point is located between the two ends. Complex rearrangements can be characterized using a combination of these structural variation features [112] (Figure 3.2).

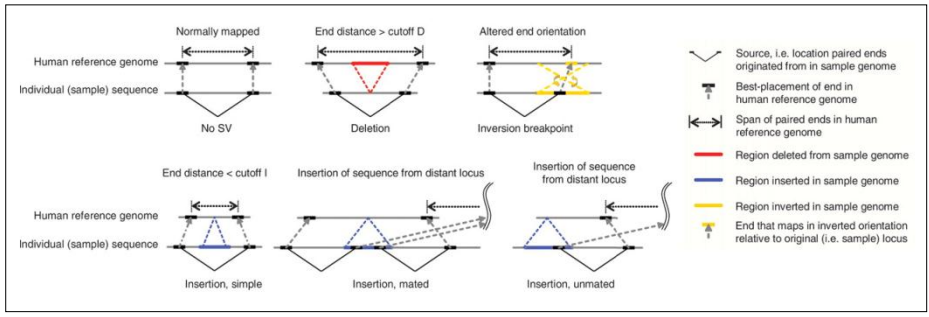


Figure 3.2: *Schematic overview of different types of structural variation that can be assessed using paired-end sequencing. Source: [112].*

3.2.5 Other uses of genomic sequencing

Genomic sequencing is not only used to characterize disease causing mutations or other forms of genetic variability. Genomic sequencing has been used to study complex processes beyond the scope of traditional research questions such as organelle capturing [113], primate evolution [114-116], migration patterns [117, 118], bacterial mutation rates [119], efficacy of vaccination [120], and location of cross-over regions [121].

3.2.6 Multiplex *PCR*

A technique often used to isolate genetic regions of interest prior to identifying variation is the so-called multiplex PCR [122]. A multiplex PCR consists of multiple primer sets within a single PCR mixture which produce different amplicons. By targeting multiple genes at once, several fragments can be amplified in a single PCR that otherwise would require several times the reagents and more time to perform. However, additional challenges arise; annealing temperatures for each of the primer sets must be optimized to work correctly within a single reaction.

3.3 Identifying transcriptional variation

Up until recently, gene expression was qualitatively or semi-quantitatively characterized using a variety of relatively low throughput methods that quantify RNA levels. Next-generation sequencing has enabled a more detailed quantitative view on the expression level of individual genes.

3.3.1 Microarrays

Microarrays contain a set of microscopic probes attached to a solid surface. Each of these probes has a specific sequence and corresponds to a short section of a gene. The probes are used to hybridize target fragments (cDNA or cRNA) from a sample to be analyzed. Probe-target hybridization is usually detected and quantified by detection of fluorophore- or chemiluminescence-labeled targets to determine the relative abundance of the target sequence in the sample.

Early microarrays contained a limited amount of cDNA probes spotted on a paper surface [123]. The technology was improved and eventually miniaturized in 1996

[124]. Currently, microarray probes are spotted on a variety of surfaces such as glass or silica and manufactured by specialized companies such as Affymetrix or Illumina. Microarrays can be used in a range of different experiments such as profiling the expression pattern of a certain sample or determining differential expression between different samples.

3.3.2 Tag sequencing

The major drawbacks of microarrays are the need for *a priori* knowledge in regards to the probe sequences and the limited quantitative abilities of the technology. To overcome the aforementioned limitations, SAGE (Serial Analysis of Gene Expression) was developed. SAGE is a methodology wherein mRNA is sequenced, rather than hybridized to predefined probes. Small sequence tags (~20 bp) are extracted from each mRNA-molecule and the tags are subsequently concatenated to form a long chain. The concatemer is then sequenced (using for example Sanger sequencing), the individual gene tags are counted, and genes can be accurately quantified using these tag counts [125] (Figure 3.3).

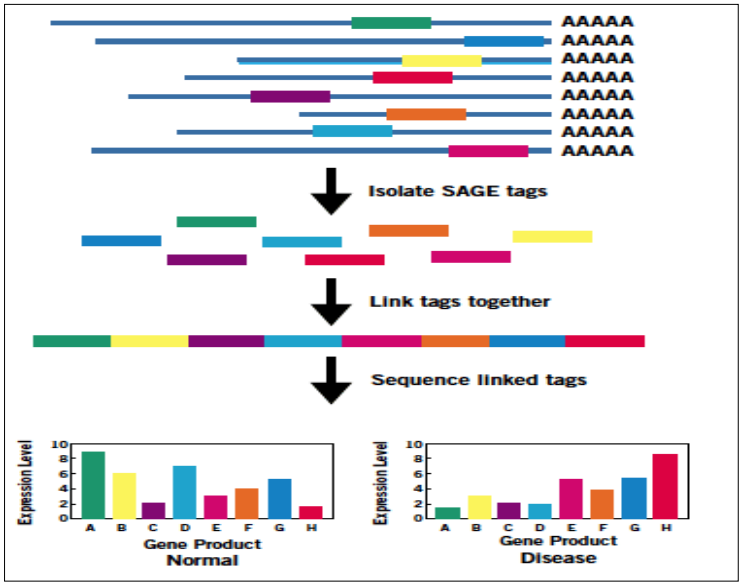


Figure 3.3: *Schematic overview of the SAGE technique. Small sequence tags originating from mRNAs are concatenated and subsequently sequenced. Each tag represents a single reference gene and the frequency of a certain tag is representative for the expression level of the corresponding gene. Source: Life Technologies.*

The mRNA-fragments are sequenced rather than hybridized, thus sequences do not need to be known *a priori*, meaning novel genes can be discovered. Quantifying gene expressions is more exact in this technique because it involves counting the number of transcripts whereas spot intensities in microarrays fall in non-discrete gradients and are prone to background noise. Several variants of the technology have been developed over the years such as LongSAGE [126] and SuperSAGE [127]. SuperSAGE uses the phage P1 EcoP15I endonuclease to cut 26 bp long sequence tags from the mRNA-molecules, expanding the tag-size by 6 bp in comparison with the older SAGE technique. The longer tag-size allows for a more precise allocation of the tag to the corresponding transcript [128].

Unlike SAGE, in which tags originate from random parts of transcripts, CAGE (Cap Analysis of Gene Expression) produces a snapshot of the 5'-ends of the mRNA transcriptome. Small fragments (~20 bp) from the beginning of mRNAs (5'-ends of capped transcripts, hence the *cap analysis*) are extracted, reverse-transcribed to cDNA, and sequenced. This technique is primarily used to locate exact transcription start sites in the genome. This knowledge in turn allows a researcher to investigate promoter structure necessary for gene expression [129].

SAGE and its variants are viable alternatives to overcome the limitations inherent to the microarray technique. However, both techniques use small tags of longer mRNAs to quantify the expression profile of a certain sample, hence these techniques offer no insights into the exact sequences of the assessed mRNAs. EST (Expressed Sequence Tag) sequencing offers more detailed insights, but at the cost of lower throughput. An EST results from one-shot sequencing of a cloned mRNA (i.e. several hundred base pairs of sequence starting from the 3'-end of a cDNA). The resulting sequence is a relatively low quality fragment whose length is limited by the sequencing length capacities of the platform that is used. Because these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes [130].

3.3.3 Expression profiling using NGS

In the aforementioned tag sequencing techniques, only a limited part of an mRNA-molecule (the tag) is used to quantify the relative abundance of an mRNA-molecule in a certain sample. The choice to use only a short tag was made mainly for practical reasons as it was the only feasible approach given the throughput and cost of the existing Sanger sequencing technology.

The development of NGS technologies changed the transcriptomics research field dramatically. The combination of next-generation sequencing and RNA-sequencing, RNA-seq, enables such a high throughput at such a low cost, making it possible to quantitatively characterize a whole human transcriptome. Rather than sequencing short tag sequences (as is the case in SAGE or CAGE) or sequencing low quality long sequences (as is the case in EST sequencing), a complete transcriptome can be sequenced in a single comprehensive effort.

RNA-seq has recently been used to characterize expression profiles [131], assess differential expression [132-134], detect fusion genes [135-137], detect post-transcriptional exon shuffling [138], analyze alternative splicing [139-141], and assess RNA editing [142, 143].

Advances in sample preparation have expanded the already broad range of RNA sequencing experiments that can be carried out. It is important to note that next-generation sequencing instruments can only process double stranded DNA-templates; hence specific RNA library preparation protocols are available on the different sequencing platforms. Recently, ingenious library preparation protocols were designed to enable directional RNA sequencing (i.e. discriminate between sense and antisense transcription) to be carried out [144, 145]. Furthermore specific library preparation protocols enable the miRNA profile of a sample to be characterized [146, 147] or non-coding RNAs (ncRNAs) to be discovered [148].

However, the existing next-generation sequencing methods typically require RNA to be reverse-transcribed to cDNA before it is actually sequenced. This conversion step has been shown to introduce biases and artifacts that may interfere with the proper characterization and quantification of transcripts. To enable an unbiased quantification of transcriptomes, Helicos BioSciences has developed the single molecule Direct RNA Sequencing (DRS) technology. DRS can sequence RNA molecules directly in a high-throughput fashion without RNA conversion to cDNA or other biasing sample manipulations such as ligation and amplification [70, 71].

Next-generation sequencing typically yields plenty of small fragments of a single mRNA molecule. Different strategies exist to aggregate the small fragment data and convert the aggregated data to a single expression level for each gene. Nanopore sequencing looks promising as complete mRNA molecules will be able to be sequenced. Full mRNA sequencing will enable researchers to have a very detailed view on splicing events and other RNA rearrangements.

3.3.4 Ribosome profiling sequencing

RNA profiling techniques such as microarrays and RNA-seq can (semi)-quantitatively characterize mRNA expression levels. However, mRNA levels are an imperfect proxy for protein production because mRNA translation is subject to extensive regulation (by for example miRNAs [149]).

Recently, a ribosome profiling technique was devised which is able to detect active ribosomes transcribing mRNA by using a sequencing-based approach. The position of a translating ribosome can be precisely determined by using the fact that a ribosome protects a ~30 nucleotides region on its mRNA template from nuclease digestion. By applying nuclease digestion, only the translational active regions are retained. Subsequent sequencing of the retained fragments allows comprehensive high-precision measurements of *in vivo* translation with subcodon precision [2] (Figure 3.4).

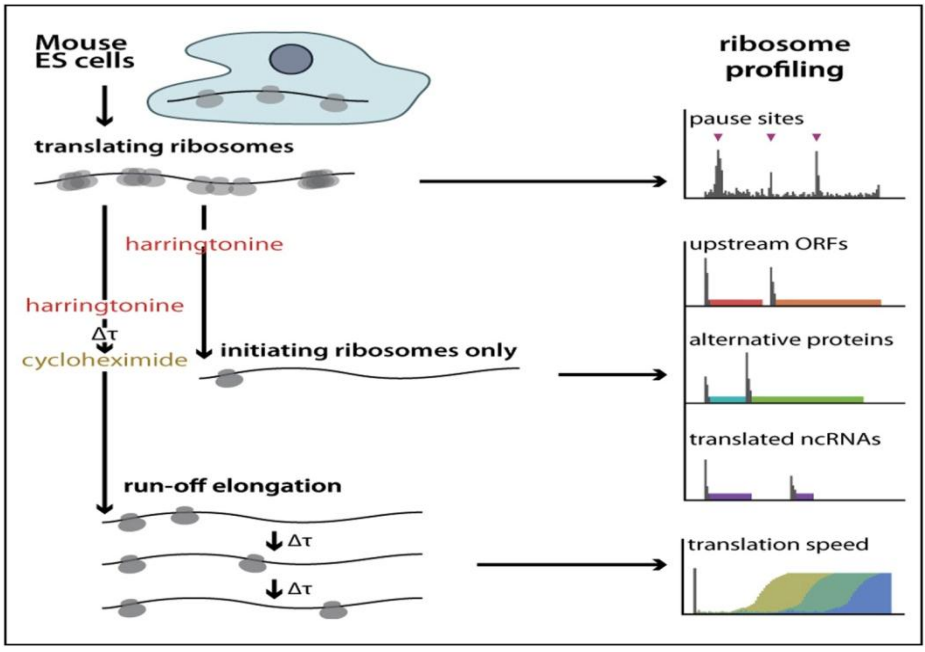


Figure 3.4: *Schematic overview of the ribosome profiling sequencing technique. Ribosome-protected mRNA fragments are sequenced to provide genome-wide maps of protein synthesis as well as a pulse-chase strategy for determining rates of translation elongation. Harringtonine causes ribosomes to accumulate at sites of translation initiation, making it possible to identify initiation sites. Source: [150].*

3.4 Identifying DNA methylation

DNA-methylation is a process wherein a methyl group is added to a genomic residue. This process has been known for a long time and was already described not many years after the initial discovery of DNA [151]. Although carbon-5-cytosine methylation (5mC) in a CpG context is considered to be the most important form of DNA methylation in humans, other forms of methylation such as carbon-5-cytosine hydroxy-methylation (5hmC) [152] and non-CpG methylation [153] have already been described. Methylated cytosines spontaneously mutate to thymines over time and most CpGs are methylated in humans [154], hence CpG dinucleotides are underrepresented in the human genome [44]. Unmethylated CpGs are frequently clustered together in CpG islands which are mainly located at the 5' regulatory region of a gene [155].

DNA-methylation is considered to be one of the many important epigenetic mechanisms involved in cell regulation. It is present in healthy cells where it is involved in different processes such as tissue differentiation [156], aging [157], memory formation [158], and X chromosome inactivation [159]. Gene promoter hypermethylation (mostly CpG island hypermethylation) is generally associated with gene silencing [160]. Both hypomethylation of oncogenes [161] and hypermethylation of tumor suppressor genes [162] are associated with cancer development. Furthermore, the methylation profile of a cell can be used as diagnostic [163], prognostic [164] or treatment response marker [165].

3.4.1 Bisulphite sequencing

Since the first discovery of DNA methylation, different methods have been used to detect and characterize methylation patterns. Bisulphite sequencing is a widely used high-accuracy, low-throughput method and allows individual methylated cytosines to be identified. Treatment of DNA with bisulphite converts cytosines to uracils, but does not affect 5mCs. Thus, bisulphite treatment introduces specific changes in the DNA sequence that depends on the methylation status of individual cytosine residues, yielding single-nucleotide resolution methylation information.

Bisulphite conversion is a three step process (Figure 3.5). In the first step, cytosine undergoes a reversible sulphonation to yield cytosine-6-sulphonate. Then, the cytosine-6-sulphonate undergoes an irreversible hydrolytic deamination to yield uracil-6-sulphonate. In a final step, the sulphonate group is removed in a reversible desul-

phonation to yield uracil. When the bisulphite-treated DNA is amplified via PCR, the uracil is amplified as thymine. 5-methylcytosines are not converted by the bisulphite process, hence methylated cytosines are amplified as cytosines [166].

After the DNA has been converted using the bisulphite treatment, the converted DNA can be sequenced. Two distinct sequencing approaches are possible. Either a certain region of interest is first isolated (using specific PCR primers or a capture approach) which is followed by sequencing [167] or the complete bisulphite converted DNA is sequenced in a genome-wide approach [168]. The former is called targeted bisulphite sequencing, the latter genome-wide bisulphite sequencing. In both approaches, the underlying principle is the same: those cytosines which are read as cytosines after sequencing represent methylated cytosines, while those that are read as thymines represent unmethylated cytosines in the genomic DNA.

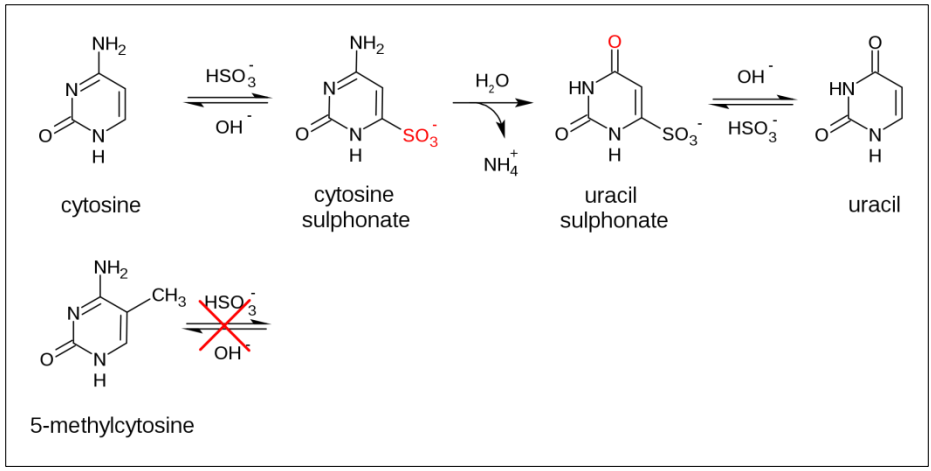


Figure 3.5: *Chemistry of bisulphite treatment of genomic DNA. Position C6 of cytosine is reversibly sulphonated. Cytosine sulphonate is irreversibly hydrolytic deaminated at position C4. 5-sulphonate-uracil is subsequently desulphonated to yield uracil. Methylation at the C5 position of the cytosine impedes sulphonation at the C6 position.*

3.4.2 Methylation capturing sequencing

Although bisulphite sequencing offers a tremendous resolution, and is thus considered the golden standard for cytosine methylation sequencing, there are drawbacks to the technology. The required coverage to accurately and quantitatively deter-

mine the methylation state of all the cytosines in the genome is too high, making it impractical to apply this method to study multiple biological samples.

Alternative approaches are based on specific enrichment of methylated portions of the genome which are captured from a mixture of methylated and unmethylated DNA-fragments (hence the name methylation capturing sequencing). One of the first methylation capturing techniques developed was *methylated DNA immunoprecipitation* (MeDIP). MeDIP is a technique wherein methylated DNA-fragments are isolated from a pool of DNA-fragments using an immunoprecipitation (IP) via an antibody raised against 5-methylcytosine [169]. At the time the technique was developed, next-generation sequencing was not yet available and consequently other methods were used to identify the isolated DNA-fragments (such as DNA microarrays [169]).

Once next-generation sequencing technology became available around 2008, next-generation sequencing and MeDIP were quickly combined in a approach named MeDIP-seq [170]. However, the antibody is raised against 5-methylcytosine and will consequently capture fragments containing a single 5-methylcytosine [171] (Figure 3.6).

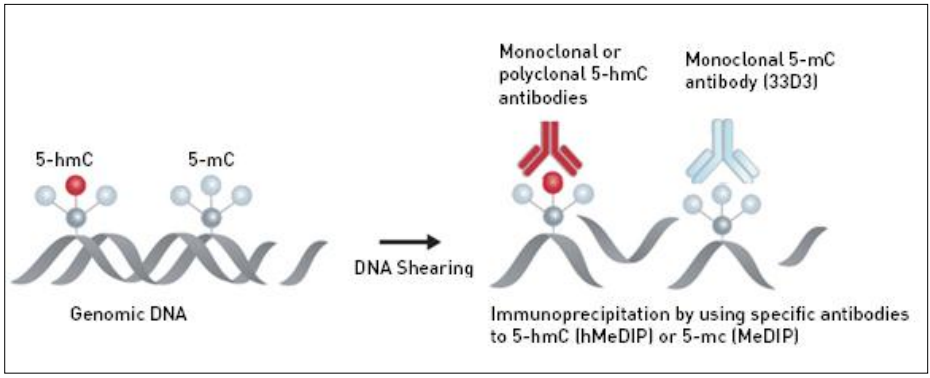


Figure 3.6: *Schematic overview of the MeDIP technique to isolate methylated and hydroxy-methylated DNA fragments. After isolation of the fragment of interest, sequencing identifies the methylated region. Source: Diagenode, Belgium.*

To alleviate the problems observed in MeDIP-seq, a novel technique was developed which makes use of a natural human protein domain. Capturing of methylated DNA-fragments using the Methyl-CpG-Binding Domain (MBD) of the human *MBD2* gene [172] was combined with next-generation sequencing in a technique named

MiGS (MBD-isolated Genome Sequencing). MBD binds with increasing affinity to multiple 5-methylcytosines in close proximity and will precipitate more biologically relevant methylated fragments as opposed to sporadically methylated CpGs of uncertain biological relevance as is the case in MeDIP [171, 173].

In principle, any molecule that interacts with specific DNA sequences and where an antibody can be raised against can be isolated using the chromatin immunoprecipitation technique (ChIP). The DNA sequence bound to the isolated molecule can be isolated and sequenced. For example, this technique has been used to identify transcription factor binding sites [174].

3.4.3 Direct methylation sequencing

Recent developments in sequencing have enabled researchers to directly sequence modified nucleotides. Rather than having to apply a capturing step or a destructive bisulphate treatment step, which both have their limitations, methylated bases can be sequenced directly using third-generation sequencing technologies. As mentioned before, both the Pacific Biosciences (Figure 3.7) and the Oxford Nanopore Technologies platforms have shown to be able to sequence both methylated and hydroxy-methylated bases [12, 74, 75, 175].

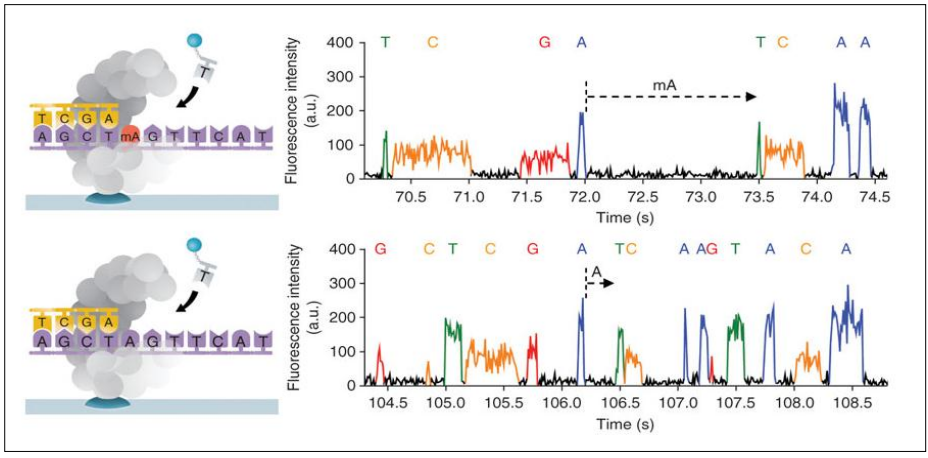


Figure 3.7: *Schematic overview of polymerase synthesis of DNA strands containing a methylated (top) or unmethylated (bottom) adenine (left). Typical SMRT sequencing (Pacific Biosciences) fluorescence traces (right). Dashed arrows indicate the interpulse duration, which is about five times larger for the methylated adenine (mA) compared to the native adenine (A). Source: [175].*

4 Analyzing next-generation sequencing data

The type of output data generated by a sequencing instrument is a constant, i.e. a set of nucleotide strings, regardless of the experiment being carried out. The type of experiment only has an influence on the downstream data processing. This makes the primary data analysis, converting the raw sequencing data to sequence strings, a constant throughout all the different types of experiments. As a consequence, we will not further discuss primary data analysis in this dissertation. However, secondary data analysis, analysis of the sequence strings, is of utmost importance.

There are two distinct types of DNA sequencing experiment classes. In the first type of experiments, an unknown genome is sequenced, i.e. sequences are obtained from a previously unknown species. In the second type of experiments, a known genome is re-sequenced, i.e. sequences from a species with a known reference genome are obtained.

Typically an unknown genome is sequenced to shed light on the genomic background of a species of interest. The individual sequence fragments generated by the sequencing instrument are assembled into bigger fragments (genes, chromosomes, plasmid, and mitochondrial genome) in a process referred to as *genome assembly*.

On the other hand, a known genome can be re-sequenced to identify certain characteristics of the genome under investigation. Identification of genomic variation is an obvious example of re-sequencing where the obtained sequence data is compared to a reference sequence. However, RNA sequencing and methylation sequencing can technically be seen as re-sequencing as well, as the obtained sequencing data is also compared to a reference sequence to both localize and quantify transcription and methylation.

4.1 Genome assembly

4.1.1 De novo assembly

De novo genome sequencing involves an assembly step to merge overlapping sequence reads into longer fragments. In an ideal situation, individual fragments can

be assembled into ungapped genes or chromosomes. Currently, myriad genome assembly programs are available.

Newbler (proprietary, Roche) is a program which is specifically designed to *de novo* assemble long-read sequences generated on a 454 sequencing instrument. Velvet [176] is one of the most popular programs to *de novo* assemble short-read sequences generated on an Illumina or SOLiD instrument. MIRA [177] is often described as the Swiss army knife of sequence assembly as it can assemble 454, Ion Torrent, Illumina, Sanger, and Pacific Biosciences data. MIRA is one of the few assembly programs that can assemble combinations of different data types, known as a hybrid assembly. Most of these genome assembly packages make use of a technique called *De Bruijn graphs* [178] to identify and organize overlaps between the sequenced reads and eventually merge the reads together into a bigger fragment (Figure 4.1).

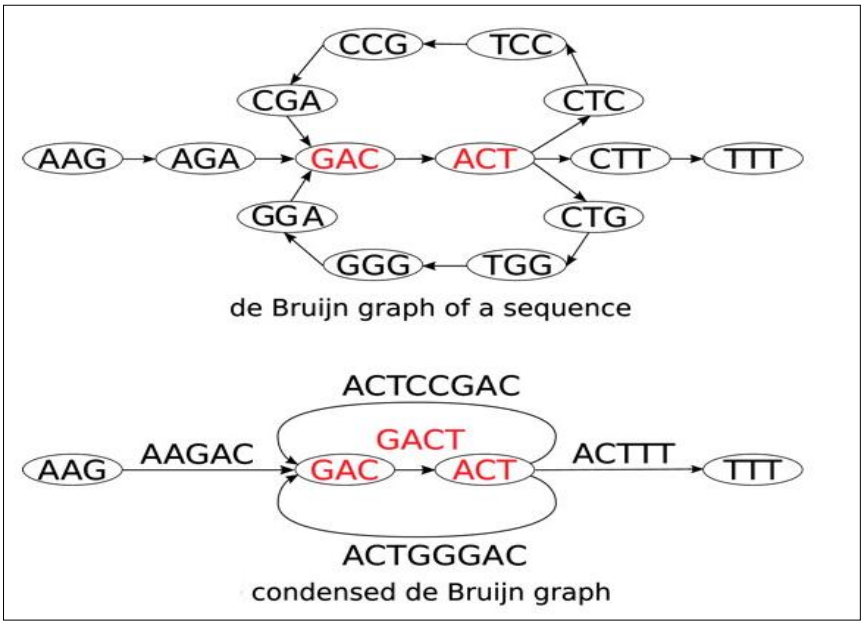


Figure 4.1: *De Bruijn graph for the AAGACTCCGACTGGGACTTT sequence. Note that GACT is a repeat (shown in red). The de Bruijn graph of a sequence contains a vertex for every* k*-mer in the sequence and an edge for every pair of consecutive (overlapping)* k*-mers in the sequence (top). The condensed de Bruijn graph replaces all paths containing non-branching vertices by a single edge labeled by the sequence that generated the path (bottom). Source: [179].*

4.1.2 Guided assembly

Besides a *de novo* assembly, which uses no prior knowledge, there are assembly methodologies, called *guided assemblies* or *mapping assemblies,* which make use of a reference genome [180]. Sequence reads are first mapped onto the related species and then assembled in a second step. These steps can be carried out separately [181] or in a single comprehensive step [182].

4.2 Aligning sequencing data

Next-generation sequencing datasets typically generate datasets containing a huge amount of sequences, typically called sequenced reads. Mapping (or aligning) sequenced reads back to a reference genome is one of the most important steps when analyzing NGS re-sequencing data.

The first algorithm specifically designed to align two different protein or nucleotide sequences was the Needleman-Wunsch algorithm, developed in 1970 by Saul Needleman and Christian Wunsch. The developed algorithm uses *dynamic programming* to make a global alignment (i.e. align the two sequences as complete as possible) and is relatively slow [183]. The dynamic programming approach allows the bigger problem, i.e. align the two sequences, to be divided into smaller subproblems by separately aligning each base of the sequence. Later, the Needleman-Wunsch algorithm was modified to allow a local alignment (i.e. align a portion of one sequence to a larger reference sequence). This lead to the development of the Smith-Waterman algorithm in 1981 [184].

The dynamic programming algorithms designed in the 1970s and 1980s generated the best possible results, but were very slow. Once datasets became relatively large, a clear need for better algorithms arose. To tackle this problem, a fast heuristic algorithm was developed in 1990. BLAST (Basic Local Alignment Search Tool) was designed to perform well on the bigger datasets typically generated by Sanger sequencing. BLAST enabled a researcher to compare a query sequence with a set of reference sequences, and identify reference sequences that resemble the query sequence above a certain threshold [185]. BLAST was used intensively (the BLAST paper was the most cited scientific paper in the 1990s) and improved [186, 187] until datasets again became too big to be processed within a reasonable timeframe.

Large parallel Sanger sequencing efforts were being carried out in the beginning of the 21st century leading to the release of draft genomes of several vertebrate species. To assist in the annotation of these genomes, there was a clear need for an algorithm that could align mRNA and DNA sequences to these large genomes. Jim Kent developed BLAT (BLAST-Like Alignment Tool) at UCSC which was much faster than BLAST and could perform spliced alignments of RNA on DNA as well [188].

Since the 454 sequencing technology has been introduced to the market, the need for algorithms that efficiently map huge amounts of reads onto reference genomes has increased rapidly. Later high-throughput sequencing methods such as Illumina and SOLiD sequencing have intensified the demand for efficient mapping algorithms.

The development of mapping methods depends on specifications and error models of the respective technologies. Unfortunately, little is known about specific error models, and models are likely to change as manufactures are constantly modifying chemistry, machinery, and processing software. Recent investigations on the accuracy of the different platforms have shown that 454 reads are more likely to include insertions and deletions while Illumina reads typically contain mismatches [57, 79].

Currently available mapping programs are specifically designed to allow for mismatches when aligning the reads to the reference genome. Most of the programs, e.g. MAQ [189], SOAP [190], and ELAND (proprietary), use seeding techniques that gain their speed from pre-computed hash look-up tables. In an effort to reduce the memory requirements, programs were developed which use BWT (Burrows-Wheeler Transform) compression [191]. These BWT programs are amongst the most popular mapping programs used and include Bowtie [192], BWA [193], and SOAP2 [194].

The matching models of MAQ, SOAP, Bowtie, and ELAND focus on mismatches and largely neglect insertions and deletions. Indels are only considered during subsequent alignment steps but not while searching for seeds. With indels accounting for more than two thirds of all 454 sequencing errors, this is a major shortcoming for these kinds of reads. Only PatMaN [195] and BWA are able to handle a limited number of indels. To handle a large number of small indels or large indels, BLAT, despite its slow speed, is still a reliable option. Bowtie2 [196], the successor to the highly popular Bowtie, can align complex indels.

Some of these programs e.g. SOAP, MAQ, and Bowtie are specifically designed to map short Illumina or SOLiD reads. Longer sequences cannot be mapped by these tools. The length of the sequenced fragments has steadily been increasing for the past few years. The short reads generated on the recent Illumina or SOLiD platform nowadays have the length of the early long reads generated on the 454 platform, eventually converging the world of short and long read sequencing. BWA, typically designed to map short read sequences, was adapted to BWA-SW and now can handle long reads as well [197]. Bowtie2 can align sequences longer than 10,000 bp.

4.3 Detecting genomic variation

4.3.1 Detecting single nucleotide variation

Detecting single nucleotide variants (SNVs) using NGS is relatively easy when a reference genome is available and the sequencing depth is sufficiently deep to discriminate between real variants and sequencing errors. With these two conditions in mind, it is obvious that detecting variants in human samples is relatively straightforward as a very accurate reference sequence is available. How deep 'sufficiently deep' is, is sometimes a topic of debate, but 20-40x coverage is generally considered to be sufficiently deep [58].

However, although most of the sequence aligners are able of mapping sequences with errors (hence putative variants) to a reference sequence and pinpointing the variants, specialized packages are available that automatically list, filter, prioritize, annotate, and visualize detected variants. Frequently used SNP analysis packages include for example AVA (proprietary, Roche), SOAPsnp [190, 194], CLCbio Genomics Workbench (proprietary, CLCbio), Alamut (proprietary, Interactive Biosoftware), and NextGENe (proprietary, softgenetics). Most of these methodologies are also capable of detecting small (mostly 1 bp) indels.

4.3.2 Detecting structural variation

Using NGS it is not only possible to detect SNVs and small indels, but also copy number variation and larger rearrangements. After single-end reads are aligned to a reference genome, a density count can reveal regions which are more abundantly present in the genome compared to the reference and hence indicate duplication. In a similar fashion, deleted regions can be detected [198]. For example, both MrFast [199] and CNV-Seq [200] allow CNVs to be detected using NGS data.

Using paired-end and mate-pair data obtained by next-generation sequencing technology, complex rearrangements such as large insertions, deletions, inversions, inter- and intra-chromosal translocations can be detected [198]. The most advanced algorithms available, capable of detecting all these structural variations, are for example BreakDancer [201] and SVDetect [202]. Structural variation is often visually summarized in circular plots which are generated using specialized packages such as Circos [203] (Figure 4.2).

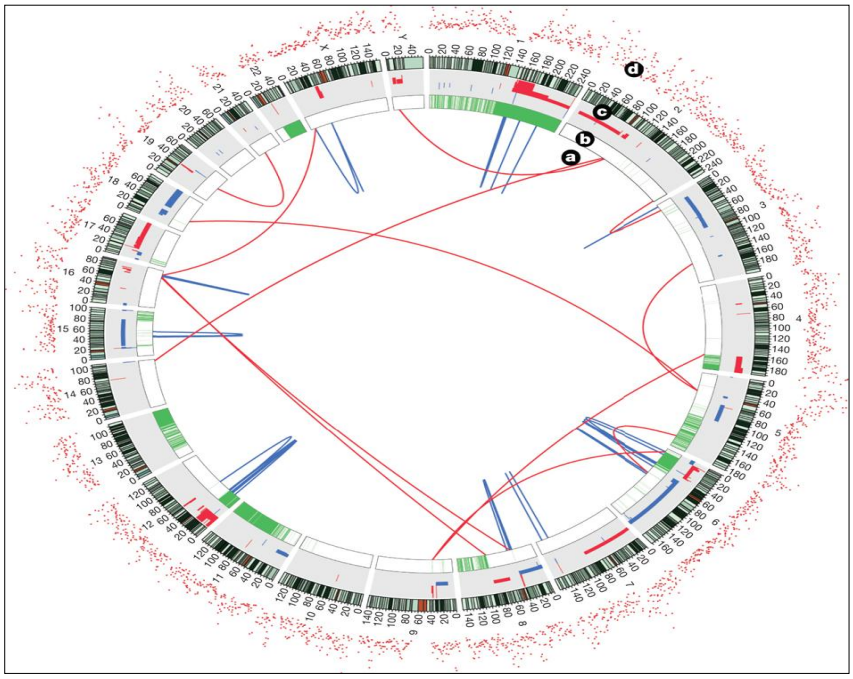


Figure 4.2: *Overview of the genomic variation in a lung cancer patient. a) Red lines indicate inter-chromosomal structural variations; blue lines represent intra-chromosomal structural variations. b) Regions of loss of heterozygosity and allelic imbalance as derived from an Affymetrix SNP 6.0 array. c) Copy number variation profile as derived from an Agilent array. d) Each red dot represents a high-confidence SNV. Source: [204].*

4.4 Analyzing methylation data

4.4.1 Analyzing bisulphite treated DNA using NGS

Mapping bisulphite converted reads to a large reference genome (such as the human genome) brings substantial computational challenges. Unmethylated cytosines are eventually converted to thymines after applying bisulphite treatment, unlike methylated cytosines which remain unchanged by the treatment. Thus, thymines in the sequenced reads could be mapped against either cytosines or thymines in the reference genome, but not the other way around, making cytosine and thymine mapping asymmetric This increases the search space significantly, making mapping bisulphite converted reads more challenging [205].

In silico bisulphite converting the reference genome is an efficient and fast method at first sight. However, an *a priori* assumption has to be made about the methylation pattern. Earlier bisulphite mapping strategies naively assumed that a sample either was methylated (and consequently all cytosines remained unconverted) or unmethylated (and consequently all cytosines were converted to thymines) [206]. Using this strategy, the sequenced reads are mapped on both the unconverted reference genome and the completely converted reference genome allowing only a limited number of mismatches. This method cannot efficiently process reads containing both unmethylated and methylated cytosines as the number of mismatches in reads originating from a sequence with both methylated and unmethylated cytosines will be too high when mapped on both the converted and the unconverted reference genome. Another strategy was proposed where the sequenced reads are converted to position–weight matrices and an alignment is carried out in probability space. Due to its computational complexity, the approach is not practical unless the reference genome is small [168, 207].

To overcome these limitations, different bisulphite mapping programs have been developed. However, the underlying principle used in most of the different programs is similar to the ineffective method described in the previous paragraph. Either the reference genome or query sequence (the sequenced read) is *in silico* bisulphite modified and then aligned using an existing mapping program. By tracing back the modification, it is possible to determine the original methylation status of the aligned sequence read.

Bismark [208] uses Bowtie to align sequenced reads to a converted reference genome. BSMAP is based on a modified version of SOAP in which the reference genome is converted to a series of typically octamer seeds on which hashing and fast lookup methods can be applied to attain efficient performance. BSMAP generates C/T and G/A converted seeds for the reference genome in which all possible methylation patterns exist for each seed. A bit-mapping strategy is applied within the program to highlight mismatches from methylation and sequencing errors [205, 209]. RMAPBS [210] is a modification of the RMAP mapping program and uses an advanced seed and hashing strategy similar to the strategy used in BSMAP.

CASA (personal communication, Trooskens) uses a completely different strategy. CASA uses a dynamic programming approach similar to the Smith-Waterman algorithm to efficiently align CG/TG di-nucleotides in the sequenced reads to a CG on a reference sequence (which needs not to be converted). The scoring matrix is optimized in such a way that converted (i.e. thymines) and unconverted cytosines receive the same score when aligned to a cytosine on the reference sequence.

4.4.2 Analyzing methylation capturing sequencing

After methylated DNA-fragments are captured and subsequently sequenced using NGS, these fragments are aligned to a reference genome. Regions with high coverage are considered to be highly methylated, thus efficiently captured away before sequencing; regions with no coverage are considered unmethylated. Several strategies are available to compare different samples and identify differential methylation patterns.

Some of the currently applied methodologies to identify differential expression have their origin in the field of RNA-sequencing analysis where coverage of a certain region is used as a proxy for the expression level of the region under investigation. Although the methodology is not optimal, due to the different profile of the signal, BaySeq [211] for example can be used to identify differentially methylated regions (personal communication, Denil; De Meyer).

Other strategies merely try to identify methylation patterns and convert the semi-continuous signal to a more discrete signal. In practice this means that a profile with peaks and troughs (according to the coverage) is converted to defined blocks by *peak callers*. What remains of the semi-continuous signal is a series of called peaks each with a certain width and height. These peaks can then be used as a binary sig-

nal (i.e. the region is either methylated or not methylated) in further analyses. Several of these peak caller algorithms exist, many of them having an origin in transcription factor based ChIP-seq analyses, such as PeakRanger [212]. To overcome the limitations inherent to repeated peak calling (e.g. called peak regions differing between samples) researchers have grouped methylation capturing data originating from different experiments and developed algorithms to divide the genome into discrete regions by calling peaks on the grouped data (personal communication, Trooskens). This has enabled researchers to compare the methylation profile of many different samples by assessing the methylation status of the functional regions.

4.5 Analyzing RNA sequencing data

Gene expression levels can be deduced via RNA-seq by means of the sequencing coverage. Sequencing coverage is the average number of reads representing each nucleotide in the sequence under investigation (i.e. a gene). For example, a hypothetical 10,000 bp gene covered by 10,000 reads with a length of 100 nucleotides will have a 100X coverage. This effectively means that the more a certain genomic region is sequenced in the RNA-seq experiment, the higher the expression is. When the transcriptome is fragmented prior to cDNA synthesis, the number of reads corresponding to the particular exon normalized by its length *in vivo* yields gene expression levels which correlate with those obtained through qPCR [213].

Once expression levels are determined per gene, normalization is required because lane-specific coverage for example can vary. The standard normalizing techniques for microarray data (e.g. quantile normalization) do not apply. However, there are few standard normalizing techniques for RNA-seq. The most popular approach RPKM [214], normalizes the gene counts in each sample by gene length and the total number of mapped reads in that sample [215]. BaySeq [211] goes a step further and uses a Bayesian method to identify differential expression.

Specialized programs have been developed to allow more detailed analyses to be carried out. TopHat [216] is a fast splice junction mapper for RNA-seq reads. TopHat aligns RNA-seq reads to a reference genomes using Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Cufflinks [217] assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-seq samples. It accepts aligned RNA-seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates

the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

Recently, several *de novo* strategies have been developed. A *de novo* transcriptome assembly strategy does not use a reference genome allowing transcriptomes to be determined of species whereof the genome sequence is not known. It also allows unknown transcripts to be discovered and 'exotic' splicing events to be picked up [218]. Most of the *de novo* transcriptome assembly programs use, just as is the case in genome assembly, de Bruijn graphs [178] to assemble the transcriptome. Among the most popular assembly programs are Abyss [219] and Trinity [220]. An extensive overview of *de novo* transcriptome assembly programs is given in the review paper by Martin and Wang [218].

Part 2: Identifying genomic variation using second-generation sequencing

Adapted from:

Joachim De Schrijver, Kim De Leeneer, Steve Lefever, Nick Sabbe, Filip Pattyn, Filip Van Nieuwerburgh, Paul Coucke, Dieter Deforce, Jo Vandesompele, Sofie Bekaert, Jan Hellemans, Wim Van Criekinge (2010), **Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline,** *BMC Bioinformatics* 11:269

Joachim De Schrijver et al. (2012), **Advancing the Variant Identification Pipeline,** in preparation

5 Analysing 454 resequencing experiments: Variant Identification Pipeline

5.1 Abstract

5.1.1 Background

Next-generation amplicon sequencing enables high-throughput genetic diagnostics, sequencing multiple genes in several patients together in one sequencing run. Currently, no open-source out-of-the-box software solution exists that reliably reports detected genetic variations and that can be used to improve future sequencing effectiveness by analyzing the PCR reactions.

5.1.2 Results

We developed an integrated database oriented software pipeline for analysis of 454/Roche GS-FLX amplicon resequencing experiments using Perl and a relational database. The pipeline enables variation detection, variation detection validation, and advanced data analysis which provides information that can be used to optimize PCR efficiency using traditional means. The modular approach enables customization of the pipeline where needed and allows researchers to adopt their analysis pipeline to their experiments. Clear documentation and training data is available to test and validate the pipeline prior to using it on real sequencing data.

5.1.3 Conclusions

We designed an open-source database oriented pipeline that enables advanced analysis of 454/Roche GS-FLX amplicon resequencing experiments using SQL-statements. This modular database approach allows easy coupling with other pipeline modules such as variant interpretation or a LIMS system. There is also a set of standard reporting scripts available.

5.2 Background

Recent DNA sequencing technology, so-called next-generation sequencing (NGS) technology, enables researchers to read a number of DNA sequences that is several orders of magnitudes bigger and at a cost that is several orders of magnitude small-

er than the previous generation DNA sequencing technologies. The cost of determining the human genome was estimated at $2.7 billion for the IHGSC genome and at $300 million for the Celera genome. Recently several human genomes were sequenced in about 1.5 months at a cost that is around $1.5 million [50, 51].

Large-scale parallel pyrosequencing from 454/Roche generates hundreds of thousands sequenced DNA reads within a matter of hours [52]. The latest version of the sequencing technology (Titanium) enables a throughput of 0.4-0.6 gigabases per 10h run [53]. The amount of data to be analyzed keeps growing at an increasing speed. Other NGS platforms such as Illumina's Genome Analyzer (San Diego, CA, USA), Applied Biosystems' SOLiD (Foster City, CA, USA) and Helicos' HeliScope (Cambridge, MA, USA) generate more than 10 gigabases in a single multi-day run. It is estimated that by 2012 genome centers around the world will generate more data per year than the expected 15 petabytes per year that is produced by CERN's Large Hadron Collider [221].

Sequencing researchers agree that data-analysis of NGS projects is the biggest challenge to make the technology accessible for biologists around the world. The cost of sequencing might go as low as $1,000 per human genome, but it will still be a lost investment if the generated data cannot be adequately analyzed. Especially sequencing platforms with a very high throughput of read lengths as short as 35–40 nucleotides face challenges for genome assembly and annotation [222].

Nevertheless, NGS is expected to have an enormous impact on diagnostics and SNP discovery [223], provided there are tools available that make variant detection and interpretation of the sequenced data straightforward and automated. A good step towards standardization and data uniformity, especially in the short-read field, was taken with the development of the SAM/BAM format [224]. This format is currently supported by BWA [193] and Bowtie [192], but is not supported by long-read (typically 454 GS-FLX data) mappers such as BLAT [188]. However, with the recent release of BWA-SW [197] one might expect a more broad transition of long-read NGS towards SAM/BAM pipelines.

There are already some tools available to analyze 454 GS-FLX experiments, but mostly with a limited analysis spectrum. Most of the present tools are designed to do a specific task, and rarely offer broad spectrum analysis. The few integrated analysis tools offering broad spectrum analysis available are commercial packages (such as CLCbio Genomics Workbench, Genomatix or NextGENe).

Mosaik [225], for example, is one of the few broader analysis packages available and has cross-platform features which makes it indeed easy to work with. Unfortunately, it lacks some additional analyzing possibilities besides mapping and assembly that are useful in amplicon resequencing based diagnostics (for example coverage per amplicon and per patient or calculation of primer dimer frequencies).

Roche's Amplicon Variant Analyzer (AVA) software is specifically designed for analysis of amplicon resequencing experiments and is user-friendly but has some limitations. Advanced coverage analysis is of paramount importance in a diagnostic setting but is lacking in AVA. Data storage is also an issue using AVA as data is not stored in a structured way (i.e. flat files instead of storage in a database).

In this paper we describe an open-source database oriented pipeline that is capable of analyzing a single GS-FLX amplicon resequencing run automatically from receiving the raw data to generating custom (variation) reports within one day. Using the database approach, the pipeline is capable of analyzing a set of runs together in a meta-analysis. While sequence variations can be assessed using this pipeline, it also allows researchers in diagnostic labs to extract additional information (e.g. frequency of primer dimers). This data can then be used to optimize the amplicon PCR reactions and general sequencing settings to improve future sequencing runs. The pipeline, a manual, a testing dataset, and example reports are available on the web at http://athos.ugent.be/VIP_pipeline.

5.3 Implementation

The Variant Identification Pipeline (VIP) is developed in a modular way so that each module can run independently from another. Each module uses data stored in a relational database (MySQL) to do its task and stores its result again in the database. The pipeline is completely written in Perl and uses the *DBI* and *DBD::mysql* packages for database interactions, bioperl package to parse mapping reports and 5 custom Perl modules (*gs_flx_functions*, *gs_flx_references*, *gs_flx_seqIO*, *gs_flx_snp_report*, and *gs_flx_validation*).

A schematic overview of the Variant Identification Pipeline and the different pipeline modules is given in Figure 5.1.

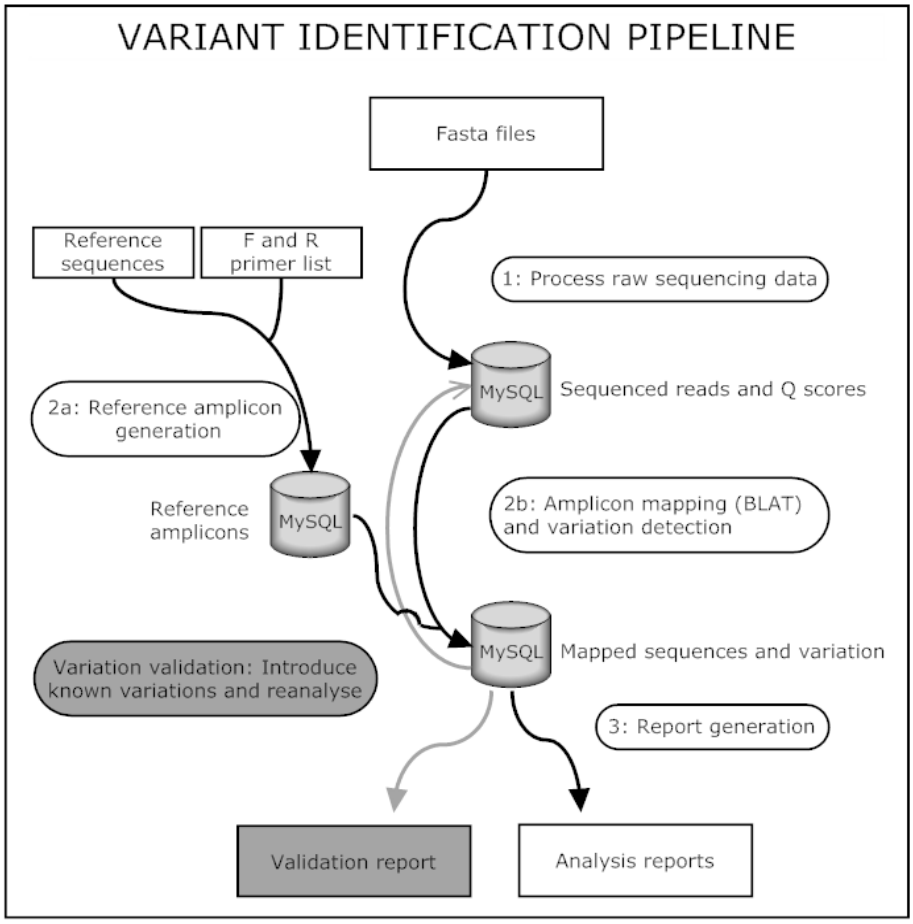


Figure 5.1: *Overview of the Variant Identification Pipeline (black arrows and white text-boxes) and the VIP Validator (grey arrows and grey text-boxes). The analysis pipeline consists of 4 modules. 1) Raw sequences are extracted from the FASTA files generated by the GS-FLX sequencer and processed into sequenced amplicons and additional information. 2a) Reference amplicons are generated using a list of reference sequences and the list of primers. 2b) Mapping is carried out with BLAT using the reference amplicons and the sequenced amplicons. Variations are detected and stored in the database. 3) The requested reports are generated. The VIP Validator introduces additional variation in the sequence reads and reanalyzes those sequences to validate the pipeline for that specific variation.*

5.3.1 Processing the raw sequence data

The raw sequencing step uses both the FASTA (sequenced reads) and QUAL (positional quality scores of the sequenced reads) files generated by the 454/Roche GS-FLX sequencer as input. Furthermore, the pipeline uses a list with multiplex identifier (MID) sequences and linker sequences. The multiplex identifier or MID is a short bar-code sequence used to label samples/patients when multiplexing, the linker sequence is a sequence used to link the PCR primer together with the MID sequence. Typically, all MID sequences have the same length l and differ substantially from each other.

In a first step, each sequence in the FASTA file is associated with a (patient) sample using the MID. After a sequence has been associated with an MID, the MID is trimmed off the DNA sequence. To associate and trim, four different algorithms are used. The different algorithms allow perfect MID and linker sequences to be trimmed but also tag sequences with sequencing errors.

The first (and fastest) algorithm takes each sequence from the FASTA file and checks whether one of the MID sequences shows a perfect match at the 5'-end of the sequence. If no match is found, a second (slower) algorithm takes the first four bases of the sequence and looks for a perfect match with the first four bases of one of the MID sequences. If still no match is found, a third algorithm takes four bases from inside the sequence at position $[l\text{-}3, l]$ (where the end of MID is supposed to be) and looks for a match in one of the MID sequences. If all these algorithms fail, a slow fourth algorithm takes the first $l+1$ bases of the sequence and looks for a match with one of the MIDs allowing 2 mismatches. The same methodology is used to trim off linker sequences.

In a second step, the trailing sequence (residual reverse complements of the MID/linker at the end of a sequence) is trimmed off. The trimming algorithm detects complete or incomplete linker or MIDs sequences at the 3'-end and trims them off.

In a third and final step, the sequences of quality scores accompaning the FASTA sequence are trimmed accordingly. Trimming both the FASTA and QUAL sequences results in more efficient mapping in later stages of the pipeline as artificial nucleotide strings are removed from the sequence before they can interfere with the mapping process. The combined length of the MID and linker is often larger than 20

bp and can cause problems in mapping, especially when amplicons are already relatively short.

5.3.2 Generating the reference amplicons

Both a reference sequence and a list with forward and reverse primers (in FASTA format) are fed to the amplicon generation module. The primer sequences are detected in the reference sequence and the reference amplicons are extracted from the reference sequence. The reference amplicons, primer information, and genomic reference information are stored in the references database.

5.3.3 Mapping the sequenced amplicons

Recently, several fast and elegant mapping algorithms have been developed such as Bowtie, BWA or SOAP [190]. Unfortunately, these algorithms were developed for short-read mapping, typically SOLiD or Illumina reads, and are thus not suitable for mapping 454/Roche GS-FLX reads. In the VIP pipeline, sequenced amplicon sequences (reads) stored in the raw sequences database are mapped onto the reference amplicons using BLAT. The output is parsed using Perl, the *Bio::Tools::BPlite* package and some custom-made Perl modules (included in the pipeline package).

In amplicon resequencing experiments spanning large exons, overlapping amplicons are used to cover the exon completely. Using the amplicon as a reference template speeds up the mapping process. However, this causes the sequenced reads to map on multiple, partially overlapping amplicons. This problem is addressed by using the mapping position on both the reference sequence and the read. The read should map completely at the beginning of a reference amplicon to be considered as a correctly mapped read. The algorithm processing the the BLAT output takes into account the mapping position and size of the mapping to identify the correct reference amplicon.

Sequences with a mapped region (including the PCR primer) shorter than a certain threshold value (default value is 40 bp) are considered to be too short to be a real amplicon. The algorithm processing the the BLAT output will actively look for another mapped region (also shorter than the threshold value) in the sequenced read. When two such regions at both the start and end of a read are detected, both are stored in the database flagged as 'short'.

Users who might prefer to map reads genome-wide – to detect PCR artifacts for example – have the possibility to map on the entire genome, rather than on reference amplicons, using the genome-wide reference database (which is included in the package). Further downstream processing is the same for both approaches.

5.3.4 Detecting variation in the mapped data

Once the sequenced reads are mapped onto the reference template and the best mapping position is determined, the possible differences between the sequenced read and the reference template along with their position relative to the reference template are determined.

Variation reported by BLAT is stored per sequence on a nucleotide resolution. For example, a three nucleotide deletion will be stored as three single nucleotide deletions. However, two additional problems need to be addressed. First of all, BLAT is sometimes inconsistent when reporting variation. Imagine a part of a reference sequence AATTTAA and part of sequenced read AATTAA. Using BLAT, different variants might be reported. BLAT might report the first, second or third T to be deleted, depending on the context around that specific sequence. To make BLAT more consistent, variants are 'pushed' to the 5'-end of the sequence, meaning the deleted T in our example will always processed as the first T of the 3Ts being deleted. Now imagine a reference sequence containing a little 2 bp repeat GGATATATCC and a sequenced read containing GGATATCC. BLAT might report the first, second or third AT to be deleted. Again, this deleted AT will be 'pushed' to the 5'-end and then split into two single nucleotide deletions.

The second problem concerns forward and reverse sequencing. All references, variants, and positions are converted to the forward orientation of the reference genome. This allows more consistent processing of variation.

Rather than storing the alignment itself in a database, the reference amplicon name and the variations are stored in the database. A small sequence window around the variation together with the quality score on the position of the variation is stored in the database as this data is used in later stages of the pipeline.

5.3.5 Reporting

Once the processed data is stored in the database, the VIP pipeline is capable of generating eight standard reports that give end-users ample information to interpret the raw data (Table 5.1). Each of these reports can be generated independently from each other, comes straight from the database and is completely independent from the raw sequence files or raw data processing. Hence, any computer that is connected to the database server can generate its own reports in a short period of time without having access to the raw data.

Table 5.1: *Overview of the standard reports that can be generated and the duration*

Report	Duration	Paral.	Threaded duration
Coverage analysis per MID	23 sec.	no	23 sec.
Length distribution	24 sec.	no	24 sec.
Short sequences analysis	56 sec.	no	56 sec.
Coverage analysis per amplicon	94 sec.	no	94 sec.
Quality score analysis	02 min.	no	02 min.
Coverage analysis per MID/amplicon	18 min.	yes	04 min.
Variation analysis	07 h.	yes	02 h.
Coverage analysis per single base	41 min.	no	41 min.
Total time needed for all reports	± 8 h.	yes	± 03h.

Duration: Time needed to generate the reports of a single GS-FLX run (300,000 reads); Paral.: Is the report generation parallelized? Threaded duration: Time needed to generate the report using multiple threads (7 threads).

Reports are mainly generated using SQL-statements and the report generating scripts are partially parallelized where possible using the Perl packages *threads* and *threads::shared*. A more detailed view of each report is given in Appendix File 5.1.

5.3.6 Reporting variation

Variation is reported as an aggregate of the variation of the individual sequences. For each of MID/amplicon combinations, all the single nucleotide variation per sequence are queried from the database. Using the aggregation module, several parameters such as coverage, absolute frequency, and relative frequency are calculated for each aggregate variant. An overview of all the calculated parameters is given in AppendixFile 5.2.

By pushing the variation to the 5'-end (as explained earlier) and looking at variation on a single nucleotide level, the VIP pipeline is capable of detecting indels on an aggregated level although deletions and insertions might not be exactly the same in each sequence of a sample. An example demonstrating this is shown in Appendix File 5.3.

5.3.7 Filtering variation

Detecting variants followed by interpretation is the most important goal of amplicon resequencing projects. The mapping step detects variation by using BLAT and stores the raw variants in the database. These include real variants but also sequencing errors. When reporting to the end-user, the VIP pipeline is able to discriminate between sequencing errors and real variants and is optimized to reduce false negatives, which is of crucial importance in a diagnostic setting.

In theory, genomic variation is homozygous or heterozygous, thus the frequency of a certain observed variant in an amplicon has to be 50% or 100% of the total reads. In practice, this variant frequency can vary due to sampling variation, sequencing errors, and biological heterogeneity. At low coverage, variants can deviate from the 50% or 100% frequency and make discrimination between homo- and heterozygous variants difficult.

It is known that much of the sequencing errors occur when the basecaller software has to determine the exact length of a stretch of the same nucleotides (homopolymer sequences) [226]. Homopolymer error rates were calculated by looking for homopolymer stretches and counting homopolymer associated deletions and insertions. Although the overall GS-FLX error rate is reported to be relative low (≈0.035%; data not shown), the homopolymer related error rate is at a considerable level (≈10% for homopolymers with length 6bp and up to 50% for homopolymers with length 8bp; data not shown) and causing difficulties when calling certain variations as a true variant or a sequencing error [57]. The art of variation analysis is to pick the true variants from the total pool of observed variants.

Taking all this into account, a simple filter was designed that automatically separates sequencing errors from real variants. Keeping in mind that most of the observed variants are indeed sequencing error, variants with absolute frequency = 1 and relative frequency < 10% are never reported and thus filtered out. Further filtering needs to be done by the end-user using filtering values for average quality

scores, coverage, and homopolymer length. Using a dataset with known variants, recommended threshold were determined. These recommended values are: variation frequency >33% and <67% or >0.95% coverage >20, hompolymer length < 6bp, and quality score >30. Using these recommended values automatically filters out unwanted variants.

However, the coverage filter value and frequency filter vlaue are dynamic and can be changed to one's needs. A more exact filter value in function of the desired detection power can be calculated using the tool developed by De Leeneer et al. [58].

5.3.8 Alignment visualizer

Sometimes, the best way to determine which variants are real and which are not is by looking directly to how the reads are aligned to the reference sequence. The VIP pipeline contains an alignment visualizer that gives a global overview of how several reads were aligned to a reference sequence. The visualizer outputs results per MID/amplicon combination which users can specify. An example output of the visualizer is given as Appendix File 5.4 .

5.3.9 The VIP validator

The VIP validator verifies whether a certain introduced variant can be detected by the pipeline in a background of existing variants and sequencing errors. Variation is introduced in sequenced reads rather than in the reference template to simulate analysis of real data.

The validator combines the raw sequences database and the mapped sequences database to retrieve a set of raw sequences that map on a certain amplicon which eventually leads to a set of raw sequences associated with the specific amplicon. These reads already contain sequencing errors and real variants. The validator introduces a new variation (SNV, deletion, insertion or combination) at a certain position in a certain amount of sequences in this set. This variation can be something random (i.e. a random variant at a random position) or a predefined variant.

These variations are introduced with a certain frequency (from 0% to 100%) and with a certain coverage by introducing it in a random subset of sequences that are present in the total set of sequences. This way a dataset is generated as if it has been sequenced. This newly generated dataset is then fed to the mapping and vari-

ation detection module of the normal analysis pipeline and the results are further processed in a validation reporting step (Figure 5.1).

5.4 Results

The modular approach makes efficient planning possible. Time-consuming reporting steps can for example be postponed to a time when there is sufficient server power available. This is especially interesting when multiple runs need to be analyzed in a short period of time and efficient server usage is required.

The database approach makes it possible to easily generate reports and draw conclusions from a subset of sequences of a sequencing run or multiple runs together. It also prevents the user from losing data as all the data is centralized in a single location.

5.4.1 Amplicon pools

The performance of the pipeline was initially assessed by analyzing two *BRCA1/2* resequencing experiments [227]. These two runs contained samples that had been analyzed before using classic HRM (high resolution melting) and consequently the variants in the samples were known. Both runs had a similar experimental design.

A total of 111 amplicons (44 *BRCA1*, 67 *BRCA2*) were equimolarly pooled together per sample (patient). PCR products of 11 patients (tagged with MIDs) were then equimolarly pooled together. Amplicon sizes ranged from 136 bp to 435 bp (mean 244 bp). The first run generated 542,532 reads (mean length 244 bp); the second run generated 261,646 reads (mean length 247 bp).

5.4.2 Processing the raw data

During the trimming step (processing the raw data), 97.37% of the raw sequences were split into MID sequence, linker sequence, amplicon sequence and trailing sequence. A small fraction of sequences had a start with too many sequencing errors to determine the correct MID, and these sequences were consequently not used in further analyses (but are nevertheless stored in the raw sequences database). At the moment there is no straightforward method to split sequences into MID, linker and amplicon sequence using the AVA software (AVA only allows MIDs to be split off),

which makes it difficult to compare the splitting algorithms, but with a 97.37% yield one can assume that few improvements can be made.

5.4.3 Mapping the reads

BLAT mapped 85.24% of the sequenced reads generated in the first run and 93.57% of the reads generated in the second run. Some of the mapped sequences were filtered out because they map to a reference amplicon in two pieces with big gap in between and appear to be clear primer-dimers (the so-called 'short' sequences). 76% of all the sequences reads mapped correctly and passed the default filters in the first run compared to 92% in the second run.

The residual portion of the sequenced reads that did not map were further investigated and appeared to be PCR artifacts such as complex primer dimers. Only 4,653 reads (completely or partially) mapped outside the target regions when mapped genome-wide and only 45 of these mapped completely (i.e. from start to stop) somewhere in the genome, from which we concluded that the PCR reactions were specific.

Coverage per amplicon and per MID was heterogeneous and between 0 and 5,134 (mean 310) for run 1, and between 0 and 3,019 (mean 175) for run 2. This heterogeneity is mainly caused by differences in PCR efficiency and suboptimal pooling of samples. This heterogeneity (or 'spread factor') can be lowered by carrying out improved pooling strategies and incorporating normalization steps (Hellemans et al., in preparation). Errors in the labwork and/or PCR reactions not working as expected caused some amplicons to be not covered (2.98% in the first run, 1.51% in the second run).

Mapping on reference amplicons rather than on a reference genome certainly improves speed, but is somewhat discussable because one might miss paralogous amplification products. However, in a diagnostic resequencing setup one is interested in variants in a certain set of genes or even a certain exon using thoroughly validated PCR reactions. Paralogous amplification products might become apparent when mapping genome-wide (and be absent when mapping only to reference amplicons) but they would lack diagnostic significance as they give no or incorrect information on the region that was intended to be screened. Users should strive to the use of validated PCR reactions and omitting genome-wide mapping, rather than

mapping genome-wide. Nevertheless, the genome-wide approach is included in the package for users desiring to screen for aspecific PCR products.

5.4.4 Meta-analyses

Meta-analyses can be carried out by modifying the existing reporting modules slightly. The SQL queries can be adapted to narrow down the analysis range or to expand the analysis range. Several useful meta-analysis scripts are available upon request. Figure 5.2 shows an example where average quality scores (Q scores) are calculated using data from three different runs. The resulting peak coverage was more than 1,000,000x.

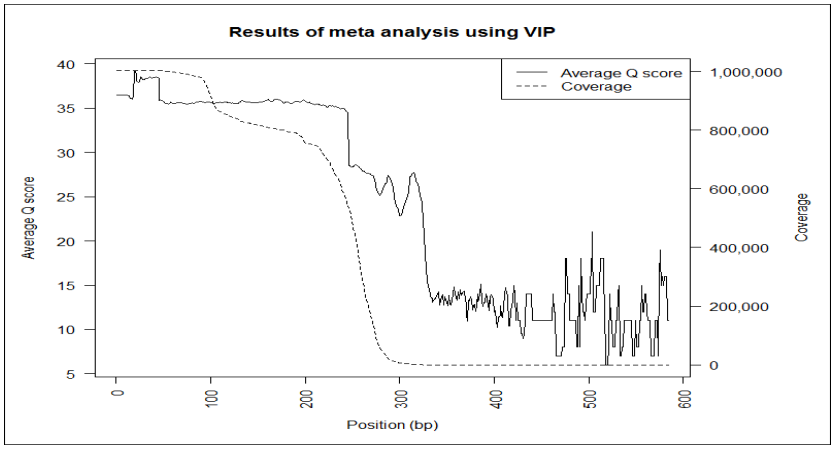


Figure 5.2: *The averages Q score is shown using the full line (left y-axis); the coverage is shown using the dashed line (right y-axis). The x-axis shows the position on a single read. Data displayed is coming from more than 1,000,000 reads that originate from 3 different runs.*

5.4.5 Calling true variants

The two *BRCA1/2* runs contained samples with known variants (132 distinct sequence variants of which 90 were deletion/insertion mutations). By setting the filters (after using the 'generate reports' script) at the recommended values one can discriminate between true variants and sequencing errors. It is difficult to reliably design a filter to only filter out faulty homopolymer variants. Discarding every variation preceding or following a homopolymeric region is not an option as real variants can also flank such regions. We have seen that for homopolymers < 6 bp there is no

problem discriminating between correctly sequenced homopolymeric stretches and homopolymer related insertions or deletions (sequencing errors) by using the quality score (Q). When using the recommended filter value Q > 30 only a minor fraction of the homopolymer related sequencing errors pass the filter (homopolymers < 6 bp). When the stretches are 6 bp long or longer the distributions of the normal and the mismatch homopolymers start to overlap and even a correctly sequenced base has a low Q score (Figure 5.3). We recommend to set the homopolymer filter at 6 bp and keep in mind that no real variants preceding or following a 6 bp homopolymeric stretch can reliably be detected using the default filter values.

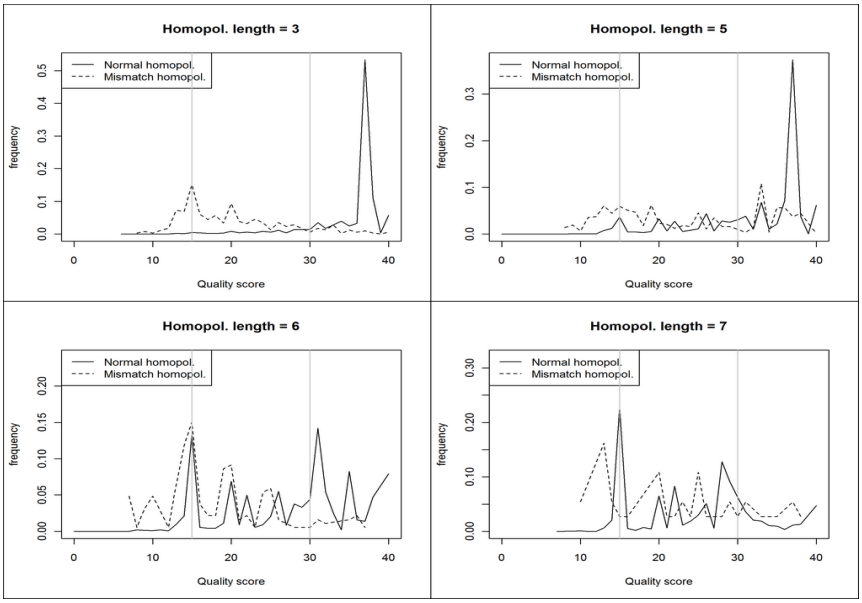


Figure 5.3: *Distribution of homopolymer related quality scores (Q score). The normal homopolymer Q score distribution is determined by making a distribution of the Q score of the homopolymer base; the mismatch homopolymer Q score distribution is determined by making a distribution of the Q score of the base preceding a homopolymer related deletion or the Q score of a homopolymer related inserted base. Distributions are shown for homopolymers with length 3, 5, 6 and 7 bp. The grey vertical lines are drawn at a Q score of 15 and 30. Distributions are based on data from the two BRCA runs.*

In total, 97% of all known variants (homozygous and heterozygous) could be detected (sensitivity). Specificity (the portion of called variants that actually are real variants) was 98.5% which means that the false positive rate is only 1.5%. All non-

detected variants were insertions/deletions in or near homopolymeric regions. We are aware of the fact that the given numbers may be overfitted to a *BRCA1/2* screening, but very similar results were obtained in other experiments (20 genes, 4 runs) as well (data not shown).

A comparison of the data pre- and post-filtering (using the recommended settings) is given in Figure 5.4. The figure clearly shows that the pre-filter data contains a lot of 'noise'. Post-filter data is concentrated around the 50% and 100% level as expected and allows easy discrimination between heterozygous and homozygous variants. At higher coverage, the data is concentrated in a band which is narrower than the specified filter settings. This indicates that at higher coverage, mainly low Q score variants and homopolymer-related variants are filtered out.

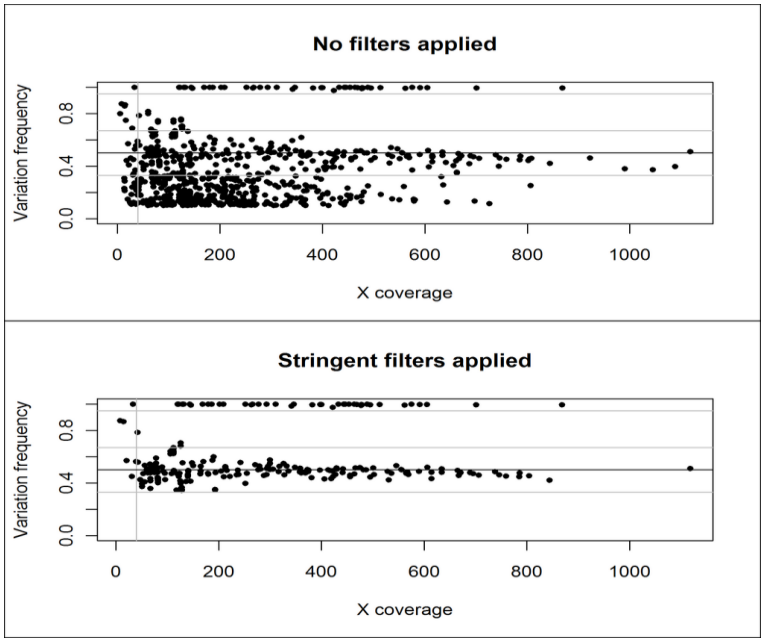


Figure 5.4: *Plot of the coverage (times a single sequence is read by the sequencing equipment) and the frequency of an observed variation. In reality, genomic variation occurs at a frequency of either 50% or 100% of total reads. The top figure gives the distribution when no filters are applied to discriminate between sequencing errors and real variants; the bottom figure gives the distribution where frequency filter, Q score filter, coverage filter and homopolymer filter are applied to discriminate the real variations from the sequencing errors. Vertical grey line indicates 40x coverage; horizontal grey lines indicate respectively 33%, 67% and 95% variation frequency.*

5.4.6 VIP compared with AVA

The performance of the VIP pipeline was compared with the performance of the AVA software (2.0.0.0). Reads from 1 patient (MID1) containing 67 different amplicons were analyzed in detail. The sample was known to contain 12 variants. A detailed overview of the variants and the comparison is given in Appendix File 5.5.

The AVA software does not handle the linker sequence between the MID adaptor and the PCR primer very well. The sequence of the MID/Linker combinations (MID1' = MID1/linker, MID2' = MID2/linker etc.) were used as new MIDs in AVA to circumvent the inability of AVA to split off the linker sequences. The VIP pipeline returned 50 variants compared to the 235 returned by AVA. This is explained by the fact that the VIP pipeline has an internal hardcoded filter which filters out any variant with a frequency lower than 10% and with a coverage of only 1. These variants are considered random sequencing errors. The fact that we still have 0 false negatives indicates this filter is not too stringent.

Setting a minimum frequency filter in AVA to 20%, 33 variants were identified, whereas 14 variants passed the different filters in the VIP pipeline. The VIP pipeline picked up all 12 known variants (including 6 difficult homopolymer related variants) and called only two false positives, both homopolymer related. The AVA software on the other hand missed all 6 homopolymer related variants and called 27 false positives (Table 5.2).

Table 5.2: *Comparison of AVA and VIP performance (1 sample, 67 amplicons, and 12 known variants)*

Parameter	AVA software	VIP pipeline
Total variants (unfiltered)	235	50
'Pass filter' variants	33	14
True variants called	6/12 (50%)	12/12 (100%)
False positives	27/33 (81.2%)	2/14 (14.2%)
False negatives	6/12 (50%)	0/12 (0%)

5.4.7 Optimizing pre-sequencing labwork

Besides generating an overview of the variants, the pipeline can also generate additional reports which make the pipeline very useful in a diagnostic setting. All the data is intelligently stored in a relational database and therefore custom analyses can be carried out using SQL statements in either an SQL browser such as HeidiSQL

[228] or by writing custom scripts in any scripting or programming language that has the ability to communicate with a MySQL database.

This database approach allowed optimization of the laboratory work of the *BRCA1/2* sequencing experiments, especially the pre-sequencing PCR reactions. It was clear that there were a lot of short sequences in the first run. Run 1 had 8.73% of total reads flagged as 'short sequence', run 2 had 1.87% of total reads flagged as such. Using this information, it was possible to pinpoint exactly which the PCR reactions were failing. Using this information obtained *in silico*, the multiplex reactions were optimized in the lab and an additional length separation was carried out. These actions improved the efficiency by reducing undesired by-products and/or primer dimers from 24% to 8% of the total sequences (Figure 5.5).

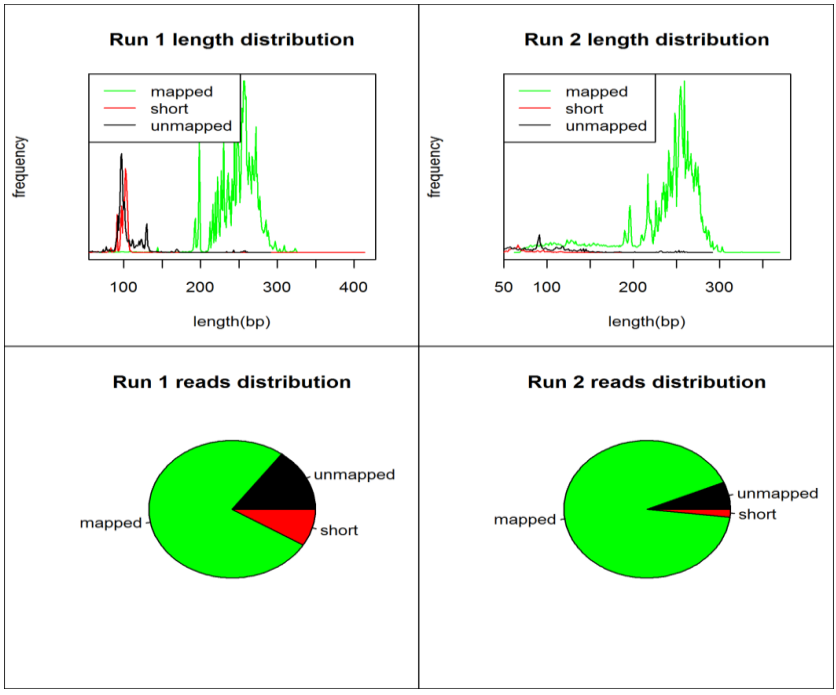


Figure 5.5: *Example of the reporting possibilities. Run 1 had many unmappable and short, mapped sequences. Length distribution showed these were mainly 60-120bp sequences. In Run 2 optimized PCR reactions and an additional length separation were carried out prior to the sequencing with a huge reduction (8% vs. 24%) of unmapped and short sequences and thus improving the cost-effectiveness.*

5.4.8 Parallelization

The report generating scripts are parallelized where possible, but one should bear in mind that hard disk accessibility and memory usage are a bigger issue than processing power when using SQL-statements.

Processing time can be reduced approximately 75% when increasing the number of parallel processes from 1 to 7. When more than seven processes are run in parallel, there is no beneficial effect on the processing time. When running between 10 and 15 parallel processes there are even possible adverse effects of increasing the number of parallel processes (Figure 5.6).

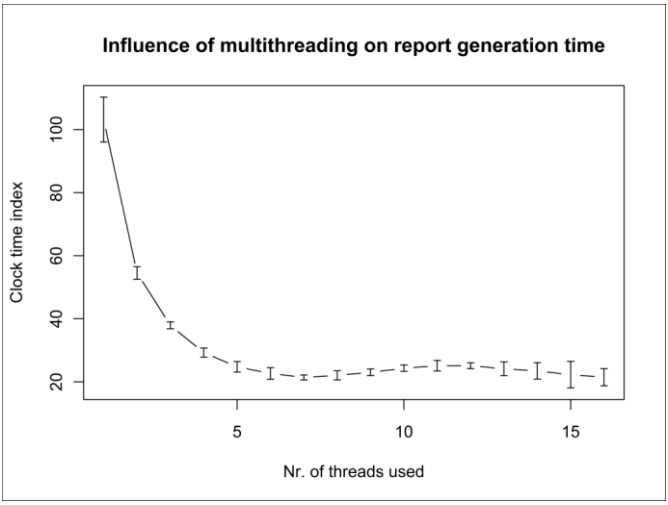


Figure 5.6: *Clock time needed to perform a reporting action using SQL statements and the effect of using multiple threads. Increasing threads reduces the clock time significantly in the beginning, but using more than seven threads has no beneficial effect. Averages and standard deviation shown for 6 repetitions.*

5.4.9 Validating variation using the VIP Validator

Random variation in the VIP Validator is defined as a random error that is introduced with a certain frequency at a certain position in a set of sequences that map on a random reference amplicon. The Validator introduces the variant into a number of sequences and then verifies in how many of these sequences this specific variant was detected at the exact location where it was introduced.

Random variation validation was initially used to optimize and validate the VIP analysis pipeline but has proven to be an effective instrument to detect problematic amplicons or problematic regions in certain amplicons. There are amplicons for which few variants have been reported yet, meaning that amplicon resequencing experiments can unravel new variants. A simulation with random variations can give a clue about regions where accurate variation detection is difficult; for example multiple variations close to each other or variations in repetitive regions can be problematic for the mapping software. The Validator does not explain why a certain region becomes a difficult region to detect variants but informs about variants that cannot be detected using the pipeline.

Random variation was introduced by choosing a random amplicon, a random position, a random variant, a random frequency, and a random coverage and then introducing the variant accordingly. This process was repeated 1000 times.

The ratio between the number of sequences in which a variation was introduced (the coverage so to speak) and the number of sequences wherein the variation was detected is the detection ratio. This ratio is independent from the frequency by which a variant was introduced.

It is clear that SNV detection with the VIP is no problem as 100% of the introduced variants can be detected by the VIP Validator with a sufficiently high detection ratio. The 67% threshold is considered as sufficiently high because the detection ratio for a heterozygous variant would be at least 33.5% and still pass the 33% filter.

Deletions (and insertions) are often not detected using the newest NGS mapping and variation detection packages. The VIP pipeline can detect >99% of random 3bp deletions with a detection ratio ≥ 67%. Determining the location of a gap appears to be relatively easy. The difficult part of gap detection is determining the exact length. Results are similar for longer (10bp) deletions and insertions (both 3bp and 10bp). An overview of the detection ratios is given in Figure 5.7.

These observations make it clear that deletions and insertions can be detected but one should keep in mind that the exact length of the insertion/deletion that passes through the filters is not necessarily the real length. This is one of the drawbacks of needing multiple reads to have a reliable call for a single nucleotide position. Nevertheless, it is detected that something is wrong, which is essential in a diagnostic

setting. The alignment visualizer can be a useful resource to manually assess the exact deletion/insertion size.

Rather than using random variation, the VIP Validator can also use a list of known variations as input. Validating known variation is the most useful application of the VIP Validator in a diagnostic setting. It gives an answer to the question "If variation x is present in an amplicon, would it then be possible to detect it using the VIP pipeline?". The answer to this question is objectively addressed using the detection ratio parameter.

Around 95% of the known *BRCA1/2* variants were detected by the Validator with a detection ratio of 100% meaning that in every single read (wherein the variant was introduced) the variant was detected at the correct position. The other 5% of the variants were detected with a detection frequency that is lower than 100% but still larger than 67%.

The VIP Validator can also be used to determine minimal needed coverage *in silico*. The number of sequences wherein variation is introduced can be altered and one can find a minimal coverage that is needed to have a sufficiently high and reproducible detection frequency.

This Validator is useful for pipeline optimization and determination of the ideal cut-off values. It allows the end-user to fine-tune the pipeline to its own needs and to objectively validate the results of a given pipeline modification. Moreover, it allows validation of the analysis software with respect to the detection of certain variation screening, which is very important in diagnostics, and it determines the detection limits of the pipeline, prior to starting a diagnostic screening.

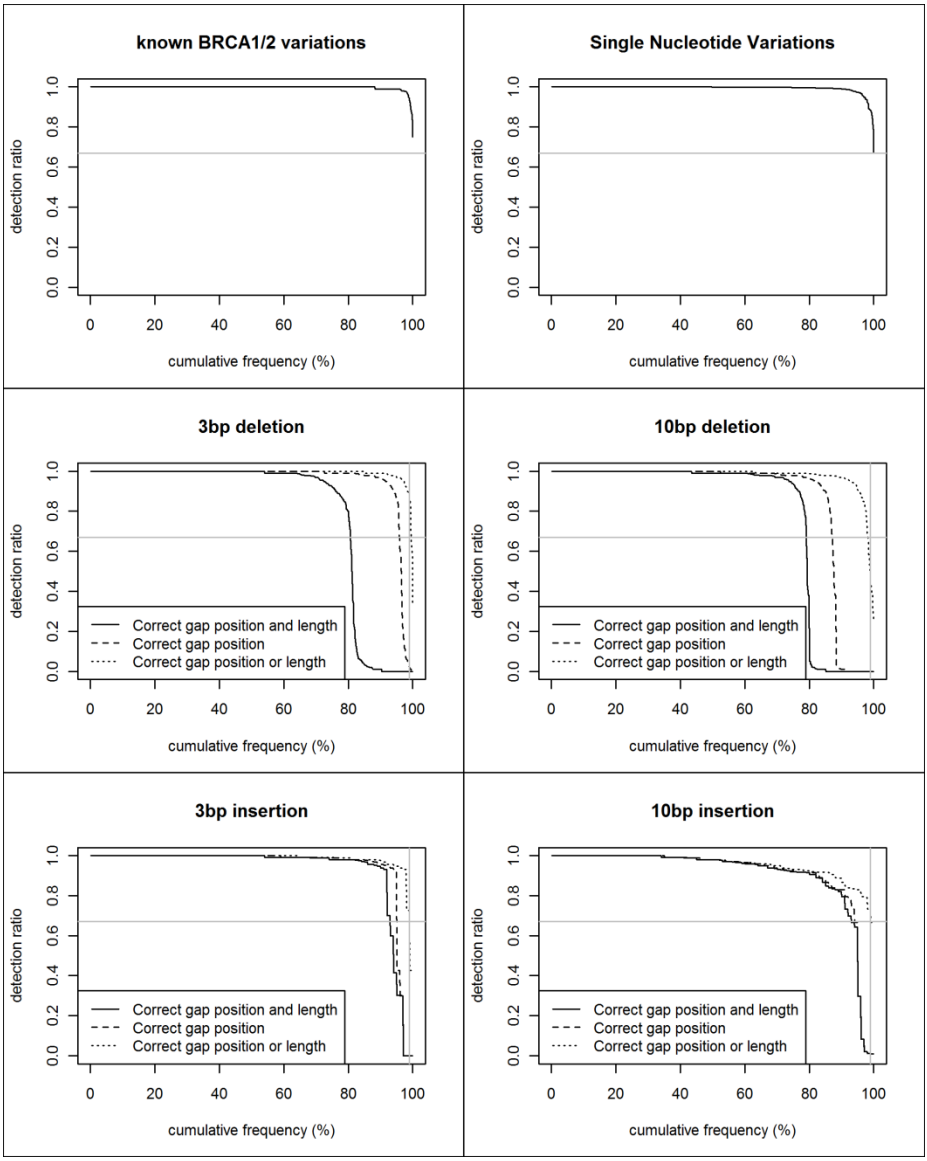


Figure 5.7: *VIP Validator detection ratios of 1000 known BRCA1/2 variants, random SNVs, random 3bp and 10bp deletions, and random 3bp and 10bp insertions. The grey horizontal lines indicate 67% detection ratio. The grey vertical lines indicate a 99% cumulative frequency.*

5.5 Current state of the Variant Identification Pipeline

The initial public version of VIP (version 1.3) has been further developed into VIP version 1.4, which has a set of improved functionalities. Compared to VIP 1.3, the current VIP 1.4 pipeline is completely built around the database, meaning that final reports are not written to text files anymore. Raw 454 sequencing data, processed intermediary data, and identified variants (together with other reports) are stored in the database. This approach allows a more robust data access, data storage, and data manipulation. However, reports can still be exported to flatfiles from the database if the end-user wishes so.

Improved analysis of genome-wide sequencing experiments is now implemented in VIP 1.4. During the reference amplicon generation step, a user can specify a region (up to a chromosome) to act as a reference. Artificial 10 kb amplicons are generated by the pipeline and stored in the database. Furhter analyses of such experiments is completely similar to analyzing regular amplicon resequencing experiments.

Furthermore, the VIP pipeline now uses a logging system, which tracks the executed commandos, the user which initiated the pipeline, timestamps etc. This logging system allows different users to analyze different datasets on a single system and guarantees data integrity and prevents data mix-ups.

Other small changes such as speed improvements, addition of new parameters, improvement of the command-line interaction with the end-user and minor bug fixes have been carried out. All the changes are described in detail in the VIP manual and change-log. VIP 1.4 is available from http://athos.ugent.be/VIP_pipeline.

5.6 Discussion

MIDs and linker sequences are trimmed efficiently, there are some drawbacks, however. The algorithm can only trim sequences when the MID used on the forward and reverse primer are the same one. This because the trimming algorithm uses the MID/linker information obtained at the 5'-end to trim off similar sequences at the 3'-end. Some users might prefer to use different MID at the forward and reverse primer to increase the number of samples that can be pooled. For example Sample1: MID1+MID2, Sample2: MID1+MID3 etc. In that case, VIP will not efficiently

trim. However, currently there are ~100 different MIDs available, decreasing the need to mix MIDs together.

Reporting variation on single nucleotide resolution allows complex indels to be reported. However, it becomes difficult to determine the exact frequency of the indel as it is not analysed in its totality (see fo example Appendix File 5.3). The frequency of each nucleotide of the indel might differ, making sometimes even impossible. An elegant solution is to use for example the lowest frequency of the entire indel. The visualizer offers the possibility to assess the complex indel manually in detail and often it becomes clear which indel really occurred. Despite its shortcoming, it still is a valuable asset of the VIP as there are many packages or pipelines available reporting no deletions or insertions alltogether.

We recommend to set the homopolymer filter to 6. In most of the cases, this will not pose any problems as there are not many homopolymer stretches in coding regions. However, one should bear in mind that there are known homopolymer related variations (for example in the *CFTR* gene) and that VIP will not perform optimal when trying to detect those. But homopolymer related sequencing errors are a known pyrosequencing issue, so trying to detect such variations using pyrosequencing technology is not a good starting point alltogether.

Most of the sequencing errors are removed from the data by the hardcoded filter implemented in the pipeline (see for example the 235 variants reported by AVA and the 50 unfiltered variants reported by the VIP in the BRCA example). The recommended filter settings will remove most of the noise data, but this approach is pragmatic, rather than perfect. Other approaches include sequencing a dataset with known variants and determining filter values using these known variants or calculating optimal filter values as mentioned earlier [58].

It is important to note that the pipeline itself only has limited hardcoded filtering rules. Further filtering needs to be done by the end-user using the extensive list of parameters which are reported back to the user. This can be done directly on the variants database using SQL-queries or once the variants are exported to an Excel file using the filter functionalities in Excel. Two important notes should be added. First of all, the rough hardcoded filtering in the VIP removes most of the false posities (as shown in table 5.2). Furthermore, commercial software such as AVA typically reports a list of variants and their relative frequency and does not include

the absolute frequencies on the forward strand and reverse strand, quality scores, homopolymer length ... making adequate filtering impossible alltogether.

The modular approach offers flexibility to the end-users. For example, users wanting to only trim off MID sequences can use only the first module. Perhaps other users want to write their own custom reporting modules or aggregate the variation data themselves. It is perfectly possible to use parts of the pipeline and integrate them into other pipelines.

We developed and tested the pipeline using human samples and human reference sequences. However, the pipeline will work with any type of sequencing data, provided there is a reference sequence available and the sequencing data is in the FASTA/QUAL format. This means viral, bacterial, and plant samples sequenced on the 454 platform can be analyzed using the VIP. Upcoming pyrosequencing technologies such as Ion Torrent sequencing produce the exact same output format and can benefit from these methodologies as well. However, the basecalling software of Ion Torrent is currently still being actively developed and it is unsure whether the current filtering recommendations can simply be used to accurately filter variants obtained using Ion Torrent data.

The read length of Illumina sequencing is steadily increasing. When the Illumina read length reaches the length typically observed in 454 sequencing, it is doubtful whether the VIP can be used to analyze Illumina experiments. Although the databases are well designed and are queried in an intelligent fashion, input/output operations will become a bottleneck as the number of sequences obtained in Illumina sequencing are typically in the ranges of billions whereas 454 sequencing yields some million sequences.

5.7 Conclusion

We have developed an open-source pipeline that is of great interest for diagnostic resequencing projects. The AVA software (the standard GS-FLX software package) is user-friendly and useful in a research setting but not exactly what is needed in a diagnostic setting. The VIP pipeline is a fully automated pipeline that enables accurate variant detection, even insertions and deletions, and can objectively quantify the identification power by using the Validator. Although the pipeline performs better than AVA when detecting homopolymer related variants, one should bear in mind that reliably detecting variants next to homopolymers will always be an issue,

even with high coverage and 'intelligent' software packages. The VIP pipeline also allows a degree of flexibility when one wants to fine-tune its performance, a feature which is lacking in AVA.

The pipeline also allows optimization of the sample preparation procedures and amplicon generating PCR reactions and offers a complete suite of reports that allows researchers in diagnostic labs to assess the reliability of a sequencing run and the detected variants.

The VIP pipeline is under continuous development and improvement as more and more sequencing data becomes available to validate and improve the pipeline. Furthermore, an additional module is in development to do an automated and integrated interpretation of the variants that are discovered using both the amplicon and the genome-wide pipeline.

5.8 Availability and requirements

- **Project name:** Variant Identification Pipeline
- **Project home page:** http://athos.ugent.be/VIP_pipeline
- **Operating system(s):** Platform independent, Unix/Linux preferred
- **Programming language:** Perl
- **Other requirements:** BLAT, MySQL
- **License:** GNU LGPL

5.9 Authors' contributions

JDS wrote the VIP scripts and packages. JDS wrote the manuscript. JDS carried out custom analyses to optimize the pre-sequencing steps. KDL carried out the sequencing runs and helped estimating the error profile. SL helped estimating the error profile. NS designed a mathematical approach to determine optimal filter values. FP participated in the design of the pipeline. FVN participated in the design of the pipeline. PC, DD, JV, SB, JH, WVC provided access to the sequencing facility and were involved in the design of the pipeline from the very beginning. All authors read and approved the final manuscript.

Adapted from:

Joachim De Schrijver et al. (2012), **Advancing the Variant Identification Pipeline**, in preparation

Joachim De Schrijver & Geert Trooskens (2010), **Next-generation high-throughput interpretation and visualization of genetic variants**, *Advances in Genomics Symposium 2010: Next Generation Sequencing* poster presentation

Sarah De Keulenaer, Jan Hellemans, Steve Lefever, Jean-Pierre Renard, Joachim De Schrijver, Hendrik Van de Voorde, Mohammad Amin Tabatabaiefar, Filip Van Nieuwerburgh, Daisy Flamez, Filip Pattyn, Bieke Scharlaken, Dieter Deforce, Sofie Bekaert, Wim Van Criekinge, Jo Vandesompele, Guy Van Camp, Paul Coucke (2012), **Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform**, *BMC Medical Genomics* 5:17

Machteld Baetens, Lut Van Laer, Kim De Leeneer, Jan Hellemans, Joachim De Schrijver, Hendrik Van De Voorde, Marjolijn Renard, Bjorn Menten, Wim Van Criekinge, Julie De Backer, Anne De Paepe, Bart Loeys, Paul Coucke (2011), **Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz Syndrome**, *Human Mutation* 12 (9): 1053 - 1062

Kim De Leeneer, Jan Hellemans, Joachim De Schrijver, Machteld Baetens, Bruce Poppe, Wim Van Criekinge, Anne De Paepe, Paul Coucke, Kathleen Claes (2011), **Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges and limitations**, *Human Mutation* 32 (3): 335 - 344

6 The Variant Interpretation Pipeline

6.1 Abstract

Due to recent developments of the next-generation sequencing (NGS) technology, novel potentially disease-causing genetic variants can rapidly be identified. As the amount of sequencing data becoming available is growing and is expected to keep growing in the future, straightforward and automated interpretation of such identified variants is needed.

We designed an open-source interpretation module, the Variant Interpretation Pipeline) for the Variant Identification Pipeline– combined the Variant Identification and Interpretation Pipeline (VIIP) – that annotates genetic variants. The combined pipeline is designed to process Roche/454-resequencing data from raw data to variant interpretation. Furthermore, VIIP can annotate external lists of variants.

VIIP is implemented in Perl, uses a MySQL database, and runs on all Unix-compatible platforms. The pipeline, a manual, and test dataset are available under the GNU LGPL license from http://athos.ugent.be/VIIP.

6.2 Background

Current next-generation sequencing (NGS) technologies are able to sequence hundreds of millions high quality bases in a single run [55, 61, 229]. Sequencing technology is expected to keep advancing and the cost of genotyping to decrease significantly in the near future. It is expected that next-generation sequencing will replace Sanger sequencing as the golden standard in the diagnostic field [223].

Currently, diagnosticians rely on a plentitude of tools to annotate genetic variation. Effects of structural alterations of proteins [230, 231] and splicing effects [232] due to genetic mutations can be assessed easily. Some effort has been put into place to combine several tools and automate batch processing. Alamut (proprietary), FANS [233], and MutationTaster [234] are examples of programs that combine different tools. However, Alamut still requires manual steps, FANS can only be run from the developer's website and focuses mainly on enhancer/silencer regions, and MutationTaster does not analyze structural alterations of proteins.

We developed the combined Variant Identification and Interpretation Pipeline (VIIP) which is able to process 454/Roche GS-FLX re-sequencing data, identify variants, and annotate variants. VIIP determines whether a variant is new using dbSNP, determines the impact of the variant on a protein level and the influence on splicing. Furthermore, the pipeline is designed to process in a batch mode in a user-friendly way.

6.3 Implementation

The Variant Interpretation Pipeline is written in Perl and uses a relational database (MySQL) as back-end. The pipeline uses the *DBI* and *DBD::mysql* packages for database interactions, Ensembl API modules to interact with Ensembl, bioperl packages, *LWP::UserAgent*, and 2 custom Perl modules (*gs_flx_interp_functions*, *gs_flx_vip_interpretation*). The pipeline is designed to run on any Linux distribution and is expected to run smoothly on standard hardware.

6.4 Methods

6.4.1 Integration Variant Identification and Variant Interpretation Pipeline

VIIP consist of two distinct modules, the Variant Interpretation Pipeline (VIP) [235] and the Variant Interpretation Pipeline. VIP identifies variants in the sequencing data, whereas the Variant Interpretation Pipeline annotates the identified variants. A schematic overview of the integration of both database structures is given in Appendix File 6.1.

6.4.2 The Variant Interpretation Pipeline

Variants identified using the VIP are automatically stored in a separate variation database. Reference species (e.g. *Homo sapiens*), reference genome (e.g. NCBI37), chromosome, chromosomal position, wild type allele, and variant allele are stored in the database, together with a unique variant identifier. Variants already identified in earlier experiments are only stored once and later referred to by the variant identifier; effectively decoupling variant identification and interpretation and avoiding redundant analyses. Variant interpretation is only carried out on yet unannotated variants present in the variation database.

All the interpretation results are stored in a MySQL database which can easily be queried to have the data visualized in for example a genome browser. In total 22 parameters are analysed and stored in the database. An overview of all the parameters is given in Appendix File 6.2.

Furthermore, the Variant Interpretation Pipeline does not only allow variants identified with the Variant Identification Pipeline to be annotated. It allows users to submit a list of variants to the pipeline, provided the format is correct. Results are also returned in flatfile format.

6.4.3 Determining protein effect

The Variant Interpretation Pipeline retrieves genomic features from the Ensembl database using the Ensembl APIs [236]. Exon, intron, and untranslated region (UTR) locations of each transcript of a gene that is impacted by the variant to be annotated are retrieved. When the variant is not located inside a gene, the variant is categorized as 'intergenic' and subsequently not further processed. When a variant is located inside a gene, it is categorized as 'genic' and further processed.

Using a reference genome stored locally, and the retrieved exon, intron, and UTR locations each reference transcript sequence is rebuilt. Using the gene structure of the rebuilt transcript, the 'genic' variants are then categorized as '5'-UTR', 3'-UTR', 'exonic' or 'intronic'. This status of the variant is given for each transcript. For example, a variant located inside a gene having 20 different transcripts will be categorized 20 times.

The 'exonic' variants are further processed by categorizing the variant as 'synonymous' or 'non-synonymous'. Both the reference transcript and the transcript with the variant inserted are translated *in silico*. When both differ, the variant is categorized as 'non-synonymous', when both are the same, the variant is categorized as 'synonymous'.

6.4.4 Determining splice site impact

The effect on splicing is calculated using the algorithm available on the fruitlfy.org website [232]. A 20 nucleotides upstream and 20 nucleotides downstream window around the variant is submitted to the algorithm using the webinterface. Forms are automatically submitted and results processed using the *LWP::UserAgent* module.

Using the obtained results, variants are analyzed for changed splice site affinity (both loss and gain of splice site affinity). The end-user has the option to analyze 'genic' variants, 'intergenic' variants or both.

6.4.5 Determining dbSNP status

Variants are checked against dbSNP [237] to categorize a certain variant as a new potential disease-causing variant or a known polymorphism. The Ensembl APIs are used to access the Ensembl Variation database and retrieve RefSNP(rs) numbers and the frequency of the RefSNP for different ethnicities.

The idea behind this approach is that disease causing variants or relatively rare compared to regular polymorphisms which are frequently present in a high percentage in certain ethnicities. Sometimes rare SNPs, which might be disease causing, are present in dbSNP which could categorize a detected disease causing variant as a regular SNP.

6.5 Results

6.5.1 Variant interpretation

An unannotated variant is taken from the variation database or the flatfile, processed, and the results are again stored in the variation database or reported in a flatfile. By default, the pipeline runs in batch-processing, which makes processing a large number of variants a very straight-forward routine.

Variant input consists of a chromosome number, chromosomal position, reference allele, and variant allele. Using Ensembl data, variants are categorized as either 5'-UTR, exonic, intronic, 3-'UTR or intergenic. Exonic variants are further categorized into synonymous and non-synonymous variants. Depending on the category wherein a variant is categorized, other parameters are reported such as transcript position, triplet position, and amino-acid position. When multiple transcripts are known for a certain gene, all the transcripts are analyzed independently from each other.

The splice site analysis categorizes a variant as 'changed splice site affinity' or 'no changes in splice site affinity'. Both introduction of new splice sites and variant based splice site skipping are reported, hence both exonic as intronic/intergenic variants are analyzed. The analysis is performed on both the forward and the

reverse strand. When a variant is known in dbSNP, the rs-number and the frequency from the ethnicity showing the lowest frequency is reported. An example output is shown in Appendix File 6.3.

6.5.2 Three-tier architecture

Three-tier architecture is a client–server architecture in which data storage, data processing, and data presentation are developed and maintained as independent modules. These three different tiers can share the same hardware, but are often present on separate hardware platforms [238]. By breaking up the application into tiers, specific tiers can be modified, rather than modifying the entire application. Furthermore, the three-tier architecture allows any of the three tiers to be upgraded or replaced independently from the other tiers. The VIIP pipeline can easily be organized to run as a classic three-tier system (Figure 6.1).

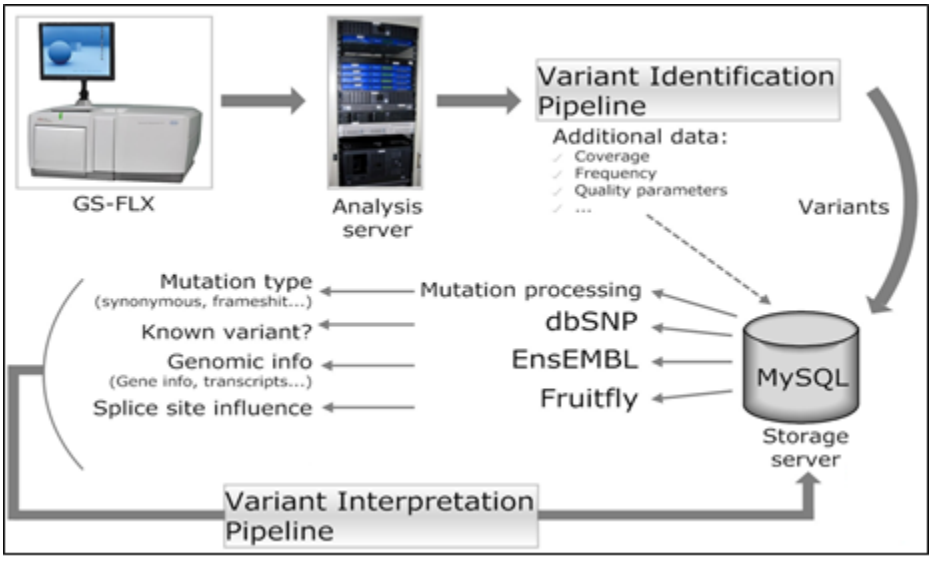


Figure 6.1: *Overview of the dataflow of VIIP. The pipeline can easily be organized as a classic three-tier architecture. Data is stored in a central MySQL database on a storage server (tier 1). Data processing is carried out by the Variant Identification Pipeline and Variant Interpretation Pipeline on the analysis server (tier 2). Data can be visualized by a presentation tier (tier 3 not shown in the image) which fetches the required data directly from the central storage server.*

All the data is stored in a MySQL database system which serves as the first tier (data storage). The stored data is retrieved from the database and processed by both the identification and interpretation modules. These processing modules can be considered as the second tier (data processing). The processed data is ready for immediate visualization using at third tier module (data presentation). An example of data presentation using a modified version of the H2G2 visualization genome browser [239] is shown in Figure 6.2.

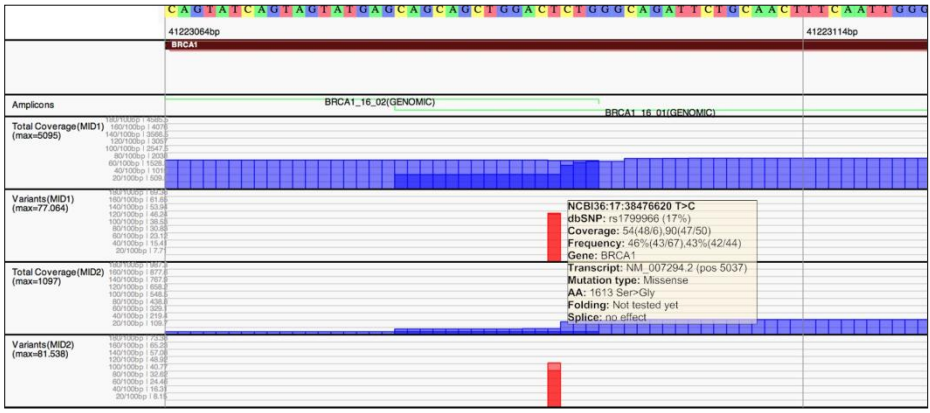


Figure 6.2: *Example of the possibilities of combining the annotated variants stored in the database with the H2G2 genome browser. The top layers show the actual DNA sequence, annotated genes, and designed amplicons. The blue bars indicate the coverage on a base-pair level. The red bars indicate the coverage of the annotated variants. The variant annotation is shown in the yellow box.*

6.6 Discovering variants in medical samples

The Variant Identification Pipeline and Variant Interpretation Pipeline were developed, tested, and optimized using three different datasets [227, 240, 241] which were provided by collaborators. Ultimately, the experiments were optimized and analysed using the VIIP.

6.6.1 Breast cancer

The breast cancer susceptibility genes (*BRCA1* and *BRCA2*) are two important genes frequently analyzed in high-risk female populations. A proof of concept study was carried out to assess the feasibility of using next-generation sequencing as a diagnostic screening tool to identify important mutations in these two genes.

Using 454 sequencing followed by an analysis using the VIIP pipeline, the sensitivity was assessed by analysis of 133 distinct sequence variants. Three (2%) deletions or duplications in homopolymers of greater than or equal to seven nucleotides remained undetected, illustrating a limitation of pyrosequencing. Furthermore, other limitations like nonrandom sequencing errors, pseudogene amplification, and failure to detect multi-exon deletions are thoroughly described.

The used workflow illustrates the potential of massive parallel sequencing of large genes in a diagnostic setting, which is of great importance to meet the increasing expectations of genetic testing. Implementation of this approach will hopefully lead to a strong reduction in turnaround times. As a consequence a wider spectrum of at risk women will be able to benefit from therapeutic interventions and prophylactic interventions.

6.6.2 Marfan and Loeys-Dietz Syndrome

The Marfan (MFS) and Loeys-Dietz (LDS) syndromes are caused by mutations in the fibrillin-1 (*FBN1*) and Transforming Growth Factor Beta Receptor 1 and 2 (*TGFBR1* and *TGFBR2*) genes, respectively. With the current conventional mutation screening technologies, analysis of this set of genes is time consuming and expensive. Using multiplex PCR followed by next-generation sequencing, discovery can be cost-effective.

In a first stage, genomic DNA from five MFS or LDS patient samples with previously identified mutations and/or polymorphisms in *FBN1* and *TGFBR1/2* were analyzed using next-generation sequencing followed by analysis using the VIIP pipeline. This analysis revealed all expected variants. In a second stage, we validated the technique on 87 samples from MFS patients fulfilling the Ghent criteria [242]. This resulted in the identification of 75 *FBN1* mutations, of which 67 were unique. Subsequent multiplex ligation-dependent probe amplification [243] (MLPA) analysis of the remaining negative samples identified four large deletions/insertions. Finally, Sanger sequencing identified a missense mutation in *FBN1* exon 1 that was not included in the NGS workflow.

In total, there was an overall mutation identification rate of 92%, which is in agreement with data published previously. We conclude that multiplex PCR of all coding exons of *FBN1* and *TGFBR1/2* followed by NGS analysis and MLPA is a robust strategy for time- and cost-effective identification of mutations.

6.6.3 Genetic deafness

Hereditary hearing loss can originate from mutations in one of many genes (>60) involved in the complex process of hearing. Identification of the genetic defects in patients is currently labor intensive and expensive. Next generation sequencing technology has the potential to be much more cost efficient. A semi-automated PCR amplification and NGS can offer high sensitivity, speed and cost efficiency.

In a proof of concept study, 15 autosomal recessive deafness genes were screened in 5 patients with congenital genetic deafness. 646 specific primer pairs for all exons and most of the UTR of the 15 selected genes were designed using primerXL. Using patient specific identifiers, all amplicons were pooled and analyzed using 454 sequencing. The obtained datasets were analyzed using the VIIP pipeline. In three patients, two new mutations in *CDH23* and *OTOF* were identified. For another patient, the etiology of deafness was unclear, and no causal mutation was found. In a fifth patient, included as a positive control, a known mutation in *TMC1* was confirmed.

6.7 Discussion

The Variant Interpretation Pipeline uses a local reference genome and gene structure data obtained online from Ensembl to construct reference transcripts. Although this approach seems unnecessary complex at first sight, it offers several advantages. First of all, data traffic is limited which avoid excessive traffic and speeds up the whole process. Secondly, this approach allows users to build their own 'mutated' reference genomes and use the gene structure data from Ensembl. This allows effects of variation to be assessed compared to a custom reference genome.

The transcript based approach allows a detailed analysis of the effects of the variants without *a priori* knowledge of the exact transcript being active. This offers two advantages. First of all, the variant interpretation is made 'future proof'. Rather than categorizing a variant as intronic or intergenic on a gene level, a variant is categorized on the transcript level. The human reference genome is relatively stable and has not changed that much over the last few years. However, new gene structures are added to Ensembl frequently. Rather than reanalyzing the whole variation database when transcripts are added to have a new conclusion per gene, interpretation of only the new transcripts can be added to the database. Secondly,

the information can be combined *a posteriori* with transcription data obtained using for example RNA-sequencing or qPCR. Combining both datasets allows a quick interpretation of the exact variant with respect to the transcript being transcribed.

Currently, the pipeline relies heavily on the gene structures obtained from Ensembl. Consequently, only human variants and variants in other species present in Ensembl can be analyzed. Unfortunately, plant, viral, and bacterial genomes are not present in the standard (vertebrate) Ensembl. However, EnsemblPlant, EnsemblBacteria and similar databases have the same structure and offer similar APIs as the regular Ensembl. The pipeline could be modified to work with these databases as well, although we have not extensively tested those approaches.

The splicing effects are calculated using the tool from the fruitfly.org website. It is a valuable tool, but it is questionable whether a generic splicing tool predicts splicing effects of different species adequately.

The Variant Interpretation Pipeline can be integrated into the VIP pipeline, which allows the annotated results to be shielded away from the raw data. Combined with a visualization system such as the genome browser we presented, allows the data to be shared with collaborating scientists around the world. The system allows specific users access to specific parts of the analysed data and guarantees data integrity.

The annotated variants correspond with the results obtained using commercial software such as Alamut. However, the Variant Interpretation Pipeline calculates the effects for other transcripts than the default transcript, does it very easily in batch, allows easy data export and perhaps most importantly, does it for free. Free software is of paramount importance in the research world as it allows tools to be used by a broader research audience.

6.8 Conclusion

We designed an open-source pipeline – the Variant Interpretation Pipeline – that can be combined with the Variant Identification Pipeline to form the VIIP. The VIIP identifies and annotates genetic variants automatically. Variants are annotated as intergenic, exonic, intronic, 3'-UTR or 5'-UTR. Exonic variants are further categorized as synonymous or non-synonymous. Furthermore, splicing effects are assessed and dbSNP status is retrieved, allowing a detailed interpretation of a variant.

The pipeline is designed to batch process Roche/454-resequencing data from raw data to variant interpretation. Furthermore, VIIP can carry out interpretation of external lists of variants.

The pipeline allows an easy integration in a visualization framework to form a three-tier structure. The pipeline has shown to be a reliable tool to identify causal mutations in several diseases such as breast cancer, Marfan and Loeys-Dietz Syndrome and genetic deafness.

6.9 Availability and requirements

- **Project name:** Variant Identification and Interpretation Pipeline
- **Project home page:** http://athos.ugent.be/VIIP
- **Operating system(s):** Unix/Linux
- **Programming language:** Perl
- **Other requirements:** MySQL, (VIP and BLAT)
- **License:** GNU LGPL

6.10 Authors' contributions

JDS developed the Variant Interpretation Pipeline and combined it with the Variant Identification Pipeline to form the VIIP, implemented the annotated results into the H2G2 genome browser, set up the three-tier architecture, analyzed the sequencing data of all the clinical sample experiments including variant identification, interpretation, and custom analyses to optimize pre-sequencing labwork. GT developed the H2G2 genome browser used to visualize results. WVC supervised development of the pipeline. JDS authored the manuscript on variant interpretation (Advancing the Variant Identification Pipeline). All other authors were involved in sample recruiting, sequencing, and (co)-authored the 3 manuscripts concerning the medical samples.

Adapted from:

Joachim De Schrijver et al. (2012), **Identifying genomic variation in a large number of samples using 3D-pooling and next-generation sequencing**, in preparation

7 SNP-CUB3: Using 3D-pooling and next-generation sequencing

7.1 Abstract

Recent advances in short-read next-generation sequencing allow genetic mutation screening in research and diagnostics. Not only has the Illumina platform passed the 100 base pair sequencing length, also data volumes per sequencing run have exploded. These recent developments allow scientists to analyze large patient numbers in a single experiment.

Using a 3D-pooling strategy, where each sample is determined by a unique set of three different sample pools, it is possible to reduce the number of samples to be analyzed dramatically, allowing a larger number of samples to be analyzed in a single sequencing experiment.

We developed a pipeline that allows easy analysis of 3D-pooled next-generation sequencing data. The pipeline contains a module to test different parameters and experimental designs by generating simulated datasets that are consequently analyzed using the pipeline. SNP-CUB3, a manual, and a dataset for testing purposes are freely available at http://athos.ugent.be/snpcube.

7.2 Introduction

Second-generation sequencing, up to a couple of months ago collectively known as next-generation sequencing (NGS), keeps advancing at a fast pace and data volumes are expanding faster than ever [244]. Currently, Illumina (San Diego, CA, USA) offers several platforms such as the Genome Analyzer II (GAII) and the HiSeq2000 which yield around 100 Gb and over 500 Gb of sequencing data per experiment respectively.

Roche/454 (Branford, CT, USA) offers the GS-FLX which has a much lower yield (around 500 Mb). However, the GS-FLX system can sequence up to 600bp fragments and the upcoming GS-FLX+ platform up to 1000bp fragments whereas Illumina is only able to sequence much shorter fragments (typically 36-150bp). The GS-

FLX is currently being explored as a tool to aid molecular diagnosticians in the analysis of PCR based single locus conditions [227, 240].

As the different Illumina platforms are gradually moving towards longer sequencing lengths (it is expected that sequencing lengths will continue to increase over time), PCR based diagnostics will become a possibility on these platforms as well.

Due to the high throughput of second-generation sequencing, PCR-based locus selection strategies are suited for multiplexing (combining different PCR products in a single experiment) or pooling (combining different samples in a single experiment) to make optimal use of the capacity of the sequencer.

By 3D-pooling different samples, one can increase the total number of samples that can be analyzed in a single experiment by a larger number. SNP-CUB3 is a pipeline that identifies genomic variation in 3D-pooled Illumina sequencing experiments.

7.3 3D-pooling

3D-pooling is a strategy where the total number of samples to be analyzed is drastically reduced by allocating each sample to three different sample pools. Imagine an experiment with 27 (3^3) samples, each of these samples could be allocated a point in a 3*3*3 cube (Figure 7.1).

In our 27 sample example, sample pools are generated each containing 9 samples. The sample pool shown in red in Figure 7.1 contains the 9 samples of the bottom xy-plane. Three such sample pools can be constructed in the xy-plane (bottom, middle, top), three in the yz-plane (left, middle, right), and three in the xz-plane (front, middle, back). 9 sample pools can be generated in total each containing 9 samples.

Each of the original 27 samples is represented by the intersection of a unique combination of 3 sample pools. For example, the grey sample in the cube in Figure 7.1 is formed by the intersection of the bottom sample pool in the xz-plane (red), the left sample pool in yz-plane (blue), and the front sample pool in the xy-plane (green). In total, 27 such intersections can be formed, and thus all 27 samples are formed by a unique intersection of 3 sample pools.

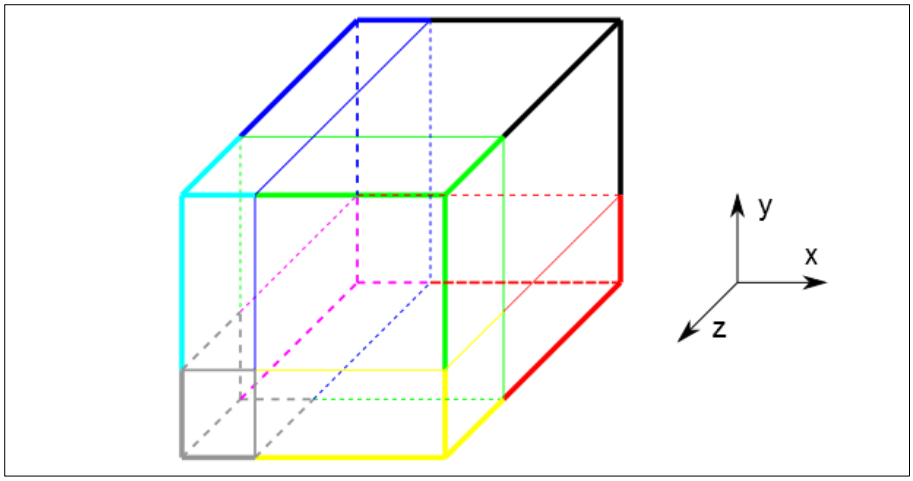


Figure 7.1: *Schematic representation of 3D-pooling using 27 samples. Each of the original 27 samples is represented by the intersection of a unique combination of 3 sample pools. The grey sample at the bottom left is formed by the intersection of the bottom sample pool in the xz-plane (red), the left sample pool in yz-plane (blue), and the front sample pool in the xy-plane (green).*

Pooling dramatically reduces the hands-on time needed in the laboratory as fewer sample pools need to be processed separately prior to sequencing. It allows more samples to be analyzed in a limited number of sample pools as fewer multiplex identifiers can be used to process a larger amount of samples. Imagine a situation where 1000 (10^3) samples need to be analyzed. It is unpractical to use 1000 multiplex identifiers and separately sequence each of these samples. However, 3D-pooling allows each sample to be represented by 3 different sample pools, and the total number of sample pools to be reduced to 30. The complexity is reduced from x^3 (total number of samples in the cube) to 3x (total number of planes in the cube) (Table 7.1).

Table 7.1: *Overview of the reduction of complexity using 3D-pooling*

Cube dimension	Samples	Pools	Samples/pool
2	8	6	4
3	27	9	9
4	64	12	16
6	6^3	18	36
n	n^3	3n	n
10	1000	30	100

Genomic variation present in one of the 27 samples should be present in each of the 3 sample pools corresponding with that specific sample. Variation which is present in a unique combination of three sample pools can be allocated to a specific sample (Figure 7.2). Variation in sample pools can be identified using for example Illumina sequencing.

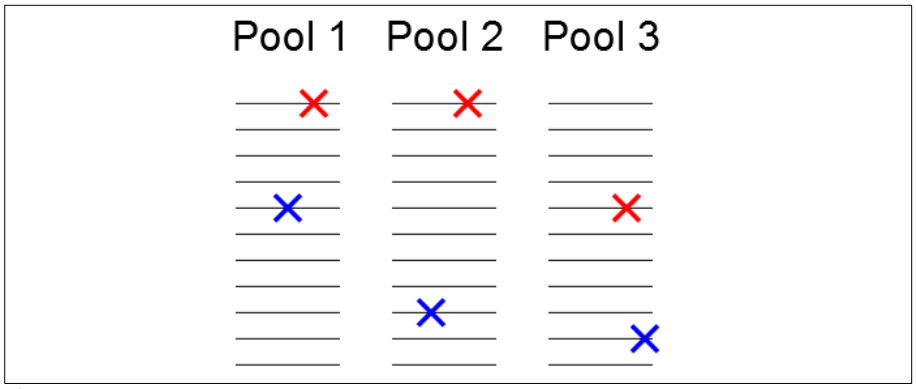


Figure 7.2: *Demonstration how pooling can detect a variant in a certain sample. Imagine the sample of interest (Sample 1) is the intersection of three pools (Pool 1, Pool 2, and Pool 3). Each of the pools contains sequences from other samples as well. The crosses indicate variation detected using sequencing. The red crosses are variations presents in each of the three pools and thus variation present in Sample 1. The blue crosses indicate variation originating from other samples which are also included into the sample pools.*

7.4 Implementation

SNP-CUB3 is an open-source pipeline written in Perl. The different scripts interact with a MySQL database (through *DBD::mysql*) and uses R (through *Statistics::R*) to calculate three-way similarities (using a chi square test).

The memory footprint is usually low (in the range of some GBs depending on the specifics of the dataset) and the database consumes very little space (MBs range) as only variant positions and coverage on the variant positions are stored.

The pipeline is available under the GNU Lesser General Public License (LGPL) from http://athos.ugent.be/snpcube. We expect to release updated versions of the pipeline in the future.

7.5 Methodology

SNP-CUB3 uses a MySQL database to store intermediary and final results. An initial Perl script will create all the necessary database structures. Furthermore, SNP-CUB3 consists of three main modules which need to be run consecutively.

In a first step, a file containing the composition of the sample pools is processed. Each sample is associated to a unique combination of three different sample pools An example of such a composition file is given in Appendix File 7.1. This processing step effectively builds the cube and links sample pools to samples.

Each sample pool should be represented by a fastq file obtained using Illumina sequencing (or a similar sequencing technique yielding the same file format). Sample pools need to sequenced separately or split into separate fastq files afterwards. In a second step, each of these fastq file is mapped separately onto a set of chosen reference sequences (either a set of short reference sequences or entire chromosomes) using Bowtie [193]. All variants reported by Bowtie are extracted from the output file and a variation database is constructed per sample pool. Rather than storing variants of each sequence separately, data is aggregated per sample pool using an algorithm similar to the one used to generate variation reports in the Variant Identification Pipeline [235]. For each variant in the sample pool, the reference position, wild type and variant allel, coverage, absolute frequency, and relative frequency are stored in the database. This is done for each of the sample pools.

In a third and final step, data is processed per sample using the sample pool composition obtained in the first step. Each of the three sample pools corresponding to a sample are grouped and analysed together. Variants specific for a certain sample should be present in each of the three associated sample pools (as demonstrated in Figure 7.2).

Assuming that variants are unique for a certain sample (which is highly probable in a screening or discovery setting) the variant frequencies should be low, but similar in each of the three pools. To assess the similarity of the variant frequencies a $\chi 2$-test is performed through R. The rationale is that unique sample variants which are present in each of the three sample pools should have similar relative frequencies.

Sample variants which are present in each of the three sample pools are stored in a separate sample database together with additional parameters such as coverage in the three samples, absolute frequencies in the three samples, the $\chi 2$ p-value etc.

These final results are stored per sample and can be accessed through any SQL-browser or dumped as a flat text file using the dumper script of the pipeline.

Furthermore, a simulation module is available to generate and analyze a dataset using certain parameters such as sample/pool composition, PCR error rate, sequencing error rate, sequencing read size, number of reads per pool and sample variant frequency. Variants are inserted randomly into the simulated data and the scripts assesses if the variants are still detectable using the pipeline.

7.6 Validation

The pipeline was validated on a dataset where a single known variant was present in one of the samples. The dataset contained 27 samples (and 9 sample pools). The pipeline identified 216 putative variants. The known variant was the only true positive when the variants were ranked using the coverage in each of the 3 pools and the p-value was taken into account. An overview of the detected variants is included in Appendix File 7.2.

7.7 Discussion

Pooling dilutes the frequency of a single variant in the pool which can make it harder to discover true variants. However, a variant that is seen in the three pools of a certain sample can be considered a true variant. The aggregate false positive rate is very low as the error rate x for false positives drops from x to x^3. Indeed, the chance that a certain sequencing error occurs in each of the 3 sequenced sample pools making up a certain sample is x^3 compared to the chance x that it occurs by accident in only one sample pool.

However, the method has some drawbacks. By combining different samples in a pool, variant frequencies of a single sample are diluted and it is obvious this dilution has limits. Homozygous variants in a single sample in a 3*3*3 cube experiment (27 samples, 9 pools, 9 samples per pool) will have a diluted frequency of ~ 11% (100/9)

in the pool. This is still detectable through the backgroud of sequencing errors and PCR error rate.

It is optimal to have a perfect cube (i.e. all dimensions are equal), but this is not a strict requirement. However, one should keep in mind that the expected frequencies of variants in each sample pool will not be equal anymore and the $\chi 2$-test will lost its value. Different reasons can lead to a non-perfect cube: the number of samples available, the number of pools that can be separated using barcoding or physical separation. However, there is a simulation module available which enables user to test the effects of different experimental setups.

The methodology works optimal on datasets where the expected number of variants is low and where the variants are unique per sample. Datasets where a variant is present in for example half of the samples will cause the variant to be detected in almost all of the sample pools and will erroneously call the variant present in almost all samples. Nevertheless, the pipeline performs well in an environment where one is looking for rare and unique variation.

7.8 Conclusion

I developed SNP-CUB3, a pipeline which allows 3D-pooled experiments followed by Illumina sequencing to be analysed and variants to be identified.

3D-pooling followed by next-generation sequencing allows the number of sample pools to be analysed to be decreased from X^3 to 3X. This drastically reduces hands-on time needed in the lab. This means that by physically separating sample pools (using different sequencing lanes) or by barcoding different pools (using pool specific tag sequences within a pool) it becomes possible to analyze thousands of samples in a single experiment.

7.9 Availability and requirements

- **Project name:** SNP-Cub3
- **Project home page:** http://athos.ugent.be/snpcube
- **Operating system(s):** Unix/Linux
- **Programming language:** Perl
- **Other requirements:** MySQL, R
- **License:** GNU LGPL

7.10 Authors' contributions

JDS wrote the pipeline and authored the manuscript.

Part 3: Identifying methylation and transcription using second-generation sequencing

Adapted from:

Joachim De Schrijver, Geert Trooskens, Linda J.W. Bosch, Pierre Dehan, Beatriz Carvalho, Leander Van Neste, Gerrit Meijer, Steve Baylin, Wim Van Criekinge (2012), **Genome-wide total RNA and MBD-sequencing in HCT116 and DKO cells as a global re-expression model**, *BMC Genomics* submitted

8 Genome-wide total RNA and MBD-sequencing in HCT116 and DKO cells

8.1 Abstract

8.1.1 Background

Methylation of the promoter or the first exon can suppress gene expression. DNA methylation is known to be causative for diseases including cancer, where expression of tumor suppressor genes can be suppressed by methylation. The genome-wide profile of several colorectal cancer cell lines (including HCT116) was determined using Methyl-Binding-Domain (MBD) capturing followed by next-generation sequencing. Directional next-generation RNA-sequencing was used for genome-wide total RNA (mRNA, miRNA, snoRNA …) profiling of the HCT116 and the *DNMT1* and *DNMT3B* double knock-out HCT116 cell line (DKO). By correlating these profiles, a genome-wide view of the interplay between methylation and RNA expression in this colorectal cancer model was obtained.

8.1.2 Results

The RNA-sequencing data confirmed a relative low abundance of natural antisense transcription in HCT116 and ruled out methylation as the regulator of antisense transcription. Combining the RNA expression data in HCT116 and DKO and DNA methylation data in HCT116, pathways were identified that are regulated by methylation. Using this approach, known methylation-regulated protein coding genes, miRNAs and snoRNAs were confirmed in a single comprehensive sequencing effort. Furthermore, a new putative colorectal cancer gene, *LAYN*, which is suppressed by DNA methylation in colorectal cancer cell lines, was identified.

8.1.3 Conclusion

RNA-sequencing combined with MBD-sequencing of both HCT116 and DKO is a powerful approach allowing a genome-wide view on the interplay between DNA methylation and RNA expression.

8.2 Introduction

DNA-methylation is a process wherein a methyl group is added to a genomic residue. This process has been described just a decade after the initial discovery of the structure of DNA [151]. Cytosine-guanine dinucleotides are only separated from each other by a single phosphate group and are often referred to as CpG dinucleotides. Although cytosine-5-methylation (5mC) in a CpG context is considered to be the most important DNA modification in humans, other changes such as cytosine-5-hydroxymethylation (5hmC) [152] and non-CpG methylation [153] have been described as well. Methylated cytosines spontaneously mutate to thymines over time and most CpGs in humans are methylated [154], hence CpG dinucleotides are underrepresented in the human genome [44]. Unmethylated CpGs are frequently clustered together in CpG islands which are mainly located at the 5' regulatory region of a gene [155]. Gene promoter hypermethylation (mostly CpG island hypermethylation) is generally associated with gene silencing [160].

DNA-methylation is considered to be a major epigenetic mechanism involved in cell regulation. In healthy cells it is involved in processes like tissue differentiation [156], aging [157], memory formation [158], and X chromosome inactivation [159].

Both hypomethylation of oncogenes [161] and hypermethylation of tumor suppressor genes [162] are associated with cancer development. Furthermore, the methylation profile of a cell can be used as diagnostic [163], prognostic [164] or treatment response marker [165].

Since the discovery of DNA methylation, different methods have been used to detect and characterize methylation patterns. Until recently, methylation specific PCR (MSP) [245] and targeted bisulfite sequencing [166] were most commonly used for this purpose. However, these techniques require an *a priori* region of interest and have a limited throughput. Discovery of methylation patterns was mostly reliant on indirect methods, e.g. gene re-expression after epigenetic reactivation [1]. More recently, methylation specific arrays such as Illumina's Infinium II platform were developed, measuring the methylation status of more than 27,000 CpG dinucleotides [246] with an increased throughput, however, still dependent on assumptions regarding the regions of interest.

Recent developments in next-generation sequencing techniques [247] allow for assessing methylation patterns on a genome-wide scale. Characterization of ge-

nome-wide DNA methylation patterns using bisulfite sequencing requires tens of gigabases of sequencing data to reach a minimal coverage, implying practical limitations even by current standards. An enrichment step prior to bisulfite sequencing [167] can offer a solution, but limits the genome-wide character. Recently, MBD-isolated genome sequencing was developed enabling the methylation profile to be determined on a genome-wide level. This technique uses the native human methyl-CpG-binding-domain (MBD) protein to capture and isolate methylated genomic regions which are subsequently sequenced using next-generation sequencing technology [171]. Using this technology, the Map of the Human Methylome (MHM), which lists all the regions of the human genome that can possibly be methylated (MHM), was developed.

Up until recently, there was only a limited amount of techniques available to assess the transcriptional state of cells. Hybridization techniques have evolved from Northern blot [248] to more sensitive techniques such as microarray chips [249]. Although recent microarrays are very sophisticated, it remains a technique that can only detect *a priori* defined transcripts. Next-generation sequencing allows genome-wide RNA-sequencing and offers the highest possible resolution combined with a high-throughput approach unseen before [250]. Recently, Illumina developed a protocol to specifically sequence single-stranded RNA sequences (rather than double-stranded cDNA which makes it impossible to discriminate between sense and antisense sequences). This approach, known as directional RNA-sequencing, enabled a more detailed and accurate analysis of the transcriptome.

The influence of the methylation status of the promoter region on expression of a downstream gene in Colorectal Cancer (CRC) cell lines such as HCT116 has been investigated before. The double knockout (DKO) cell line, a HCT116 cell line where both methyltransferases *DNMT1* and *DNMT3B* are knocked out, shows a minimal methylation activity [251]. Comparison of the expression in HCT116 and DKO and analysis of the methylation profile of HCT116 has identified important genes in CRC, of which the transcription is under methylation control.[1, 252]. However, the methylation status is often verified by methylation specific PCR (MSP) or array-based methods, while expression is typically quantified using microarrays or reverse transcriptase PCR (RT-PCR). We analyzed the epigenome and transcriptome of HCT116 and DKO cell lines using MBD-sequencing and RNA-sequencing enabling a genome-wide scale analysis of the relationship between methylation and expression in CRC. Total RNA was sequenced directionally offering a broad and accurate

picture of the transcription in HCT116 and DKO. Furthermore, transcription and methylation in 7 other CRC cell lines was assessed as well.

8.3 Methods

8.3.1 Cell line culturing

HCT116, SW1398, SW480, RKO, LS513, HT29, Colo320 and Colo205 colorectal cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM; BioWitthaker, Lonza, Verviers, Belgium) supplemented with 10% heat-inactivated fetal calf serum, 100 U/ml penicillin (Astellas Pharma BV, Leierdorp, The Netherlands), 100 g/L streptomycin (FisioPharma, Palomonte, Italy), and 2 mmol/L L-glutamine.

The DKO cell line was cultured according to the protocol of the Johns Hopkins University School of Medicine [251].

8.3.2 RNA and DNA extraction

Total RNA from cell lines was isolated using TRIzol reagent (Invitrogen, Breda, The Netherlands), and subjected to purification using RNeasy Mini Kit (Qiagen, Venlo, The Netherlands) according to the manufacturer's instructions. DNA from cell lines was extracted using a column based method (QIAamp microkit; Qiagen, Venlo, The Netherlands) and sheared by sonication.

8.3.3 Total RNA sequencing

Total RNA fragmentation - to obtain 200bp fragments on average - was performed on a Covaris S2 using the following settings: duty cycle 10%; intensity 5; 200 cycles per burst during 270 sec.

Further library preparation was carried out using a modified version of the Illumina 'Directional mRNA-Seq Sample Preparation' protocol using total RNA instead of mRNA. In short, the following steps were performed. RNA phosphatase treatment, followed by a PNK treatment. Illumina v1.5 small RNA 3' adapters ligation, followed by the 5' adapter ligation. Reverse transcription and amplification using RT-PCR. DNA fragment enrichment with AMPure XP beads.

Ribosomal DNA (rRNA) was depleted from the RNA fraction using Illumina's Duplex-Specific Thermostable Nuclease (DSN) normalization protocol for bidirectional mRNA sequencing (application note 15014673). The time used for ssDNA renaturation, an essential step in the protocol, differed for the different HCT116 samples. Renaturation time was equal to 5, 6, 7, 8, 9, 10 or 11h depending on the sample. Renaturation time for DKO, SW1398, SW480, RKO, LS513, HT29, Colo320 and Colo205 was 5h.

Single read clusters were generated on the Cluster Station following the manufacturer's guidelines. Single read 36bp sequencing was carried out on a Genome Analyzer II (GAIIx).

8.3.4 Methylation sequencing

Methylation profiles of the cell lines were determined using MBD-sequencing. A custom MBD-protein and protocol was used to capture and isolate the methylated DNA sequences in the HCT116 cells. The MethylCollector kit (Acive Motif, La Hulpe, Belgium) was used to capture and isolate the methylated DNA sequences in the SW1398, SW480, RKO, LS513, HT29, Colo320 and Colo205 cell lines.

Paired-end read clusters were generated on the Cluster Station following the manufacturer's guidelines. Paired-end 36bp sequencing was carried out on an Illumina GAIIx.

8.3.5 Mapping

36bp sequence reads (both the single-end and paired-end reads) were mapped onto the human reference genome (NCBI build 37.3) using Bowtie (software version 0.12.7) [192] allowing 0 mismatches in the seed (i.e. first 28 nucleotides) and requiring unique mapping.

8.3.6 Quantification of gene expression

Gene expression was assessed on a transcript level, rather than on gene level. The expression level of each available transcript of each gene was quantified without assuming *a priori* whether a certain transcript was transcribed or not. The Ensembl [253] database (version 56_37) contains 979,901 different transcripts which were consequently quantified.

RNA-expression levels of each transcript were determined by calculating the average base coverage of each transcript. To compare samples - with different total coverages - the expression in each sample was normalized using the total number of mapped reads of the first sample as a baseline number and the ratio of the number of mapped reads of the first sample and the other samples as a normalization factor. The expression level of each transcript was then multiplied by that normalization factor.

8.3.7 Correlation analyses

For the correlation analyses, expression values were rescaled using a log2 transformation. Using the expression level of each transcript as a data point, pair-wise correlations were calculated between all samples.

8.3.8 Map of the Human Methylome

Version 2 of the Map of the Human Methylome (MHM), which contains data from 80 different samples, was used. The Map of the Human Methylome was obtained from the public MHM website http://h2g2.ugent.be/mhm.

8.3.9 Gene ontology analysis

Gene ontology (GO) analysis to determine enriched processes was carried out using the web service GOrilla [254].

8.3.10 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) was carried out using the GSEA tool [255, 256]. Analyses were performed using version 2.07 with 1000 permutations and standard settings. KEGG, Biocarta and Reactome gene sets were downloaded from the GSEA website (http://www.broadinstitute.org/gsea/msigdb/genesets.jsp). A p-value of 5% was used to retain significantly enriched gene sets.

8.4 Results & discussion

8.4.1 Data overview

RNA-sequencing experiments were carried out in different sequencing runs. The RNA from the HCT116 cell lines was processed using seven different DSN renaturation times to verify whether it was possible to enrich the RNA sample with the most or least abundant RNA molecules using the DSN procedure. The RNA originating from these seven different RNA samples yielded 43,707,489 reads on average with an average mapping percentage of 44.62%, which is a commonly observed percentage given the rather stringent mapping settings. The DKO sample yielded 65,760,903 reads with a mapping percentage of 39.96%. A third run contained seven additional cell lines (SW1398, SW480, RKO, LS513, HT29, Colo320 and Colo205) and yielded 23,353,168 reads per lane on average with an average mapping percentage of 55.06%. A detailed overview of the data is given in Table 8.1.

Table 8.1: *Overview of the RNA-sequencing statistics of the different samples*

ID	Type	DSN†	Total reads	Read length	Map %	Mapped bases	Norm. factor‡
1	HCT116	5 h	44,655,791	36 bp	46.51	727 Mb	1.0000
2	HCT116	6 h	41,864,891	36 bp	50.44	741 Mb	0.9835
3	HCT116	7 h	44,234,094	36 bp	43.45	677 Mb	1.0806
4	HCT116	8 h	44,922,006	36 bp	49.53	784 Mb	0.9335
5	HCT116	9 h	44,786,272	36bp	51.72	819 Mb	0.8967
6	HCT116	10 h	44,374,213	36 bp	36.35	569 Mb	1.2874
7	HCT116	11 h	41,117,667	36 bp	34.35	498 Mb	1.4704
8	DKO	5h	65,760,903	36 bp	39.96	946 Mb	0.7904
9	HT29	5h	18,774,280	36 bp	56.13	379 Mb	1.9709
10	RKO	5h	28,294,657	36 bp	56.59	576 Mb	1.2988
11	LS513	5h	26,499,509	36 bp	60.16	574 Mb	1.3027
12	Colo205	5h	28,443,713	36 bp	57.40	588 Mb	1.2721
13	Colo320	5h	28,544,354	36 bp	58.74	603 Mb	1.2387
14	SW480	5h	26,563,670	36 bp	51.19	490 Mb	1.5273
15	SW1398	5h	25,126,270	36 bp	45.25	409 Mb	1.8269

† The incubation time that was used during the ribosomal RNA (rRNA) removal step. This procedure is described in detail in the material & methods section.

‡ A factor to rescale coverages to compensate unequal sequencing depths.

8.4.2 Total RNA DSN normalization & reproducibility

Pair-wise correlations between all seven HCT116 samples themselves, with different DSN incubation times, are very high (R^2-values between 97% and 98%). Pair-wise correlations between each of the HCT116 samples and the DKO sample are lower than the R^2-values of the HCT116 samples (approximately 80%). The renaturation time in the DSN process has no clear effect on the composition of the RNA pool indicating that the process is fast and incubating longer than 5 hours does not invoke removal of less abundant sequences in the further steps of the DSN protocol, therefore this incubation time was chosen for processing the other cell lines.

Figure 8.1 illustrates that similar renaturation times results in apparent smoother correlations. However, there is no straightforward effect of the DSN process due to increased removal of either high or low abundant transcripts, as a clear rotation of the regression cannot be observed when comparing Sample 1 with Sample 2 - 7 (i.e. going from left to right in the top row of Figure 8.1).

8.4.3 Overview of expression in HCT116 and DKO

As expected, the HCT116-DKO correlation R^2-values are lower than the correlation R^2-values among the HCT116 samples themselves. A large number of transcripts showing no expression in HCT116 show expression in DKO, which is in correspondence with the previous findings that (promoter) methylation generally suppresses gene expression. In HCT116 there were 21,912 genes with no expression, i.e. no coverage in any of the available transcripts of a particular gene, compared with only 17,841 genes in DKO, corresponding with a net increase of expressed genes of 9.45%.

8.4.4 Antisense expression

Previous genome studies have proposed that natural antisense transcription (NAT) could be a common phenomenon with a regulatory function. The mode of action of these NAT transcripts is suspected to be, among other modes of actions, similar to the one of small RNAs in the RNA interference (RNAi) process [257][258].

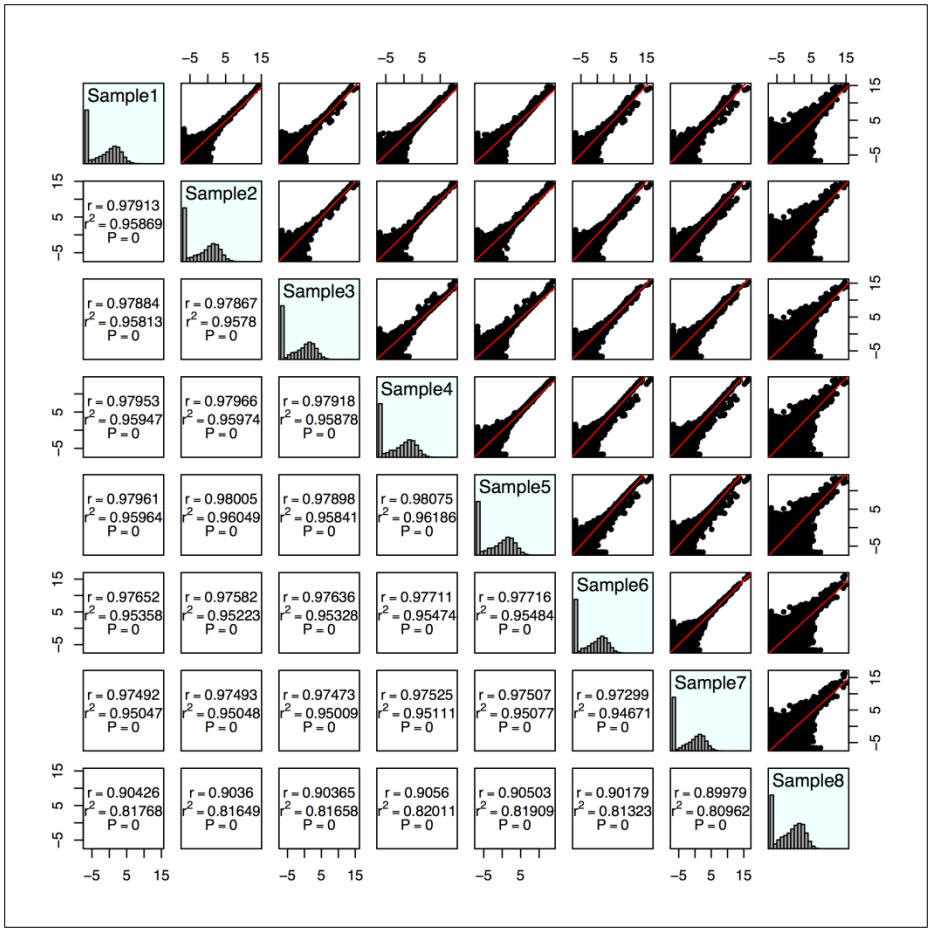


Figure 8.1: *Overview of the correlation of the expression values (measured as log2 of the expression value of each transcript) between the different samples. The upper half of the image shows the pair-wise correlation plots. The upper left panel in the image shows the correlation between Sample 1 and Sample 2, the upper right panel shows the correlation between Sample 1 and Sample 7 etc. The diagonal shows transcript expression frequency histogram plots for each individual sample. The bottom half of the image shows the correlation coefficients and P-values corresponding with the correlation plot on the other side of the diagonal.*

In HCT116, 99 Mb of the entire human genome (~3% of the genome) was detected as transcribed from the Watson strand, compared with 106 Mb from the Crick strand. Despite this relative abundant transcription, only 1.58Mb (~1.5%) of this transcription overlapped on both strands. The exact proportion of the genome

showing transcription in both sense and antisense direction is even lower as the majority of the regions showing both sense and antisense transcription are regions having overlapping genes on both strands, essentially showing sense transcription in both orientations.

In DKO, the numbers are very similar with the Watson strand having 143Mb transcriptionally active regions and the Crick strand 155Mb. The percentage of regions showing transcription on both strands is slightly higher than in the native HCT116 cells, but still low (3.4 Mb, ~2%).

Thus, at first sight it appears that there is no widespread antisense transcription in HCT116 or DKO. In addition, there is no indication that the presence of methylation in HCT116 cells is suppressing antisense transcription or that the lack of methylation in DKO is inducing antisense transcription, as the percentages of NAT are similar in both HCT116 and DKO.

However, because the mode of action of NAT is similar to the RNAi process, transcription of a short antisense RNA, complementary to a longer transcript, could have a regulatory function. Therefore, an additional analysis was carried out to find out whether NAT is a focused event, rather than a dispersed one affecting a large number of genes. On the transcript/gene level, the analysis is complicated by the presence of overlapping genes. To alleviate this problem, only genes not overlapping with another gene on the other strand were considered in the subsequent analyses.

In the native HCT116 cell line, 15,402 genes showed expression in the sense orientation, compared with 3,062 genes showing antisense expression. Approximately half of the genes (49.46%) showing antisense expression show some degree of expression in both orientations. Taken together, out of the 15,402 genes showing sense transcription, 1,520 genes show both sense and antisense transcription (9.87%), similar to previous observations [259].

In DKO, 17,799 genes showed expression in the sense orientation, compared with 4,288 genes showing antisense expression. 1,948 (45.43%) of the genes showing antisense expression show both sense and antisense transcription. Of the 17,799 genes showing transcription in the sense orientation, 1,948 genes (10.94%) show both sense and antisense transcription.

8.4.5 Global re-expression (direct and indirect re-expression)

The expression level of each transcript in HCT116 was compared with the expression level in DKO. A transcript was considered to be re-expressed when one of the 2 following conditions was met: A) no expression in HCT116, while the expression level in DKO > 0.1 or B) a ratio of the expression level in DKO over that in HCT116 > 10. Using these criteria, a total of 8,479 transcripts were found to be re-expressed in DKO originating from 3,671 different genes.

As total RNA was sequenced, rather than enriched mRNA or miRNAs, a complete picture could be formed of the type of genes being re-expressed. As expected, the majority of the genes are protein coding genes comprising 58.16% of the total number of genes being re-expressed (Figure 8.2).

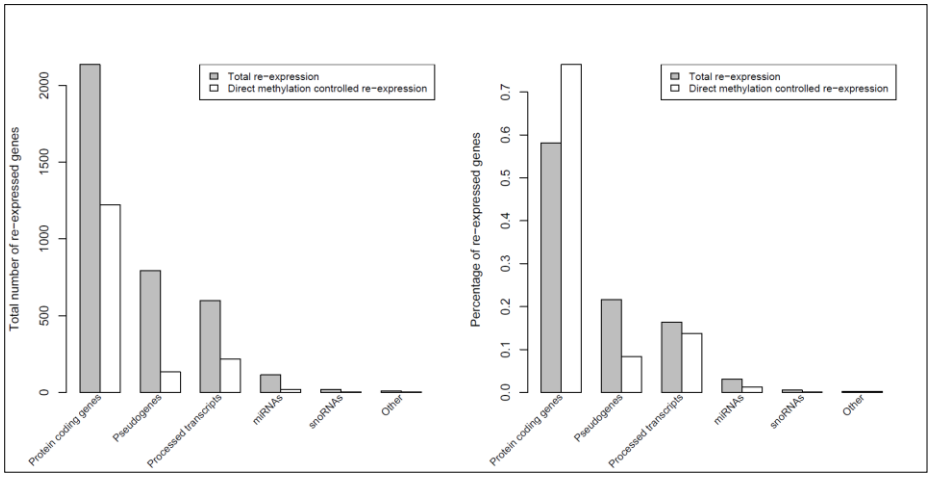


Figure 8.2: *Overview of the absolute number of genes re-expressed in DKO split by type of gene (left) and relative portion of each gene type in the total number of re-expressed genes (right). The grey bars shows data for global re-expression, the white shows data for re-expressed genes showing promoter methylation (direct re-expression).*

An overview of the functions of the re-expressed genes was obtained using Gene Ontology (GO) terms. The most enriched processes included immunological responses (e.g. cytokine pathway), responses to cellular damage (e.g. responses to ionizing radiation), processing of cellular metabolites (e.g. metabolic processing, biosynthetic processes), and more general cell (cycle) regulation (e.g. regulation of

gene expression, regulation of nuclear division). An overview of the top-20 enriched GO processes is given in Table 8.2; a more extensive overview is given in Appendix File 8.1.

Table 8.2: *Top-20 enriched Gene Ontology (GO) processes in DKO (global re-expression)*

GO number	Gene Ontology Description	P-value
GO:0010468	regulation of gene expression	8.98E-8
GO:0010212	response to ionizing radiation	8.99E-8
GO:0060255	regulation of macromolecule metabolic process	3.09E-7
GO:2000112	regulation of cellular macromolecule biosynthetic process	4.46E-7
GO:0051252	regulation of RNA metabolic process	6.34E-7
GO:0006355	regulation of transcription, DNA-dependent	9.49E-7
GO:0010556	regulation of macromolecule biosynthetic process	1.17E-6
GO:0019219	regulation of nucleobase-containing compound metabolic process	1.46E-6
GO:0080090	regulation of primary metabolic process	1.51E-6
GO:2001141	regulation of RNA biosynthetic process	1.53E-6
GO:0051171	regulation of nitrogen compound metabolic process	3.44E-6
GO:0019222	regulation of metabolic process	6.56E-6
GO:0009889	regulation of biosynthetic process	1.1E-5
GO:0031326	regulation of cellular biosynthetic process	1.24E-5
GO:0022402	cell cycle process	1.61E-5
GO:0034097	response to cytokine stimulus	1.75E-5
GO:0031323	regulation of cellular metabolic process	1.94E-5
GO:0045840	positive regulation of mitosis	2.39E-5
GO:0051785	positive regulation of nuclear division	2.39E-5
GO:0051246	regulation of protein metabolic process	3.71E-5

Despite the genetic differences between HCT116 and DKO being relatively small (2 dysfunctional genes), the cellular differences caused by the lack of methylation appear to be fairly large. This can be explained by the fact that DKO cells have largely lost their DNA methylation functionality that triggers a set of regulatory pathways and in turn changes the cell's expression pattern completely. This is expected, as promoters methylated in HCT116 are no longer methylated in DKO, subsequently activating these genes. But the re-expression effects in DKO cells are not limited to genes directly suppressed by methylation in HCT116. Several genes, not methylated in HCT116, become active in DKO upon global demethylation, without evidence that promoter demethylation lies at the basis.

8.4.6 HCT116 methylation profile

To obtain a subset of re-expressed genes that are directly regulated by methylation, a methylation profile of HCT116 needed to be constructed. The Map of the Human Methylome (MHM) was used to obtain such a profile and was generated using a wide variety of different cell types and tissue types and is in effect an overview of all the regions of the human genome that can possibly be methylated. Through the use of the MHM functional methylation units were constructed. To determine the actual methylation profile of HCT116, only HCT116 MBD-peaks overlapping with the MHM peaks were retained for further analyses, while non-overlapping peaks were discarded. The idea behind this approach is that the MHM contains validated regions that can be methylated in biological samples.

This approach yielded 283,276 (18.65%) HCT116-methylated regions out of a total of 1,518,879 regions present in the MHM. An analysis of the promoter regions of all the annotated genes revealed that methylation is mainly concentrated in the promoter and first exon (a region spanning from 1000bp upstream to 1000bp downstream the TSS) (Figure 8.3).

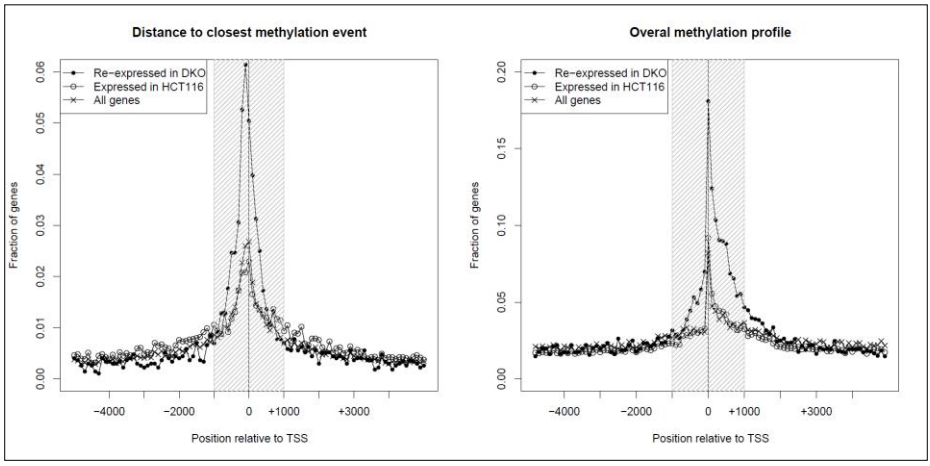


Figure 8.3: *Histogram of the distance to the closest methylation event both upstream and downstream relative to the transcription start site (left) and global methylation pattern (right). The grey shaded area indicates the region (-1000bp to +1000bp) that was considered as the regulatory region.*

8.4.7 Re-expression under direct control of methylation

Transcription under direct methylation control can be distinguished from global re-expression by combining the re-expression data with the methylation data. Re-expressed transcripts under direct methylation control are transcripts having a methylated promoter in HCT116 and are re-expressed in DKO. The promoter was defined as the region spanning 1000bp up and downstream the transcription start site (TSS). Re-expression was determined in the same manner as described before.

Through combining methylation and re-expression data, 3,672 transcripts were found to be re-expressed in DKO, originating from 1,600 different genes. The composition of the genes being re-expressed differs from the composition obtained in the global re-expression analysis. The percentage of re-expressed protein coding genes is higher (76.38% vs. 58.12%) and the percentages of other types of genes are lower. In general there is a bias towards protein coding genes (Figure 8.2). Re-expression of protein coding genes is more directly regulated by methylation than pseudogenes, processed transcripts or regulatory transcripts such as miRNAs or snoRNAs. A possible explanation is that environmental or intracellular signals trigger the expression of regulatory genes which in fact regulate the methylation of the protein coding transcripts. Promoter methylation of protein coding genes is a downstream effect of a response to a changing environment which in turn has an effect on the expression pattern of the cell.

Analysis of the GO terms revealed that different types of genes are re-expressed when considering promoter methylation compared to global re-expression. In the case of direct methylation control, the enriched GO terms are mostly linked to methylation (e.g. macromolecule methylation), cell cycle regulation (e.g. cell cycle phase) and regulation of the genomic structure (e.g. protein-DNA complex assembly, DNA modification). However, there are no 'metabolic' or 'biosynthetic' GO terms enriched in the list of genes under direct methylation control, indicating that metabolic and biosynthetic pathways are not directly regulated by methylation.

A possible explanation might be the following. In absence of methylation functionality, as is the case in DKO, cells are substantially deregulated. As methylation is no longer suppressing several genes, the increased expression of these genes will lead to the presence of undesired or potentially damaging proteins. So, although the tumor suppressor genes are expressed in DKO, which can revert the cancerous phenotype, the cell still undergoes a significant amount of stress. Other forms of feed-

back/regulation kick in to try to restore balance in the cell. The cell reacts to this hostile intracellular environment by activating all sorts of pathways to get rid of potentially damaging proteins (hence the enriched inflammatory and metabolic functionalities observed in the global re-expression analysis). The top 20 GO terms is given in Table 8.3; a more extensive overview is given in Appendix File 8.2.

Table 8.3: *Top-20 enriched Gene Ontology (GO) processes in DKO (direct methylation control)*

GO number	Gene Ontology Description	P-value
GO:0010212	response to ionizing radiation	1.02E-6
GO:0022403	cell cycle phase	5.7E-6
GO:0022402	cell cycle process	1.66E-5
GO:0032259	methylation	5.83E-5
GO:0043414	macromolecule methylation	5.83E-5
GO:0034728	nucleosome organization	1.02E-4
GO:0006334	nucleosome assembly	1.02E-4
GO:0010468	regulation of gene expression	1.1E-4
GO:0071824	protein-DNA complex subunit organization	1.26E-4
GO:0065004	protein-DNA complex assembly	1.26E-4
GO:0071901	negative regulation of protein serine/threonine kinase activity	1.7E-4
GO:0043407	negative regulation of MAP kinase activity	1.7E-4
GO:0006730	one-carbon metabolic process	2.02E-4
GO:0071514	genetic imprinting	2.08E-4
GO:0009629	response to gravity	2.53E-4
GO:0040029	regulation of gene expression, epigenetic	2.77E-4
GO:0006304	DNA modification	2.77E-4
GO:0006306	DNA methylation	2.77E-4
GO:0006305	DNA alkylation	2.77E-4
GO:0051171	regulation of nitrogen compound metabolic process	3.16E-4

8.4.8 Transcriptional activation under methylation control

It is generally accepted that promoter methylation silences downstream gene expression. However, it was recently shown that exon methylation causes the methylated exon to be retained when splicing the pre-mRNA, hence methylation is activating transcription on the exon level [260]. To investigate whether methylation might activate expresion on the gene level, genes that are silenced in DKO and expressed in HCT116 were identified. Using analogous criteria as before, i.e. expression in DKO, no expression in HCT116, and methylation in the promoter, 339 different

genes were identified, most of them i.e. barely meeting the criteria. Indeed, when a gene ontology analysis was performed, only one GO term (regulation of secretion; p-value = 3.86 10E-4) was significantly enriched. Gene silencing in DKO is not likely to be a regulated phenomenon.

8.4.9 Comparative pathway analysis

12 KEGG, 11 Reactome, and 11 Biocarta pathways were upregulated in DKO. Interestingly, most of the upregulated pathways are related to the immune system such as interleukin pathways, complement system pathways, hematopoietic pathways, toll like receptor signaling etc. An overview of all the upregulated pathways is given in Appendix File 8.3.

The previous GO analyses indicated that many of the re-expressed genes had a metabolic function. However, when looking at the pathway level, many immunological pathways are re-expressed in DKO. This could back the previous claim that DKO cells undergo stress caused by the increased transcription, which is countered by immunological responses. However, this could also have a more fundamental biological functions as it has been shown recently that a dysfunctional immune system could lead to cancerous cells [261]. Methylation could indirectly suppress the immune system, which in turn could lead to a cancerous situation. Demethylation in DKO removes the suppression, eventually leading to an upregulation of immunological pathways.

Several known cancer pathways (PI3K/AKT/mTOR, MAPK/ERK, P53 and WNT) were investigated using the gene sets present in Biocarta. None of these four pathways was significantly up or downregulated in DKO.

8.4.10 Overview of the re-expressed protein coding genes

The most re-expressed protein coding gene with HCT116 promoter methylation, *UCHL1*, has been described before as being silenced in cancerous cells. In a study by Schuebel et al. [1] *UCHL1* was the most re-expressed gene in DKO. Recently, it was shown that *UCHL1* can induce cell cycles arrest and apoptosis and that the gene is frequently silenced in breast cancer [262]; clearly showing a link with cancerous phenotypes.

The second most re-expressed gene, *LAYN*, is relatively unknown. This is the first time this gene was identified as a potential marker for CRC, silenced by DNA methylation in HCT116. Notably, *LAYN* has recently been described as a p53 downstream mediator of cellular senescence. Methylation of *LAYN* bypasses the senescence of the cell, and causes cells to keep growing and potentially leading to cancerous conditions [263]. However, the exact role of this gene in the development of CRC is topic of further investigation.

Interestingly, *LAYN* was never identified in earlier re-expression studies in CRC. A possible explanation is that a large proportion of expression analyses was and still is carried out using microarrays, and this was the sole method of analyzing transcription genome-wide before the introduction of next-generation sequencing. A frequently used microarray platform, the Agilent microarray (Agilent Technologies), has only one single probe located in the *LAYN* gene. The probe is at the start of the first exon of the gene at location 111,411,233 on chromosome 11. In DKO we were able to detect transcription throughout the whole gene, but not at the location where this probe is located. Several reasons can explain the potentially erroneous probe location i.e. the gene was not annotated correctly or an unannotated splice variant is being produced in DKO. A detailed overview of the *LAYN* gene location is given in Figure 8.4, showing a clear methylation peak at the promoter region and re-expression in DKO. Similar observations were made for other genes, e.g. *UCHL1*, *VIM*, *CFTCL*, and *PAGE1* (data not shown).

Another top candidate, *VIM*, is a well-known type III intermediate filament (IF) protein gene. Previously, it was shown that promoter and first exon methylation of *VIM* can act as a colon cancer specific marker [264]. Although a completely different methodology was used here compared to the study performed by Chen et al. [264] the suppression of *VIM* in HCT116 by promoter methylation could be confirmed, as illustrated by the high *VIM* expression in DKO.

CFTCL (also known as *BORIS*), known to be methylated in cancer cells, can regulate alternative transcription start sites [265]. Very low expression was observed in HCT116 together with promoter methylation. In DKO however, strong expression was detected.

Furthermore, *PAGE1*, a member of the GAGE family of CT-genes, coding for tumor antigene proteins, shows a strong re-expression. GAGE genes are known to be actively transcribed in certain cancer types, and can be re-expressed using chemical

treatment in cells without expression, indicating that they are probably regulated by methylation [266].

Finally, *SPG20* was recently described as a potential early detection marker for colorectal cancer [267], *TDRD12* as chemically re-expressible in salivary gland adenoid cystic carcinoma [268] and *DPEP3* as being involved in breast cancer [269]. *TUSC3* (also known as *N33*) is a tumor suppressor gene which has been previously been observed as epigenetically silenced in ovarian cancer [270]. Figure 8.5 gives an overview of the 20 most re-expressed genes in DKO with promoter methylation in HCT116. An overview of all the re-expressed genes is given in Appendix File 8.4.

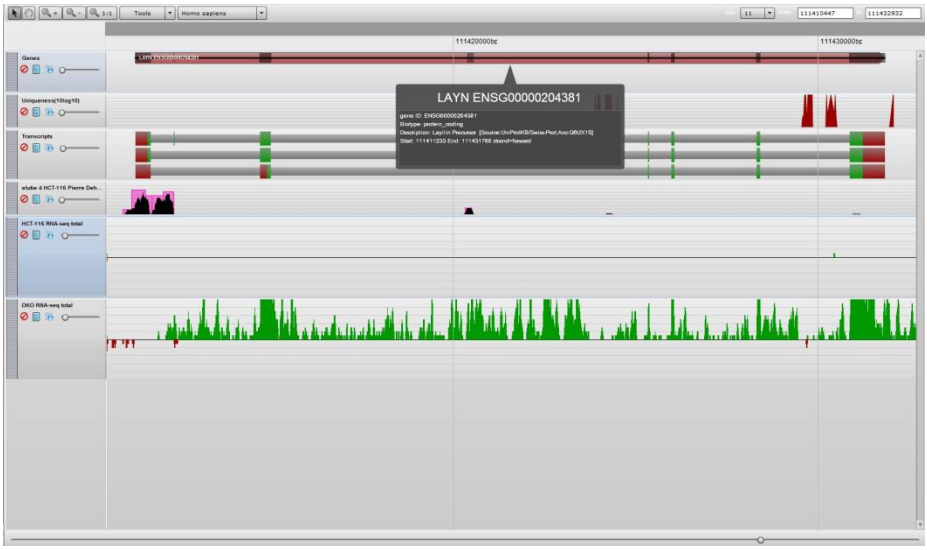


Figure 8.4: *Tracks are described from top to bottom. The 'Uniqueness track' indicates the degree of repetitiveness of a region. The 'Transcripts' track displays the different transcripts that are available in Ensembl. The 'elutie 4 HCT-116 Pierre Dehan' track displays the methylation status. A peak indicates that methylation was detected at that location. The 'HCT-116 RNA-seq total' track shows the expression of the HCT116 cell line. Green peaks above the central horizontal bar indicate transcription in the sense direction. Red peaks under the central horizontal bar indicate transcription in the antisense direction. As LAYN is located on the sense orientation, we should expect only 'green transcription'. The DKO RNA-sequencing total track shows the expression of the DKO cell line. From these tracks it is clear that there is a methylation peak in the promoter region of the* LAYN *gene. There is no expression in the wild type (HCT116) cell line and there is strong re-expression in the DKO cell line.*

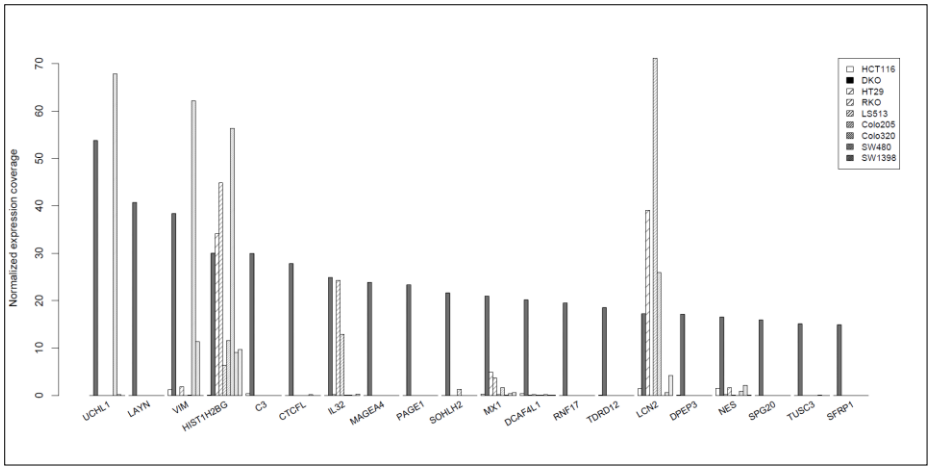


Figure 8.5: *These 20 protein coding genes show re-expression in DKO compared to HCT116 and have a methylation event in the promoter in HCT116. Expression is shown for HCT116, DKO, HT29, RKO, LS513, Colo205, Colo320, SW480 and SW1398 cell lines. The genes are ordered by the expression in DKO.*

8.4.11 Overview of the ncRNA results

Interestingly, the most abundantly (0 vs. 191.17 coverage in HCT116 and DKO respectively) re-expressed gene in DKO is the small nucleolar RNA (snoRNA) *SNORD116*. Although its biological function is currently still uncertain, it has been linked with the development of the Prader-Willi Syndrome (PWS), a genetic imprinting disease [271]. A total of 19 snoRNAs are re-expressed in DKO with 4 of them (*SNORD116*, *SNORD64*, *SNORD123*, *SNORD107*) clearly stronger re-expressed than the others. An overview of all the re-expressed snoRNAs is given in Appendix File 8.5.

Several snoRNAs, including *SNORD116*, have recently been implicated in alternative splicing [272]. The fact that the expression of these genes is regulated by methylation and these genes have a potential role in the regulation of alternative splicing, confirms the link between alternative splicing and methylation, which has been reported using different strategies [260, 273].

miRNA 146a (*hsa-mir-146a*) is one of the top ten re-expressed genes and is the most re-expressed microRNA in this HCT116/DKO experiment and is known to be involved in different cancers e.g. pancreatic cancer and breast cancer [274]. Several

studies indicate that miRNA 146a is downregulated in cancerous samples, including pancreatic cancer [275], hepatocellular carcinoma (HCC) [276] and NK/T cell lymphoma [277]. Although the underlying mechanism is unclear, it has been shown that the invasive character of pancreatic cancer can be suppressed by treating the cancer cells with isoflavones leading to an upregulation of *hsa-mir-146a* [275]. Isoflavones are known suppressors of DNA-methylation [278, 279]. This study confirms the re-expression of the important miRNA *hsa-mir-146a* in DKO which is suppressed in HCT116. Although the miRNA is upregulated in DKO, there is no detectable promoter methylation in HCT116 indicating that the re-expression is an effect of global demethylation, rather than the effect of specific promoter demethylation,

8.4.12 Comparison with DAC/TSA re-expression in HCT116

Gene re-expression in HCT116 has been studied before using 5-aza-2'-deoxycytidine (DAC) and trichostatin A (TSA). In an extensive micro-array study by Schuebel et al. [1], re-expression in HCT116 following DAC/TSA treatment was categorized as either top tier or second tier depending on the magnitude of re-expression in the treated cells. Of the 31 validated HCT116 to tier genes, 29 (93.5%) were confirmed in this study. One of the two genes not confirmed here, *PPP1R14A*, was slightly re-expressed, but did not pass the re-expression criteria. The other gene, *TLR2*, was re-expressed but has no methylation peak in the vicinity of the promoter. Possibly, the *TLR2* transcript being expressed is not yet annotated as there is upstream expression which commences downstream of a densely methylated region. A *de novo* assembly of the expressed *TLR2* transcript using Velvet [176] failed because the coverage in that region was too low. Therefore, it remains to be found out which transcript is transcribed for *TLR2*. An overview of the *TLR2* gene region is given in Appendix File 8.6.

The entire list of first and second tier genes from Schuebel at al. [1] contains 897 non-validated methylation candidates. Of those 897 430 (47.93%) were confirmed as re-expressed and under control of promoter methylation. Of the 1,600 identified genes, 1,170 were newly identified (i.e. not identified in the study by Schuebel et al.). There are several reasons why certain genes are not confirmed. Firstly, a gene can be silent in HCT116, but not (sufficiently) re-expressed in DKO. Secondly, a gene can be sufficiently re-expressed, but not readily associated with a methylation event. ENSG00000151572 can be used as example of lack of methylation. While this gene is re-expressed (0.14 vs. 1.96 coverage in HCT116 and DKO respectively), the

nearest methylation peak is 100kb upstream, making it hard to assume this peak has a direct effect on the expression of the gene.

8.4.13 Stability of promoter methylation in other CRC cell lines

The re-expressed genes obtained by comparing HCT116 and DKO were analyzed in other CRC cell lines, i.e. SW1398, SW480, RKO, LS513, HT29, Colo320 and Colo205. For each of these cell lines, a methylation and expression profile was determined using total RNA and MBD-sequencing. An overview of the RNA data of the additional cell lines is given in Table 1. In addition, correlating expression of all the transcripts between the different cell lines is determined as described earlier. The seven expression profile of the additional cell lines are similar to expression profile of the HCT116 cell line with R^2-values ranging from 0.72 (Colo320) to 0.78 (SW480) (Figure 8.6). The R^2-values obtained by comparing the expression values of the cancer cell lines to the expression values of DKO are lower than the values obtained by individually comparing the expression values to the HCT116 expression profile. As expected, this shows that CRC cell lines are more similar to one another than to DKO.

As mentioned, 3,672 transcripts were re-expressed in DKO. The methylation pattern of these transcripts was assessed in all seven additional CRC cell lines. Individual methylation profiles clearly differ, with Colo205 showing the largest overlap (86.27%) with the HCT116 promoter methylation events and LS513 the smallest (73.58%). Of the 25,704 transcripts in all the CRC cell lines combined, 20,066 were methylated in the promoter (78.06%), indicating that the methylation pattern across different CRC cell lines is very similar. This becomes even more apparent when compared to a number of random transcripts where only 27.74% exhibits promoter methylation.

The expression of the genes re-expressed in DKO was verified in the other cell lines. Out of the 20,066 transcripts having a methylated promoter, 18,066 (90.03%) are silenced, compared to 4,227 out of 7,130 (59.28%) in the set of random transcripts. Details for all the different cell lines are given in Table 8.4.

Many of the genes silenced in HCT116 and re-expressed in DKO upon demethylation are silenced in the other cell lines. Colo320 appears to differ from the other cell lines as both UCHL1 and VIM are not silenced (Figure 8.5).

The top genes of this study (e.g. *UCHL1*, *VIM* and *LAYN*) and several cancer genes previously described in the literature [280-282] were investigated. The genes present in the PI3K/AKT/mTOR, MAPK/ERK, P53 and WNT pathways were also characterized in all the discussed cell lines. In total, expression of 67 genes was quantified in HCT116, DKO and the additional cell lines. An overview of the expression and methylation status of these 67 genes is given in Appendix File 8.7. A separate overview of the genes involved in the mentioned pathways is given in Appendix File 8.8.

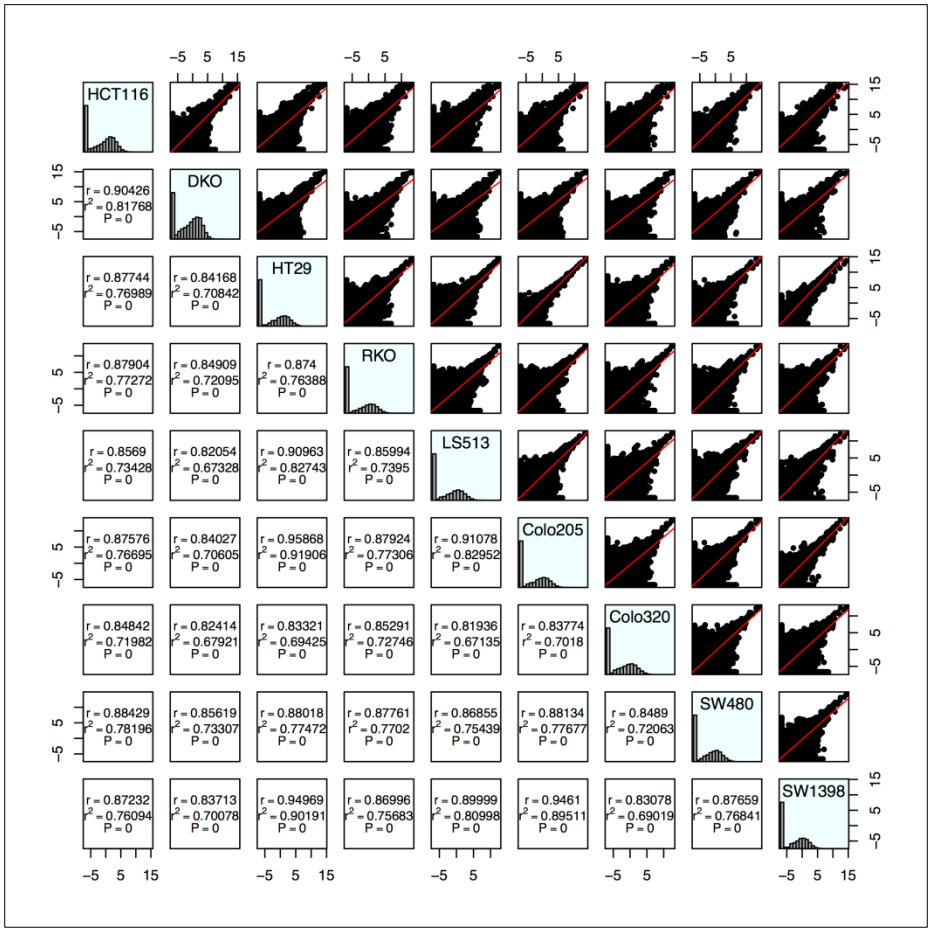


Figure 8.6: *Overview of the correlation of the expression values (measured as log2 of the expression value of each transcript) between the HCT116, DKO, HT29, RKO, LS513, Colo205, Colo320, SW480 and SW1398.*

8.5 Conclusion

This current study analyzed the expression profile of both HCT116 and DKO using next-generation directional RNA sequencing. Using total RNA, rather than poly-(A) enriched mRNA, all types of genes can be quantified in a reproducible fashion. With the development of directional RNA sequencing, it is, for the first time, possible to quantify the expression of overlapping genes (Appendix File 8.9).

The HCT116/DKO comparison allowed the identification of a plethora of (epigenetically) silenced, cancer-associated genes. Among the top hits, several known cancer, more specifically colorectal cancer-related (e.g. *VIM*) genes, were observed, indicating that this next-generation sequencing approach is a powerful and robust method to detect relevant genes.

Through combining MBD- and RNA-sequencing, it is possible to distinguish direct and indirect re-expression in DKO. Genes under direct methylation control have enriched functions of methylation, DNA interaction or genetic imprinting. Genes under indirect control, i.e. re-expressed in the case of global demethylation but no promoter methylation, are linked to metabolic activity or cellular defense responses.

Furthermore, the genome-wide sequencing experiments offer advantages over other approaches such as microarrays or methylation specific PCR (MSP) to respectively measure expression and methylation. As evidenced by the identification of *LAYN*, the information used to determine the position of microarray probes is of key importance. Genome-wide sequencing is independent of assumptions regarding potential transcriptional active regions and exact annotations of transcripts. As mentioned before, alternative splicing has recently been linked to methylation [273, 283]. This could lead to a situation where a certain microarray probe is located on an exon that can be differentially methylated, and thus could be skipped or retained depending on its methylation status. Using microarrays in re-expression experiments, this could lead to certain transcripts not being detected as re-expressed when in fact only a single exon was methylated and, thus, skipped. Although expression microarray manufacturers try to cope with this by incorporating additional probes for known alternative transcripts, hits problem is completely alleviated by sequencing.

When comparing both the methylation profile and expression profile of HCT116 to other CRC cell lines, both profiles show a high degree of similarity. Especially the methylation pattern is very similar in all the different cell lines. Indeed, the methylation pattern is more stable than the expression pattern and is therefore a prime candidate to be used as a marker.

Table 8.4: *Comparison of HT29, RKO, LS513, Colo205, Colo320, SW480 and SW1398 to HCT116 and DKO*

Cell line	HCT116 methylation cores	Silencing promoter methylation	R^2 RNA HCT116	R^2 RNA DKO
HT29	2,930 (79.79%)	2,641 (90.14%)	0.76989	0.70842
RKO	2,938 (80.01%)	2,733 (93.02%)	0.77273	0.76385
LS513	2,702 (73.58%)	2,507 (92.78%)	0.73428	0.67328
Colo205	3,168 (86.27%)	2,843 (89.74%)	0.76695	0.70605
Colo320	2701 (73.56%)	2,287 (84.67%)	0.71982	0.67921
SW480	2,752 (74.95%)	2,390 (86.85%)	0.78196	0.73307
SW1398	2,875 (78.30%)	2,665 (92.70%)	0.76094	0.70078
Sum of 7 CRC cells	20,066 (78.06%)	18,066 (90.03%)	-	-
Random gene set	7,130 (27.74%)	4,227 (59.28%)	-	-

8.6 Authors' contributions

JDS analyzed the RNA sequencing experiments, the re-expression experiments, and carried out all additional analyses. GT helped generating the methylation cores. LB, PDH and BC carried out the RNA and MBD-sequencing labwork. LVN was involved in the comparison with previous experiments and helped designing the study. LVN and SB provided the DKO cells and were involved in numerous discussions concerning gene re-expression. WVC and GM were involved from the beginning of the study JDS and WVC designed the study. WVC supervised the study. JDS authored the manuscript. All co-authors read and approved the manuscript.

Part 4: Conclusions and future perspectives

Conclusions

Where Sanger sequencing has been used in the past to confirm known and identify new genomic variants of simple genetic diseases at a reasonable cost, next-generation sequencing allows more complex diseases, more samples, and more possible variants to be identified at a lower cost and a faster speed. It is expected that next-generation sequencing will replace Sanger sequencing as the golden standard in diagnostics in the near future [97, 223], especially with the recent roll-out of the smaller benchtop instruments [79]. However, easy to use data analysis tools are required to allow this transition to take place.

The Variant Identification Pipeline (VIP) [235] I developed and subsequent improvements allow 454 re-sequencing experiments to be analyzed using a user-friendly command-line interface. The combination of 454 sequencing and the VIP pipeline is clearly a powerful combination to analyse multiplexed amplicon re-sequencing experiments, demonstrated by the efficient identification of variants in breast cancer [58, 227, 235], Marfan and Loeys-Dietz Syndrome [240], and genetic deafness [241]. These proof of concept studies open to the door to cheaper, more accurate, and faster (next-generation) diagnostics in the clinic.

The VIP was initially developed using human sequencing data. However, the application scope is not limited to human samples. VIP is perfectly capable of analysing viral, bacterial, and plant sequencing data, provided a reference sequence is available. For example, mutations specific for a certain plant phenotype or mutations specific for certain bacterial subpopulations could be identified using the VIP.

The VIP outperforms existing commercial software packages such as AVA, especially when detecting (small) indels. This can lower the cost of diagnostic sequencing as fewer samples will need to be analysed using more costly methodologies such as Sanger sequencing, which are nowadays typically used to detect indels.

Furthermore, several of the modules of the VIP pipeline can be used to process pyrosequencing data in settings different than diagnostic or variant identification. For example, I have used the MID trimming module to trim off MIDs of different bacterial genomes which were sequenced together in a single experiment. After

efficient trimming, these different bacterial genomes were successfully assembled, reducing the cost of *de novo* sequencing. As shown in chapter 5, the VIP can be used to optimize multiplexed PCR reactions. Several of the generated reports can pinpoint in which sample which PCRs are failing. In general, the Variant Identification Pipeline allows a thorough assessment of the pre-sequencing labwork and can improve experimental (diganostic) setups alltogether.

Recently, Ion Torrent technology entered the market. Although the underlying sequencing methodology is completely different than the methodology used in 454 sequencing, both sequencing data types are relatively similar. An initial analysis of Ion Torrent (PGM) datasets indicate that the VIP can be used to analyse such datasets. Currently, it remains unclear whether the error profile and homopolymer related issues are completely similar. As Ion Torrent sequencing is cheaper than 454 sequencing, this technology could further push down the cost of sequencing based diagnostics, or variant identification alltogether.

Identified variants can be annotated using the Variant Interpretation Pipeline. This is of special interest in a diagnostic setting as it easily allows variants to be categorized as having a biological impact or not. Many commercial packages currently available still do not offer automated batch processing and make the interpretation of long list of identified variants both tedious and expensive. The annotated variants data is optimized to be visualized using for example the H2G2 genome browser [239]. This genome browser improves the interaction with clinicians, which frequently have a limited experience with data analysis.

However, there are some challenges when using the Variant Identification Pipeline. The pipeline is specifically designed to analyse pyrosequencing data and inherent homopolymer issues remain a problem. Since the publication of the VIP, other researchers have confirmed the findings that variants near homopolymeric regions should be treated with caution [284, 285]. However, treating homopolymers with caution does not remove the inherent problems. Continuous work is done to try to improve the basecalling accuracy and reduce some of the homopolymer related problems alltogether. For example, recently HPCall was developed, a methodology which improves the accuracy of homopolymers. The methodology can improve 454 sequencing data and most probably Ion Torrent sequencing data as well [286].

Another specific VIP challenge relates to the database approach. When datasets become too big, storing each of these sequences (sometimes billions of sequences)

in a database becomes unfeasible. This problem is partially offset by the recent development of smaller benchtop machines, which generate smaller datasets than before, allowing the VIP to be used.

Generally, Illumina offers a throughput which is too high to allow amplicon sequencing to be practical and cost-effective as coverages are unnecessary high. Using SNP-CUB3, the total number of samples that can be analyzed in a single experiment is drastically increased, drastically reducing the cost per sample and allowing a more cost-effective alternative. SNP-CUB3 allows rare variants to be identified in large pooling experiments on Illumina sequencing instruments. This has important clinical implications; rare variants can be identified at a very low cost. Furthermore the variants identified by SNP-CUB3 can be annotated using the Variant Interpretation Pipeline, as the Variant Interpretation Pipeline can process external lists of variants.

Second-generation sequencing not only allows genomic variation to be identified, but also allows quantification of transcription and identification of methylation patterns. Using a combination of RNA- and MBD-sequencing in both a colorectal cancer cell line (HCT116) and a methylation knock-out (DKO) cell line, superior results were obtained in comparison with older techniques. Genes known to be under control of methylation could be identified faster and cheaper. Furthermore, new (cancer) genes under the control of methylation could be identified, such as *LAYN*.

Using next-generation sequencing to quantify transcription offers advantages over hybridization based approaches such as arrays. For example, there is no need for *a priori* information in regards to the sequences of the probes included onto an array (as demonstrated in the case of the *LAYN* gene). Furthermore, the genome-wide approach offers an unbiassed look and allows specific transcripts of certain genes to be identified and characterised.

Furthermore, an important distinction could be observed between global re-expression and promoter regulated re-expression in the HCT116/DKO experiment. Global re-expression activates metabolic and defensive gene functions where promoter re-expression activates genes typically associated with methylation such as for example cell cycle regulation. Further research in primary tumor samples can tell us to what extent this is a cell line specific, colorectal cancer specific or cancer specific phenomenon. Nevertheless, these findings could have an impact on the

understanding of the association between methylation/demethylation and cancer biology.

The directional total RNA sequencing data has provided interesting insights into the global transcription biology of cancer cell lines. Earlier microarray studies indicated that large portions of the genome were transcribed and that antisense transcription was frequently present. Data obtained in the HCT116/DKO experiment has indicated that background transcription is much lower than could be expected from these microarray experiments.

However, the sequencing approach suffers from some challenges. First of all, there are some difficulties with the normalization of the MBD- and RNA-sequencing data as the sequencing capacity is fixed and the sequencing signal is thus diluted over the number of signals (i.e. the genes that show expression or methylation). It is expected that a combination of identification using sequencing and quantification using array based methods will be the preferred approach.

Furthermore, cell lines are non-perfect substitutes for primary samples. It is often not feasible to do both MBD- and RNA-sequencing genome-wide on a large numbers of primary tumours. Despite the drawbacks, cell lines offer a good starting point to investigate methylation biology and identify putative methylation markers. The (new) methylation markers detected in the HCT116/DKO system offer a good starting point for more detailed research and validation in primary samples. Novel methylation markers (methylated in colorectal cancer and not in normal samples) could be used as a biomarker for early detection of colorectal cancer. Furthermore, some non-coding RNAs showed a very intense re-expression in DKO compared to HCT116. These RNA molecules could also function as a potential biomarker for early detection of colorectal cancer.

The analysis in the HCT116/DKO system has shown the potential of second-generation sequencing in identifying new methylation markers in colorectal cancer. This approach could be transferred to other types of cancer, provided a methylation knock-out is available.

Future perspectives

The development of methylation capturing strategies such as MBD-sequencing has revolutionized the field of epigenetics. MBD-sequencing offers clear advantages over for example ChIP-on-chip approaches. As is the case in RNA-sequencing, the lack of need for probe sequences allows researchers to have a broader view. However, normalization can be difficult; again, a combination of sequencing and array based methodologies allow an elegant combination of qualitative and quantitative approaches. However, genome-wide bisulphite sequencing would eliminate these problems.

Currently, genome-wide bisulphite sequencing is still not feasible. However, researchers are testing the boundaries and combining enrichment methods with bisulphite sequencing to have semi-genome-wide bisulphite sequencing [167]. As bisulphite sequencing offers the highest possible resolution, it still remains the golden standard of methylation sequencing. Combining a capturing enrichment (using MBD) with subsequent bisulphite sequencing would offer an unprecedented genome-wide high-resolution methylation overview. Currently, sample preparations are not efficient enough to allow such experiments to be carried out. It is expected that in the future such experiments will be feasible. Single base resolution methylation data would eliminate the need for methylation cores and combined with the expression data obtained in the HCT116/DKO experiment would allow a more detailed analysis to be carried out.

Complex sample preparations and ingenious sequencing strategies have enabled researchers to investigate other sources of variation such as internal ribosome entry sites (IRES), alternative translation initiation sites (aTIS), alternative splicing, RNA editing, ribosome profiling, and much more. It is expected that next-generation sequencing will be a valuable asset in future genetics research and broader life science research.

Several exciting opportunities arise. The HCT116/DKO dataset could be combined with a ribosomal sequencing dataset. This would add a third layer of information and would allow the relationship to be determined between methylation, transcription, and transcripts actively being translated into proteins.

Third-generation sequencing instruments are already available on the market and it is expected that other innovative instruments will follow in the coming years. These third-generation instruments have the possibility to be a driver of the discovery of new biology as they allow single cells to be completely profiled, modifications such as hydroxy-methylation to be identified, and complete RNA molecules to be sequenced. Full length RNA sequences would unlock the door to unbiassed splicing analysis. Currently, full length RNA sequences are often rebuilt from short reads provided by the sequencing instruments using algorithms [216]. Although a good approximation of the reality, it is not perfect. Analysing the full length RNA sequences in the HCT116/DKO system would make it possible to investigate the link between methylation patterns and specific splicing events. Furthermore, it could lead to the discovery of rare and cancer specific transcripts which could be used as a stratifying or early detection biomarker.

Although the sequencing throughput obtained with the early second-generation sequencing instruments was significantly larger than the sequencing output obtained using Sanger sequencing, early efforts to improve the next-generation sequencing technology were initially focused on increasing the throughput further. However, nowadays, a single large genome (e.g. the human genome), and especially an exome, can easily be sequenced with sufficient depth in a single experiment. Recently, the development of even larger throughputs has been abandoned, or at least slowed down, by for example Illumina. More efforts are put into improving the *user experience* of the sequencing instruments. In effect, the sample preparation is being improved, turn-over time reduced, sample pooling improved, and costs reduced . This development led to the release of smaller, faster, cheaper and more flexible benchtop instruments such as Illumina's MiSeq and Roche's GS Junior. These instruments have a purchase price of approximately $100,000 and an operating cost of approximately $1000 per experiment, making the technology accessible to many smaller research organizations [79]. Once these technologies spread to many of such smaller research organizations, easy-to-use tools such as the Variant Identification and Interpretation Pipeline will enable them to analyze small sequencing experiments, speeding up existing research and enabling new research to be conducted.

A remarkable feature of the next-generation sequencing technology is the co-evolution of wet-lab and *in silico* lab developments. Bioinformaticians were very quick with the development of specialized tools to analyze the datasets obtained by second-generation sequencing. For example, improved sequence aligners such as

Bowtie and BWA were specifically designed to align sequences obtained using the revolutionary second-generation sequencing instruments. There often was, and still is, an important symbiotic relationship between the wet-lab and the *in silico* lab to develop library preparations that ease data processing and to develop tools that allow easy library preparations. Next-generation sequencing is a truly interdisciplinary field that needs specialist both in the lab and both in front of the computer.

As data processing and data analysis is taking more and more time (as data sets become bigger and bigger) and effective sequencing times are becoming shorter and shorter, there is a clear need for easy, fast, and standardized analysis pipelines. Some efforts have been put into developing standard frameworks such as the Galaxy framework [287, 288]. Galaxy allows pipelines to be built using a click-and-drag interface, allowing the sometimes complex command-line packages to be used by computer novices and second-generation sequencing laymen.

Furthermore, processing times can be reduced by using cloud services such as Amazon's or Google's cloud service. These behemoths offer huge amounts of their processors to be used to analyze biological sequencing data in return for a small fee. Currently, many of the familiar analysis packages offer the possibility to use the Amazon services under the hood [289].

Taking this all together, the combination of next-generation sequencing and the advanced bioinformatics strategies available have revolutionized life science research and will continue to do so. Early second-generation sequencing efforts were mainly focused on the analysis of human samples. Now, sequencers have become cheaper than ever before, operational costs are lower, and data analysis is standardized, opening the door for a giant breakthrough of second-generation sequencing into all branches of biology.

Summary & samenvatting

Summary

Individuals differ from each other, both in obvious visual characteristics such as eye color and non-visual characteristics such as the predisposition for certain diseases. The phenotypical differences are the result of different sets of proteins, and other biological molecules, being active within the cells of these differing individuals. Many of these phenotypical differences are the result of distinct genotypes, i.e. individuals having distinct genotypes will produce different biological products, and as a consequence, will differ from each other. However, epigenetic differences, such as DNA methylation and chromatin remodeling, are known to alter the expression pattern of certain genes and the corresponding proteins and thus can cause individuals to differ from each other even when they share the same genetic background.

Many genetic and epigenetic signatures are specific for certain diseases or disease stages and can be used as a valuable tool in, amongst others, identification of disease causing (epi)genetic profiles and early detection and prognosis of certain diseases. This means the analysis of the (epi)genome of unhealthy individuals can help us in both determining the underlying cause and an optimal treatment; the analysis of healthy individuals can help us in identifying high-risk individuals. For example, *BRCA1* and *BRCA2* are two genes known to dramatically increase the risk of both breast and ovarian cancer when a certain genetic signature is present. Early and accurate genetic screening is the key to locating high-risk individuals and early detection of the disease, which increases outcome altogether.

Clinicians and other (life) science scientists around the world have been using the Sanger sequencing technique since the early seventies to determine the nucleotide sequence of specific genes. Only in the beginning of the 21st century, the three billion bases that make up the entire human genome were sequenced and a draft genome became available. This reference allowed researchers to compare the genetic background of disease individuals to this human reference and identify (causal) variants.

Only a few years later, around 2007, next-generation sequencing (NGS) technology became available and revolutionized the world of genome sequencing by allowing a hitherto unprecedented throughput at unprecedented speeds and unprecedented low costs. However, with the exploding throughput and intrinsically associated

exploding data volume, intelligent computational strategies are required to both process and interpret this data.

The Variant Identification and Interpretation Pipeline (VIIP) was developed to specifically tackle this aforementioned problem of next-generation sequencing data processing and data interpretation. The VIIP processes raw sequencing data generated by the 454-sequencing instrument, at the time the most used next-generation sequencing platform, using a modular and database oriented approach. The final results of the pipeline consists of a list of genetic variants identified in the sequenced samples, interpretation of the identified variants, and many additional analysis reports allowing a thorough and detailed analysis of the sequenced samples. Furthermore, the pipeline is built modularly, allowing a classic three-tier architecture to be built and allowing good scalability. Using the VIIP, several samples were successfully analyzed and new causal variants were identified for breast cancer, Marfan and Loeys-Dietz syndrome, and genetic deafness.

For the Illumina sequencing technique, SNP-CUB3 was developed, allowing the analysis of 3D-pooled samples and variants to be identified. SNP-CUB3 builds the bridge between the high-throughput available on the Illumina platform and the relatively low throughput needed for targeted genomic resequencing, allowing a large number of samples to be sequenced and subsequently analyzed in an efficient fashion.

The superior throughput of the Illumina platform allows the methylation and expression profile of human samples to be characterized genome-wide relatively easy. Methylated DNA-fragments can be isolated from the unmethylated DNA using the methyl-binding domain (MBD) protein which only captures methylated fragments; subsequent sequencing allows the methylation profile of the sample to be determined. RNA-sequencing of the same sample allows the expression profile of that sample to be determined. By integrating both the methylation and the expression profile, the relationship between the two can be elucidated.

The methylation and expression profile of a colorectal cancer (CRC) cell line (HCT116) and the expression profile of the same CRC cell line lacking methylation functionalities (DKO) were combined to unravel the exact relationship between methylation and expression in a cancer model. The experiments confirmed the standard methylation/silencing paradigm on a genome-wide level, but also identified specific genes silenced by methylation in HCT116, which could play a causal role

in the disease phenotype. Most of the identified genes were known cancer genes, observed previously as silenced by methylation in either colorectal cancer or other cancer types. However, a new putative cancer gene, *LAYN*, was identified as a possible methylation marker.

This research has proven that next-generation sequencing (or recently referred to as second-generation sequencing) can, and most probably will, be a valuable asset in unraveling both the genetic and epigenetic background of specific disease phenotypes. Over the past few months, sequencing technology has evolved even further and third-generation sequencing instruments are becoming widely available. Third-generation sequencing has many advantages of which direct methylation sequencing and more detailed and quantitative RNA sequencing appear to be the most promising in respect to the previously described experiments. The developed methods can readily be transformed to analyze third-generation sequencing data. The VIIP has successfully been used to analyze Ion Torrent sequencing data, which has become available over the last few months. It is expected that more detailed methylation and expression data obtained by using third-generation sequencing instruments will only improve the results obtained using the Illumina technology, using the previously developed analysis pipelines.

Samenvatting

Individuen vertonen onderling zowel zichtbare verschillen zoals oogkleur als niet-zichtbare verschillen zoals de aanleg voor bepaalde ziektes. De fenotypische verschillen zijn het gevolg van verschillende sets eiwitten en andere biologische moleculen die actief zijn in de cellen van deze verschillende individuen. Vele van deze fenotypische verschillen zijn te wijten aan verschillende achterliggende genotypes; individuen met verschillende genotypes produceren verschillende biologische producten and zullen als gevolg daarvan van elkaar verschillen. Echter is er ook epigenetica, zoals DNA methylatie en chromatine hermodellering, die het expressie patroon van bepaalde genen en de corresponderende proteïnen kan veranderen. Bijgevolg kunnen individuen van elkaar verschillen ook al bezitten ze dezelfde genetische achtergrond.

Veel genetische en epigenetische signaturen zijn specifiek voor bepaalde ziektes of stadia van bepaalde ziektes en kunnen o.a. gebruikt worden ter identificatie van ziekte veroorzakende (epi)genetische profielen en vroege detectie en prognose van bepaalde ziektes. Dit leidt er toe dat het analyseren van het (epi)genoom van zieke individuen kan helpen om de achterliggende genetische oorzaak te achterhalen en een optimale therapie te bepalen en het analyseren van gezonde individuen de hoge-risico individuen kan situeren. Bijvoorbeeld, *BRCA1* en *BRCA2* zijn twee kankergenen die gekend zijn het risico op borst- en eierstokkanker sterk te verhogen indien bepaalde genetische varianten aanwezig zijn. Vroege en accurate genetische screening is de sleutel tot de lokalisatie van hoge-risico individuen en vroege detectie van de ziekte die beide uiteindelijk de kans op overleven sterk verhogen.

Dokters en andere onderzoekers hebben sinds de jaren '70 Sanger sequencing gebruikt om de nucleotide sequentie te bepalen van bepaalde genen. Slechts in het begin van de 21^{e} eeuw is men er in geslaagd om de drie miljard basen van het menselijke genoom te sequeneren en werd een kladversie van dit genoom beschikbaar gesteld. Dit referentie genoom laat onderzoekers toe om het genetisch profiel van bepaalde zieke individuen te vergelijken met deze referentie en causale varianten op te sporen.

Rond 2007, nog maar enkele jaren geleden, kwam de next-generation sequencing (NGS) technologie beschikbaar, die de wereld van het genoom sequeneren op zijn

kop zette. Deze technologie laat een tot voorheen ongezien hoge doorvoer toe aan een ongeziene hoge snelheid en aan ongezien lage kosten. Echter, de explosief toegenomen hoge doorvoer en ermee gepaard gaande grote data volumes maken intelligente computationele strategien noodzakelijk om de data te verwerken en te interpreteren.

De Variant Identification and Interpretation Pipeline (VIIP) werd ontwikkeld om specifiek dit probleem van data verwerking en data interpretatie op te lossen. De VIIP verwerkt de ruwe sequenering data van de 454-sequencing platform, indertijd het meest gebruikte platform, gebruikmakend van een modulair en databank georiënteerde benadering. De finale resultaten van de pipeline bestaan uit een lijst met genetische varianten die geïdentificeerd werden in het staal, interpretatie van de geïdentificeerde varianten en vele verschillende additionele rapporteren die een gedetailleerde analyse van het staal mogelijk maken. De pipeline is op zo'n wijze modulair gebouwd dat een klassieke 3-lagen structuur opgezet kan worden die goed schaalt. Verschillende ziektebeelden zijn succesvol geanalyseerd met de VIIP en causale varianten werden geïdentificeerd voor borst kanker, Marfan en Loeys-Dietz syndroom en genetische doofheid.

SNP-CUB3 werd ontwikkeld om 3D-gepoolde Illumina sequenering data te kunnen verwerken. SNP-CUB3 bouwt de brug tussen de hoge doorvoer aanwezig op het Illumina platform en de relatief lage doorvoer die nodig is voor beperkte hersequeneringsexperimenten. Deze ontwikkelde methodologie laat toe om een heel groot aantal stalen gelijktijdig te sequeneren en de data nadien efficiënt te analyseren.

De superieure doorvoer van het Illumina platform laat toe om zowel methylatie als expressie profielen van humane stalen relatief gemakkelijk genoomwijd te karakteriseren. Gemethyleerde DNA fragmenten kunnen geïsoleerd worden van de niet gemethyleerde fragmenten met behulp van het methyl-binding domain (MBD) proteïne dat enkel gemethyleerde fragmenten bindt. Door vervolgens enkel deze gemethyleerde fragmenten te sequeneren, kan het methylatie profiel van een staal bepaald worden. RNA-sequenering toegepast op hetzelfde staal laat toe om het expressie profiel van dat staal te bepalen. Als beide profielen geïntegreerd worden, kan de relatie tussen methylatie en expressie opgehelderd worden.

Het methylatie en expressie profiel werd bepaald van een colorectale kanker (CRC) cellijn (HCT116) en het expressie profiel werd bepaald van dezelfde cellijn waarin de

methylatie functionaliteiten uitgeschakeld (DKO) werden. Beide profielen werden nadien gecombineerd om de exacte relatie tussen methylatie en expressie te bepalen in een kanker model. De experimenten bevestigden het standaard methylatie/repressie paradigma genoomwijd, maar konden ook specifieke genen identificeren die door methylatie onderdrukt worden in HCT116 en een rol zouden kunnen spelen bij het ontwikkelen van de ziekte. Het merendeel van de geïdentificeerde genen waren gekende kanker genen, eerder geobserveerd als onderdrukt door methylatie in colorectale kanker of andere kankers. Er werd echter ook een nieuw kandidaat kanker gen geïdentificeerd, *LAYN*, dat door methylatie onderdrukt wordt en dus zou kunnen dienen als methylatie merker.

Dit onderzoek heeft aangetoond dat next-generation sequencing (of tweede-generatie sequenering) een belangrijke rol kan, en waarschijnlijk zal, spelen in het ontrafelen van de genetische en epigenetische achtergrond van specifieke ziektebeelden. De laatste maanden zijn nog geavanceerdere technieken beschikbaar gekomen, de zogenaame derde-generatie sequenering technnologie. Derde-generatie sequenering biedt verscheidene voordelen ten opzichte van tweede-generatie sequenering, waarbij directe methylatie sequenering en meer accurate en kwantitatieve RNA-sequenering het meest belovend lijkt. De ontwikkelde methodes kunnen snel getransformeerd worden om overweg te kunnen met de data geproduceerd door de derde-generatie sequenering instrumenten. De VIIP is reeds succesvol gebruikt om IonTorrent data, die in recente maanden beschikbaar gekomen is, te analyseren. Het is zeer waarschijnlijk dat de meer gedetailleerde methylatie en expressie data die beschikbaar zal komen dankzij de derde-generatie sequenering instrumenten de eerder bekomen resultaten enkel zal kunnen verbeteren, gebruikmakend van de eerder ontwikkelde technieken.

References

1. Schuebel KE, Chen W, Cope L, Glockner SC, Suzuki H, Yi J-M, Chan TA, Van Neste L, Van Criekinge W, van den Bosch S *et al*: **Comparing the DNA hypermethylome with gene mutations in human colorectal cancer**. *PLoS Genetics* 2007, **3**(9):1709-1723.

2. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**(5924):218-223.

3. **Genetics and the organism: Introduction**. In: *An Introduction to Genetic Analysis.* Edited by Griffiths AJ, Miller JH, Suzuki DT, Lewontin RC, Gelbart WM, 7th edn. New York, NY, USA: W.H. Freeman; 2000.

4. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. *Nature* 1953, **171**(4356):737-738.

5. Weier H-UG: **DNA fiber mapping techniques for the assembly of high-resolution physical maps**. *Journal of Histochemistry & Cytochemistry* 2001, **49**(8):939-948.

6. Crick F: **Central dogma of molecular biology**. *Nature* 1970, **227**(5258):561-563.

7. **Harvard Computational Biology PBWiki Textbook** [http://compbio.pbworks.com/]

8. Holliday R: **DNA methylation and epigenetic mechanisms**. *Cell Biophysics* 1989, **15**(1-2):15-20.

9. Bird A: **Perceptions of epigenetics**. *Nature* 2007, **447**(7143):396-398.

10. Herman JG, Baylin SB: **Gene silencing in cancer in association with promoter hypermethylation**. *New England Journal of Medicine* 2003, **349**(21):2042-2054.

11. Bestor TH: **The DNA methyltransferases of mammals**. *Human Molecular Genetics* 2000, **9**(16):2395-2402.

12. Wallace EVB, Stoddart D, Heron AJ, Mikhailova E, Maglia G, Donohoe TJ, Bayley H: **Identification of epigenetic DNA modifications with a protein nanopore**. *Chemical Communications* 2010, **46**(43):8195-8197.

13. Marks PA, Rifkind RA, Richon VM, Breslow R, Miller T, Kelly WK: **Histone deacetylases and cancer: causes and therapies**. *Nature Reviews Cancer* 2001, **1**(3):194-202.

14. Kouzarides T: **Chromatin modifications and their function**. *Cell* 2007, **128**(4):693-705.

15. Ellis L, Atadja PW, Johnstone RW: **Epigenetics in cancer: Targeting chromatin modifications**. *Molecular Cancer Therapeutics* 2009, **8**(6):1409-1420.

16. Jenuwein T, Allis CD: **Translating the Histone Code**. *Science* 2001, **293**(5532):1074-1080.

17. van Dijk K, Marley KE, Jeong B-r, Xu J, Hesson J, Cerny RL, Waterborg JH, Cerutti H: **Monomethyl histone H3 lysine 4 as an epigenetic mark for silenced euchromatin in Chlamydomonas**. *The Plant Cell Online* 2005, **17**(9):2439-2453.

18. Osoegawa K, Mammoser AG, Wu C, Frengen E, Zeng C, Catanese JJ, de Jong PJ: **A bacterial artificial chromosome library for sequencing the complete human genome**. *Genome Research* 2001, **11**(3):483-496.

19. Zhang Rg, Joachimiak A, Lawson CL, Schevitz RW, Otwinowski Z, Sigler PB: **The crystal structure of trp aporepressor at 1.8 A shows how binding tryptophan enhances DNA affinity**. *Nature* 1987, **327**(6123):591-597.

20. Robertson KD, Wolffe AP: **DNA methylation in health and disease**. *Nature Reviews Genetics* 2000, **1**(1):11-19.

21. Black DL: **Mechanisms of alternative pre-messenger RNA splicing**. *Annual Review of Biochemistry* 2003, **72**(1):291-336.

22. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nature Genetics* 2008, **40**(12):1413-1415.

23. Brennicke A, Marchfelder A, Binder S: **RNA editing**. *FEMS Microbiology Reviews* 1999, **23**(3):297-316.

24. Kable ML, Heidmann S, Stuart KD: **RNA editing: getting U into RNA**. *Trends in Biochemical Sciences* 1997, **22**(5):162-166.

25. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J: **A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine**. *Cell* 1987, **50**(6):831-840.

26. Agranat L, Raitskin O, Sperling J, Sperling R: **The editing enzyme ADAR1 and the mRNA surveillance protein hUpf1 interact in the cell nucleus**. *Proceedings of the National Academy of Sciences* 2008, **105**(13):5028-5033.

27. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization**. *Nature Reviews Genetics* 2007, **8**(6):413-423.

28. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**(5830):1484-1488.

29. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function**. *Cell* 2004, **116**(2):281-297.

30. Wang H, Iacoangeli A, Lin D, Williams K, Denman RB, Hellen CUT, Tiedge H: **Dendritic BC1 RNA in translational control mechanisms**. *The Journal of Cell Biology* 2005, **171**(5):811-821.

31. Beltran M, Puig I, Peña C, García JM, Álvarez AB, Peña R, Bonilla F, de Herreros AG: **A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition**. *Genes & Development* 2008, **22**(6):756-769.

32. Yu W, Gius D, Onyango P, Muldoon-Jacobs K, Karp J, Feinberg AP, Cui H: **Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA**. *Nature* 2008, **451**(7175):202-206.

33. Kozak M: **Regulation of translation via mRNA structure in prokaryotes and eukaryotes**. *Gene* 2005, **361**(0):13-37.

34. Bazykin GA, Kochetov AV: **Alternative translation start sites are conserved in eukaryotic genomes**. *Nucleic Acids Research* 2011, **39**(2):567-577.

35. Takahashi K, Maruyama M, Tokuzawa Y, Murakami M, Oda Y, Yoshikane N, Makabe KW, Ichisaka T, Yamanaka S: **Evolutionarily conserved non-AUG translation initiation in NAT1/p97/DAP5 (EIF4G2)**. *Genomics* 2005, **85**(3):360-371.

36. Hellen CUT, Sarnow P: **Internal ribosome entry sites in eukaryotic mRNA molecules**. *Genes & Development* 2001, **15**(13):1593-1612.

37. Steiner DF, Oyer PE: **The biosynthesis of insulin and a probable precursor of insulin by a human islet cell adenoma**. *Proceedings of the National Academy of Sciences* 1967, **57**(2):473-480.

38. Khoury GA, Baliban RC, Floudas CA: **Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database**. *Scientific Reports* 2011, **1**.

39. Menschaert G, Vandekerckhove TTM, Landuyt B, Hayakawa E, Schoofs L, Luyten W, Van Criekinge W: **Spectral clustering in peptidomics studies helps to unravel modification profile of biologically active peptides and enhances peptide identification rate**. *Proteomics* 2009, **9**(18):4381-4388.

40. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences* 1977, **74**(12):5463-5467.

41. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA**. *Nature* 1977, **265**(5596):687-695.

42. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SB, Hood LE: **Fluorescence detection in automated DNA sequence analysis**. *Nature* 1986, **321**(6071):674-679.

43. Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M: **Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis**. *Nucleic Acids Research* 1987, **15**(11):4593-4602.

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

45. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science* 2001, **291**(5507):1304-1351.

46. **DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program** [www.genome.gov/sequencingcosts]

47. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays**. *Nature Biotechnology* 2000, **18**(6):630-634.

48. Adessi Cl, Matton G, Ayala G, Turcatti G, Mermod J-J, Mayer P, Kawashima E: **Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms**. *Nucleic Acids Research* 2000, **28**(20):e87.

49. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM: **Accurate multiplex polony sequencing of an evolved bacterial genome**. *Science* 2005, **309**(5741):1728-1732.

50. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J *et al*: **The diploid genome sequence of an Asian individual**. *Nature* 2008, **456**(7218):60-65.

51. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y-J, Makhijani V, Roth GT *et al*: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**(7189):872-876.

52. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2008, **437**(7057):376-380.

53. Roche: **GS FLX Titanium series reagents: New reagents for the Genome Sequencer FLX instrument**. 2008.

54. Zheng Z, Advani A, Melefors Ãj, Glavas S, Nordström H, Ye W, Engstrand L, Andersson AF: **Titration-free massively parallel pyrosequencing using trace amounts of starting material**. *Nucleic Acids Research* 2010, **38**(13):e137.

55. Ansorge W: **Next-generation DNA sequencing techniques**. *New biotechnology* 2009, **25**(4):195-203.

56. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P: **Real-time DNA sequencing using detection of pyrophosphate release**. *Analytical Biochemistry* 1996, **242**(1):84 - 89.

57. Quinlan AR, Stewart DA, Stromberg MP, Marth GT: **Pyrobayes: an improved base caller for SNP discovery in pyrosequences**. *Nature Methods* 2008, **5**(2):179-181.

58. De Leeneer K, De Schrijver J, Clement L, Baetens M, Lefever S, De Keulenaer S, Van Criekinge W, Deforce D, Van Nieuwerburgh F, Bekaert S *et al*: **Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics**. *PLoS ONE* 2011, **6**(9):e25531.

59. Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT: **Accurate determination of microbial diversity from 454 pyrosequencing data**. *Nature Methods* 2009, **6**(9):639-641.

60. **SCOPE++ website** [http://code.google.com/p/scopeplusplus/]

61. Mardis ER: **Next-generation DNA sequencing methods**. *Annual Review of Genomics and Human Genetics* 2008, **9**(1):387-402.

62. **Illumina website: System overview** [http://www.illumina.com/systems/]

63. Janitz M: **Next-generation genome sequencing: towards personalized medicine**. Hoboken, NJ, USA: John Wiley & Sons; 2008.

64. Strausberg RL, Levy S, Rogers Y-H: **Emerging DNA sequencing technologies for human genomic medicine**. *Drug Discovery Today* 2008, **13**(13-14):569-577.

65. Metzker ML: **Sequencing technologies - the next generation**. *Nature Reviews Genetics* 2010, **11**(1):31-46.

66. McKernan KJ, Peckham HE, Costa G, McLaughlin S, Tsung E, Fu Y, Clouser C, Dunkan C, Ichikawa J, Lee C *et al*: **Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding**. *Genome Research* 2009.

67. **Applied Biosystems (ABI) website: Next-generation systems** [http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html]

68. Mardis ER: **The impact of next-generation sequencing technology on genetics**. *Trends in Genetics* 2008, **24**(3):133-141.

69. Braslavsky I, Hebert B, Kartalov E, Quake SR: **Sequence information can be obtained from single DNA molecules**. *Proceedings of the National Academy of Sciences* 2003, **100**(7):3960-3964.

70. Ozsolak F, Milos PM: **Single-molecule direct RNA sequencing without cDNA synthesis**. *Wiley Interdisciplinary Reviews: RNA* 2011, **2**(4):565-570.

71. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM: **Direct RNA sequencing**. *Nature* 2009, **461**(7265):814-818.

72. Russell DH: **The waveguide below-cutoff attenuation standard**. *Microwave Theory and Techniques, IEEE Transactions on* 1997, **45**(12):2408-2413.

73. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B *et al*: **Real-time DNA sequencing from single polymerase molecules**. *Science* 2009, **323**(5910):133-138.

74. Clark TA, Murray IA, Morgan RD, Kislyuk AO, Spittle KE, Boitano M, Fomenkov A, Roberts RJ, Korlach J: **Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing**. *Nucleic Acids Research* 2011.

75. Song C-X, Clark TA, Lu X-Y, Kislyuk A, Dai Q, Turner SW, He C, Korlach J: **Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine**. *Nature Methods* 2012, **9**(1):75-77.

76. Purushothaman S, Toumazou C, Ou C-P: **Protons and single nucleotide polymorphism detection: A simple use for the Ion Sensitive Field Effect Transistor**. *Sensors and Actuators B: Chemical* 2006, **114**(2):964-968.

77. Rusk N: **Torrents of sequence**. *Nature Methods* 2011, **8**(1):44-44.

78. Pennisi E: **Semiconductors inspire new sequencing technologies**. *Science* 2010, **327**(5970):1190.

79. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ: **Performance comparison of benchtop high-throughput sequencing platforms**. *Nature Biotechnology* 2012, **30**:434-439.

80. Howorka S, Cheley S, Bayley H: **Sequence-specific detection of individual DNA strands using engineered nanopores**. *Nature Biotechnology* 2001, **19**(7):636-639.

81. Kasianowicz JJ, Brandin E, Branton D, Deamer DW: **Characterization of individual polynucleotide molecules using a membrane channel**. *Proceedings of the National Academy of Sciences* 1996, **93**(24):13770-13773.

82. Stoddart D, Heron AJ, Klingelhoefer J, Mikhailova E, Maglia G, Bayley H: **Nucleobase recognition in ssDNA at the central constriction of the αHemolysin pore**. *Nano Letters* 2010, **10**(9):3633-3637.

83. Garaj S, Hubbard W, Reina A, Kong J, Branton D, Golovchenko JA: **Graphene as a subnanometre trans-electrode membrane**. *Nature* 2010, **467**(7312):190-193.

84. Maglia G, Heron AJ, Stoddart D, Japrung D, Bayley H: **Analysis of single nucleic acid molecules with protein nanopores**. *Methods in Enzymology* 2010, **475**:591-623.

85. Korlach J, Turner SW: **Going beyond five bases in DNA sequencing**. *Current Opinion in Structural Biology* 2012, **22**(3):251-261.

86. Astier Y, Braha O, Bayley H: **Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter**. *Journal of the American Chemical Society* 2006, **128**(5):1705-1710.

87. **Oxford Nanopore Technologies: Press release 17 Feb 2012** [http://www.nanoporetech.com/news/press-releases/view/39]

88. Berglund E, Kiialainen A, Syvanen A-C: **Next-generation sequencing technologies and applications for human genetic history and forensics**. *Investigative Genetics* 2011, **2**(1):23.

89. Hampton OA, Miller CA, Koriabine M, Li J, Den Hollander P, Carbone L, Nefedov M, Ten Hallers BFH, Lee AV, De Jong PJ *et al*: **Long-range massively parallel mate pair sequencing detects distinct mutations and similar patterns of structural mutability in two breast cancer cell lines**. *Cancer Genetics* 2011, **204**(8):447-457.

90. Akinsheye I, Alsultan A, Solovieff N, Ngo D, Baldwin CT, Sebastiani P, Chui DHK, Steinberg MH: **Fetal hemoglobin in sickle cell anemia**. *Blood* 2011, **118**(1):19-27.

91. Allison AC: **Genetic control of resistance to human malaria**. *Current Opinion in Immunology* 2009, **21**(5):499-505.

92. Watson JD, Crick FH: **The structure of DNA**. *Cold Spring Harbor Symposia on Quantitative Biology* 1953, **18**:123-131.

93. Smith D, Docter J, Ferrier P, Frias J, Spock A: **Possible localisation of the gene for Cystic Fibrosis of the pancreas to the short arm of chromosome 5**. *The Lancet* 1968, **292**(7563):309-311.

94. Shannon Danes B, Bearn AG: **Localisation of the Cystic-Fibrosis gene**. *The Lancet* 1968, **292**(7581):1303.

95. Dallaire L, Destine ML: **Localisation of the Cystic-Fibrosis gene**. *The Lancet* 1969, **293**(7591):419-420.

96. Oetting WS: **Exome and genome analysis as a tool for disease identification and treatment: The 2011 human genome variation society scientific meeting**. *Human Mutation* 2012, **33**(3):586-590.

97. Natrajan R, Reis-Filho JS: **Next-generation sequencing applied to molecular diagnostics**. *Expert Review of Molecular Diagnostics* 2011, **11**(4):425-444.

98. King M-C, Marks JH, Mandell JB, The New York Breast Cancer Study G: **Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2**. *Science* 2003, **302**(5645):643-646.

99. Compston A, Coles A: **Multiple sclerosis**. *The Lancet* 2008, **372**(9648):1502-1517.

100. Lusis AJ: **Genetics of atherosclerosis**. *Trends in Genetics* 2012(0).

101. Tsuang M, Francis T, Minor K, Thomas A, Stone W: **Genetics of smoking and depression**. *Human Genetics* 2012:1-11.

102. Ober C, Hoffjan S: **Asthma genetics 2006: the long and winding road to gene discovery**. *Genes and Immunity* 2006, **7**(2):95-100.

103. Döring Y, Noels H, Weber C: **The use of high-throughput technologies to investigate vascular inflammation and atherosclerosis**. *Arteriosclerosis, Thrombosis, and Vascular Biology* 2012, **32**(2):182-195.

104. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: **Natural selection has driven population differentiation in modern humans**. *Nature Genetics* 2008, **40**(3):340-345.

105. Sawyer SA, Hartl DL: **Population genetics of polymorphism and divergence**. *Genetics* 1992, **132**(4):1161-1176.

106. Fang S, Fang X, Xiong M: **Psoriasis prediction from genome-wide SNP profiles**. *BMC Dermatology* 2011, **11**(1):1.

107. Schwarz UI: **Clinical relevance of genetic polymorphisms in the human CYP2C9 gene**. *European Journal of Clinical Investigation* 2003, **33**:23-30.

108. Kim J, Kim JH, Park TJ, Bae JS, Lee JS, Pasaje C, Park BL, Cheong HS, Park J-S, Park S-W *et al*: **Positive association between aspirin-intolerant asthma and genetic polymorphisms of FSIP1: a case-case study**. *BMC Pulmonary Medicine* 2010, **10**(1):34.

109. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing**. *Nature* 2010, **467**(7319):1061-1073.

110. Feuk L, Carson AR, Scherer SW: **Structural variation in the human genome**. *Nature Reviews Genetics* 2006, **7**(2):85-97.

111. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome**. *Nature Genetics* 2004, **36**(9):949-951.

112. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L *et al*: **Paired-end mapping reveals**

extensive structural variation in the human genome. *Science* 2007, **318**(5849):420-426.

113. Stegemann S, Keuthe M, Greiner S, Bock R: **Horizontal transfer of chloroplast genomes between plant species**. *Proceedings of the National Academy of Sciences* 2012, **109**(7):2434-2438.

114. Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, Hobolth A, Lappalainen T, Mailund T, Marques-Bonet T *et al*: **Insights into hominid evolution from the gorilla genome sequence**. *Nature* 2012, **483**(7388):169-175.

115. Hvilsom C, Qian Y, Bataillon T, Li Y, Mailund T, Sallé B, Carlsen F, Li R, Zheng H, Jiang T *et al*: **Extensive X-linked adaptive evolution in central chimpanzees**. *Proceedings of the National Academy of Sciences* 2012, **109**(6):2054-2059.

116. George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH: **Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection**. *Genome Research* 2011, **21**(10):1686-1694.

117. Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, The Genomes P, Bustamante CD: **Demographic history and rare allele sharing among human populations**. *Proceedings of the National Academy of Sciences* 2011, **108**(29):11983-11988.

118. Zhan S, Merlin C, Boore J, Reppert S: **The Monarch butterfly genome yields insights into long-distance migration**. *Cell* 2011, **147**(5):1171-1185.

119. Ford CB, Lin PL, Chase MR, Shah RR, Iartchouk O, Galagan J, Mohaideen N, Ioerger TR, Sacchettini JC, Lipsitch M *et al*: **Use of whole genome sequencing to estimate the mutation rate of Mycobacterium tuberculosis during latent infection**. *Nature Genetics* 2011, **43**(5):482-486.

120. Wang D, Li F, Freed DC, Finnefrock AC, Tang A, Grimes SN, Casimiro DR, Fu T-M: **Quantitative analysis of neutralizing antibody response to human cytomegalovirus in natural infection**. *Vaccine* 2011, **29**(48):9075-9080.

121. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen J-E, Weigel D, Andersen SU: **SHOREmap: simultaneous**

mapping and mutation identification by deep sequencing. *Nature Methods* 2009, **6**(8):550-551.

122. Chamberlain JS, Gibbs RA, Ranier JE, Nguyen PN, Caskey CT: **Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification**. *Nucleic Acids Res* 1988, **16**(3205741):11141-11156.

123. Kulesh DA, Clive DR, Zarlenga DS, Greene JJ: **Identification of interferon-modulated proliferation-related cDNA sequences**. *Proceedings of the National Academy of Sciences* 1987, **84**(23):8453-8457.

124. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.

125. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial Analysis of Gene Expression**. *Science* 1995, **270**(5235):484-487.

126. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE: **Using the transcriptome to annotate the genome**. *Nature Biotechnology* 2002, **20**(5):508-512.

127. Matsumura H, Reich S, Ito A, Saitoh H, Kamoun S, Winter P, Kahl G, Reuter M, Krüger DH, Terauchi R: **Gene expression analysis of plant host-pathogen interactions by SuperSAGE**. *Proceedings of the National Academy of Sciences* 2003, **100**(26):15718-15723.

128. Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Krüger DH, Terauchi R: **SuperSAGE**. *Cellular Microbiology* 2005, **7**(1):11-18.

129. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T *et al*: **Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage**. *Proceedings of the National Academy of Sciences* 2003, **100**(26):15776-15781.

130. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al*: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991, **252**(5013):1651-1656.

131. Wickramasinghe S, Rincon G, Islas-Trejo A, Medrano J: **Transcriptional profiling of bovine milk using RNA sequencing**. *BMC Genomics* 2012, **13**(1):45.

132. Jabbari A, Suarez-Farinas M, Dewell S, Krueger JG: **Transcriptional profiling of psoriasis using RNA-seq reveals previously unidentified differentially expressed genes**. *Journal of Investigative Dermatology* 2012, **132**(1):246-249.

133. Beane J, Vick J, Schembri F, Anderlind C, Gower A, Campbell J, Luo L, Zhang XH, Xiao J, Alekseyev YO *et al*: **Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-seq**. *Cancer Prevention Research* 2011, **4**(6):803-817.

134. Fu Y, Sun Y, Li Y, Li J, Rao X, Chen C, Xu A: **Differential genome-wide profiling of tandem 3' UTRs among human breast cancer and normal cells by high-throughput sequencing**. *Genome Research* 2011, **21**(5):741-747.

135. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**(7234):97-101.

136. Lee C-H, Ou W-B, Mariño-Enriquez A, Zhu M, Mayeda M, Wang Y, Guo X, Brunner AL, Amant Fdr, French CA *et al*: **14-3-3 fusion oncogenes in high-grade endometrial stromal sarcoma**. *Proceedings of the National Academy of Sciences* 2012, **109**(3):929-934.

137. Ju YS, Lee W-C, Shin J-Y, Lee S, Bleazard T, Won J-K, Kim YT, Kim J-I, Kang J-H, Seo J-S: **A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing**. *Genome Research* 2012, **22**(3):436-445.

138. Al-Balool HH, Weber D, Liu Y, Wade M, Guleria K, Nam PLP, Clayton J, Rowe W, Coxhead J, Irving J *et al*: **Post-transcriptional exon shuffling events in humans can be evolutionarily conserved and abundant**. *Genome Research* 2012, **21**(11):1788-1799.

139. Jia Y, Mu J, Ackerman S: **Mutation of a U2 snRNA gene causes global disruption of alternative splicing and neurodegeneration**. *Cell* 2012, **148**(1-2):296-308.

140. Gabut M, Samavarchi-Tehrani P, Wang X, Slobodeniuc V, O'Hanlon D, Sung H-K, Alvarez M, Talukder S, Pan Q, Mazzoni Esteban O *et al*: **An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming**. *Cell* 2011, **147**(1):132-146.

141. Rosel TD, Hung L-H, Medenbach J, Donde K, Starke S, Benes V, Ratsch G, Bindereif A: **RNA-Seq analysis in mutant zebrafish reveals role of U1C protein in alternative splicing regulation**. *The EMBO Journal* 2011, **30**(10):1965-1976.

142. Daneck P, Nellaker C, McIntyre R, Buendia-Buendia J, Bumpstead S, Ponting C, Flint J, Durbin R, Keane T, Adams D: **High levels of RNA-editing site conservation amongst 15 laboratory mouse strains**. *Genome Biology* 2012, **13**(4):R26.

143. Peng Z, Cheng Y, Tan BC-M, Kang L, Tian Z, Zhu Y, Zhang W, Liang Y, Hu X, Tan X *et al*: **Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome**. *Nature Biotechnology* 2012, **30**(3):253-260.

144. Maruyama R, Shipitsin M, Choudhury S, Wu Z, Protopopov A, Yao J, Lo P-K, Bessarabova M, Ishkin A, Nikolsky Y *et al*: **Altered antisense-to-sense transcript ratios in breast cancer**. *Proceedings of the National Academy of Sciences* 2012, **109**(8):2820-2824.

145. Vivancos AP, Güell M, Dohm JC, Serrano L, Himmelbauer H: **Strand-specific deep sequencing of the transcriptome**. *Genome Research* 2010, **20**(7):989-999.

146. Schotte D, Moqadam FA, Lange-Turenhout EAM, Chen C, van Ijcken WFJ, Pieters R, den Boer ML: **Discovery of new microRNAs by small RNAome deep sequencing in childhood acute lymphoblastic leukemia**. *Leukemia* 2012, **25**(9):1389-1399.

147. Li B, Qin Y, Duan H, Yin W, Xia X: **Genome-wide characterization of new and drought stress responsive microRNAs in Populus euphratica**. *Journal of Experimental Botany* 2011, **62**(11):3765-3779.

148. Jin G, Sun J, Isaacs SD, Wiley KE, Kim S-T, Chu LW, Zhang Z, Zhao H, Zheng SL, Isaacs WB *et al*: **Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk**. *Carcinogenesis* 2011, **32**(11):1655-1659.

149. Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP: **The impact of microRNAs on protein output**. *Nature* 2008, **455**(7209):64-71.

150. Ingolia NT, Lareau LF, Weissman Jonathan S: **Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes**. *Cell* 2011, **147**(4):789-802.

151. Gold M, Hurwitz J, Anders M: **The enzymatic methylation of RNA and DNA**. *Biochemical and Biophysical Research Communications* 1963, **11**:107-114.

152. Münzel M, Globisch D, Carell T: **5-hydroxymethylcytosine, the sixth base of the genome**. *Angewandte Chemie International Edition* 2011, **50**(29):6460-6468.

153. Haines TR, Rodenhiser DI, Ainsworth PJ: **Allele-specific non-CpG methylation of the Nf1 gene during early mouse development**. *Developmental Biology* 2001, **240**(2):585-598.

154. Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, Gehrke C: **Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells**. *Nucleic Acids Research* 1982, **10**(8):2709-2721.

155. Bird AP: **CpG-rich islands and the function of DNA methylation**. *Nature* 1986, **321**(6067):209-213.

156. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J *et al*: **Dynamic changes in the human methylome during differentiation**. *Genome Research* 2010.

157. Langevin SM, Houseman EA, Christensen BC, Wiencke JK, Nelson HH, Karagas MR, Marsit CJ, Kelsey KT: **The influence of aging, environmental exposures and local sequence features on the variation of DNA methylation in blood**. *Epigenetics* 2011, **6**(7):908-919.

158. Miller CA, Sweatt JD: **Covalent modification of DNA regulates memory formation**. *Neuron* 2007, **53**(6):857-869.

159. Migeon BR: **Concerning the role of X-inactivation and DNA methylation in fragile X syndrome**. *American Journal of Medical Genetics* 1992, **43**(1-2):291-298.

160. Dobrovic A, Gareau JLP, Ouellette G, Bradley WEC: **DNA methylation and genetic inactivation at thymidine kinase locus: Two different mechanisms for silencing autosomal genes**. *Somatic Cell and Molecular Genetics* 1988, **14**(1):55-68.

161. Goodman JI, Counts JL: **Hypomethylation of DNA: a possible nongenotoxic mechanism underlying the role of cell proliferation in carcinogenesis**. *Environmental Health Perspectives* 1993, **101 Suppl 5**:169-172.

162. Herman JG, Baylin SB: **Promoter-region hypermethylation and gene silencing in human cancer**. *Current Topics in Microbiology and Immunology* 2000, **249**:35-54.

163. Woodson K, O'Reilly KJ, Hanson JC, Nelson D, Walk EL, Tangrea JA: **The usefulness of the detection of GSTP1 methylation in urine as a biomarker in the diagnosis of prostate cancer**. *The Journal of Urology* 2008, **179**(2):508-512.

164. Henrique R, Ribeiro FR, Fonseca D, Hoque MO, Carvalho AL, Costa VL, Pinto M, Oliveira J, Teixeira MR, Sidransky D *et al*: **High promoter methylation levels of APC predict poor prognosis in sextant biopsies from prostate cancer patients**. *Clinical Cancer Research* 2007, **13**(20):6122-6129.

165. Hegi ME, Diserens A-C, Gorlia T, Hamou M-F, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L *et al*: **MGMT gene silencing and benefit from temozolomide in glioblastoma**. *New England Journal of Medicine* 2005, **352**(10):997-1003.

166. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL: **A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands**. *Proceedings of the National Academy of Sciences* 1992, **89**(5):1827-1831.

167. Lee E-J, Pei L, Srivastava G, Joshi T, Kushwaha G, Choi J-H, Robertson KD, Wang X, Colbourne JK, Zhang L *et al*: **Targeted bisulfite sequencing by solution hybrid selection and massively parallel sequencing**. *Nucleic Acids Research* 2011.

168. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE: **Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning**. *Nature* 2008, **452**(7184):215-219.

169. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D: **Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells**. *Nature Genetics* 2005, **37**(8):853-862.

170. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM *et al*: **A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis**. *Nature Biotechnology* 2008, **26**(7):779-785.

171. Serre D, Lee BH, Ting AH: **MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome**. *Nucleic Acids Research* 2011, **38**(2):391-399.

172. Hendrich B, Bird A: **Identification and characterization of a family of mammalian methyl-CpG binding proteins**. *Molecular and Cellular Biology* 1998, **18**(11):6538-6547.

173. Yegnasubramanian S, Lin X, Haffner MC, DeMarzo AM, Nelson WG: **Combination of methylated-DNA precipitation and methylation-sensitive restriction enzymes (COMPARE-MS) for the rapid, sensitive and quantitative detection of DNA methylation**. *Nucleic Acids Research* 2006, **34**(3):e19.

174. Elo LL, Kallio A, Laajala TD, David Hawkins R, Korpelainen E, Aittokallio T: **Optimized detection of transcription factor-binding sites in ChIP-seq experiments**. *Nucleic Acids Research* 2011.

175. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW: **Direct detection of DNA methylation during single-molecule, real-time sequencing**. *Nature Methods* 2010, **7**(6):461-465.

176. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Research* 2008, **18**(5):821-829.

177. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S: **Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs**. *Genome Research* 2004, **14**(6):1147-1159.

178. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly**. *Proceedings of the National Academy of Sciences* 2001, **98**(17):9748-9753.

179. Chaisson MJ, Brinza D, Pevzner PA: **De novo fragment assembly with short mate-paired reads: Does the read length matter?** *Genome Research* 2009, **19**(2):336-346.

180. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nature Methods* 2011, **8**(6):469-477.

181. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N *et al*: **Reference-guided assembly of four diverse Arabidopsis thaliana genomes**. *Proceedings of the National Academy of Sciences* 2011, **108**(25):10249-10254.

182. Pop M, Phillippy A, Delcher AL, Salzberg SL: **Comparative genome assembly**. *Briefings in Bioinformatics* 2004, **5**(3):237-248.

183. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *Journal of Molecular Biology* 1970, **48**(3):443-453.

184. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *Journal of Molecular Biology* 1981, **147**(1):195-197.

185. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403 - 410.

186. Altschul SF, Madden TL, SchÃ¤ffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389-3402.

187. Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Research* 2001, **29**(14):2994-3005.

188. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Research* 2002, **12**(4):656-664.

189. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**(11):1851-1858.

190. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program**. *Bioinformatics* 2008, **24**(5):713-714.

191. Vyverman Ml, De Baets B, Fack V, Dawyndt P: **Prospects and limitations of full-text index structures in genome analysis**. *Nucleic Acids Research* 2012.

192. Langmead B, Trapnell C, Pop M, Salzberg S: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biology* 2009, **10**(3).

193. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.

194. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**(15):1966-1967.

195. Prüfer K, Stenzel U, Dannemann M, Green RE, Lachmann M, Kelso J: **PatMaN: rapid alignment of short sequences to large databases**. *Bioinformatics* 2008, **24**(13):1530-1531.

196. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nature Methods* 2012, **9**(4):357-359.

197. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

198. Xi R, Kim T-M, Park PJ: **Detecting structural variations in the human genome using next generation sequencing**. *Briefings in Functional Genomics* 2010, **9**(5-6):405-415.

199. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O *et al*: **Personalized copy number and segmental duplication maps using next-generation sequencing**. *Nature Genetics* 2009, **41**(10):1061-1067.

200. Xie C, Tammi M: **CNV-seq, a new method to detect copy number variation using high-throughput sequencing**. *BMC Bioinformatics* 2009, **10**(1):80.

201. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP *et al*: **BreakDancer: an algorithm for high-resolution mapping of genomic structural variation**. *Nature Methods* 2009, **6**(9):677-681.

202. Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, Barillot E: **SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data**. *Bioinformatics* 2010, **26**(15):1895-1896.

203. Krzywinski M, Schein J, Birol Än, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics**. *Genome Research* 2009, **19**(9):1639-1645.

204. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D *et al*: **The mutation spectrum revealed by paired genome sequences from a lung cancer patient**. *Nature* 2010, **465**(7297):473-477.

205. Chatterjee A, Stockwell PA, Rodger EJ, Morison IM: **Comparison of alignment software for genome-wide bisulphite sequence data**. *Nucleic Acids Research* 2012.

206. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR: **Highly integrated single-base resolution maps of the epigenome in Arabidopsis**. *Cell* 2008, **133**(3):523-536.

207. Harris EY, Ponts N, Levchuk A, Roch KL, Lonardi S: **BRAT: bisulfite-treated reads analysis tool**. *Bioinformatics* 2010, **26**(19):2499.

208. Krueger F, Andrews SR: **Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications**. *Bioinformatics* 2011, **27**(11):1571-1572.

209. Xi Y, Li W: **BSMAP: whole genome bisulfite sequence MAPping program**. *BMC Bioinformatics* 2009, **10**(1):232.

210. Smith AD, Chung W-Y, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ: **Updates to the RMAP short-read mapping software**. *Bioinformatics* 2009, **25**(21):2841-2842.

211. Hardcastle T, Kelly K: **baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics* 2010, **11**(1):422.

212. Feng X, Grossman R, Stein L: **PeakRanger: A cloud-enabled peak caller for ChIP-seq data**. *BMC Bioinformatics* 2011, **12**(1):139.

213. Li H, Lovci MT, Kwon Y-S, Rosenfeld MG, Fu X-D, Yeo GW: **Determination of tag density required for digital transcriptome analysis: Application to an androgen-sensitive prostate cancer model**. *Proceedings of the National Academy of Sciences* 2008, **105**(51):20179-20184.

214. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nature Methods* 2008, **5**(7):621-628.

215. Auer PL, Srivastava S, Doerge RW: **Differential expression - the next generation and beyond**. *Briefings in Functional Genomics* 2011.

216. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.

217. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nature Biotechnology* 2010, **28**(5):511-515.

218. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nature Reviews Genetics* 2011, **12**(10):671-682.

219. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE *et al*: **De novo transcriptome assembly with ABySS**. *Bioinformatics* 2009, **25**(21):2872-2877.

220. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length**

transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* 2011, **29**(7):644-652.

221. Flicek P: **The need for speed**. *Genome Biology* 2009, **10**(3):212.

222. Pop M, Salzberg SL: **Bioinformatics challenges of new sequencing technology**. *Trends in Genetics* 2008, **24**(3):142-149.

223. Voelkerding KV, Dames SA, Durtschi JD: **Next-generation sequencing: from basic research to diagnostics**. *Clinical Chemistry* 2009, **55**(4):641-658.

224. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.

225. **Mosaik Website** [http://bioinformatics.bc.edu/marthlab/Mosaik]

226. Huse S, Huber J, Morrison H, Sogin M, Welch D: **Accuracy and quality of massively parallel DNA pyrosequencing**. *Genome Biology* 2007, **8**(7):R143.

227. De Leeneer K, Hellemans J, De Schrijver J, Baetens M, Poppe B, Van Criekinge W, De Paepe A, Coucke P, Claes K: **Massive parallel amplicon sequencing of the breast cancer genes BRCA1 and BRCA2: opportunities, challenges, and limitations**. *Human Mutation* 2011, **32**(3):335-344.

228. **HeidiSQL Website** [http://www.heidisql.com/]

229. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nature Biotechnology* 2008, **26**(10):1135-1145.

230. Adzhubei I, Schmidt S, Peshkin L, Ramensky V, Gerasimova A, Bork P, Kondrashov A, Sunyaev S: **A method and server for predicting damaging missense mutations**. *Nature Methods* 2010, **7**(4):248-249.

231. Ng P, Henikoff S: **Predicting the effects of amino acid substitutions on protein function**. *Annual Review of Genomics and Human Genetics* 2006, **7**(1):61-80.

232. Reese MG, Eeckman FH, Kulp D, Haussler D: **Improved splice site detection in Genie**. *Journal of Computational Biology* 1997, **4**(3):311-323.

233. Liu C-K, Chen Y-H, Tang C-Y, Chang S-C, Lin Y-J, Tsai M-F, Chen Y-T, Yao A: **Functional analysis of novel SNPs and mutations in human and mouse genomes**. *BMC Bioinformatics* 2008, **9 Suppl 12**.

234. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D: **MutationTaster evaluates disease-causing potential of sequence alterations**. *Nature Methods* 2010, **7**(8):575-576.

235. De Schrijver J, De Leeneer K, Lefever S, Sabbe N, Pattyn F, Van Nieuwerburgh F, Coucke P, Deforce D, Vandesompele J, Bekaert S *et al*: **Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline**. *BMC Bioinformatics* 2010, **11**(1):269.

236. **Perl API Documentation** [http://www.ensembl.org/info/docs/api/index.html]

237. Smigielski EM, Sirotkin K, Ward M, Sherry ST: **dbSNP: a database of single nucleotide polymorphisms**. *Nucleic Acids Research* 2000, **28**(1):352-355.

238. Martin F: **Patterns of enterprise application architecture**: Addison-Wesley Longman Publishing Co., Inc.; 2002.

239. **H2G2 Genome Browser Website** [http://h2g2.ugent.be/biobix.html]

240. Baetens M, Van Laer L, De Leeneer K, Hellemans J, De Schrijver J, Van De Voorde H, Renard M, Dietz H, Lacro RV, Menten B *et al*: **Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes**. *Human Mutation* 2011, **32**(9):1053-1062.

241. De Keulenaer S, Hellemans J, Lefever S, Renard J-P, De Schrijver J, Van de Voorde H, Tabatabaiefar MA, Van Nieuwerburgh F, Flamez D, Pattyn F *et al*: **Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform**. *BMC Medical Genomics* 2012, **5**(1):17.

242. Loeys B, De Backer J, Van Acker P, Wettinck K, Pals G, Nuytinck L, Coucke P, De Paepe A: **Comprehensive molecular screening of the FBN1**

gene favors locus homogeneity of classical Marfan syndrome. *Human Mutation* 2004, **24**(2):140-146.

243. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G: **Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification**. *Nucleic Acids Research* 2002, **30**(12):e57.

244. Pareek C, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing**. *Journal of Applied Genetics* 2011, **52**(4):413-435.

245. Herman JG, Graff JR, Myöhänen S, Nelkin BD, Baylin SB: **Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands**. *Proceedings of the National Academy of Sciences* 1996, **93**(18):9821-9826.

246. Weisenberger D, Van Den Berg D, Pan F, Berman B, Laird P: **Comprehensive DNA methylation analysis on the illumina infinium assay platform**. *Illumina application note* 2008.

247. Pareek C, Smoczynski R, Tretyn A: **Sequencing technologies and genome sequencing**. *Journal of Applied Genetics* 2011:1-23.

248. Towbin H, Staehelin T, Gordon J: **Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications**. *Proceedings of the National Academy of Sciences* 1979, **76**(9):4350-4354.

249. Maskos U, Southern EM: **Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ**. *Nucleic Acids Research* 1992, **20**(7):1679-1684.

250. Su Z, Li Z, Chen T, Li Q-Z, Fang H, Ding D, Ge W, Ning B, Hong H, Perkins RG *et al*: **Comparing Next-Generation Sequencing and Microarray Technologies in a Toxicological Study of the Effects of Aristolochic Acid on Rat Kidneys**. *Chemical Research in Toxicology* 2011, **24**(9):1486-1493.

251. Rhee I, Bachman KE, Park BH, Jair K-W, Yen R-WC, Schuebel KE, Cui H, Feinberg AP, Lengauer C, Kinzler KW *et al*: **DNMT1 and DNMT3b cooperate to silence genes in human cancer cells**. *Nature* 2002, **416**(6880):552-556.

252. Suzuki H, Gabrielson E, Chen W, Anbazhagan R, van Engeland M, Weijenberg MP, Herman JG, Baylin SB: **A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer**. *Nature Genetics* 2002, **31**(2):141-149.

253. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S *et al*: **Ensembl 2011**. *Nucleic Acids Research* 2011, **39**(suppl 1):D800-D806.

254. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists**. *BMC Bioinformatics* 2009, **10**(1):48.

255. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E *et al*: **PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes**. *Nature Genetics* 2003, **34**(3):267-273.

256. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.

257. Vanhée-Brossollet C, Vaquero C: **Do natural antisense transcripts make sense in eukaryotes?** *Gene* 1998, **211**(1):1-9.

258. Carmichael GG: **Antisense starts making more sense**. *Nature Biotechnology* 2003, **21**(4):371-372.

259. Lapidot M, Pilpel Y: **Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms**. *EMBO Reports* 2006, **7**(12):1216-1222.

260. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing**. *Nature* 2011, **479**(7371):74-79.

261. Hanahan D, Weinberg R: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.

262. Xiang T, Li L, Yin X, Yuan C, Tan C, Su X, Xiong L, Putti TC, Oberst M, Kelly K *et al*: **The ubiquitin peptidase UCHL1 induces G0/G1 cell cycle arrest and apoptosis through stabilizing p53 and is frequently silenced in breast cancer**. *PLoS ONE* 2012, **7**(1):e29783.

263. Rovillain E, Mansfield L, Lord C, Ashworth A, Jat P: **An RNA interference screen for identifying downstream effectors of the p53 and pRB tumour suppressor pathways involved in senescence**. *BMC Genomics* 2011, **12**(1):355.

264. Chen W-D, Han ZJ, Skoletsky J, Olson J, Sah J, Myeroff L, Platzer P, Lu S, Dawson D, Willis J *et al*: **Detection in fecal DNA of colon cancer-specific methylation of the nonexpressed vimentin gene**. *Journal of the National Cancer Institute* 2005, **97**(15):1124-1132.

265. Renaud Sp, Pugacheva EM, Delgado MD, Braunschweig R, Abdullaev Z, Loukinov D, Benhattar J, Lobanenkov V: **Expression of the CTCF-paralogous cancer-testis gene, brother of the regulator of imprinted sites (BORIS), is regulated by three alternative promoters modulated by CpG methylation and by CTCF and p53 transcription factors**. *Nucleic Acids Research* 2007, **35**(21):7372-7388.

266. De Backer O, Arden KC, Boretti M, Vantomme Vr, De Smet C, Czekay S, Viars CS, De Plaen E, Brasseur F, Chomez P *et al*: **Characterization of the GAGE genes that are expressed in various human cancers and in normal testis**. *Cancer Research* 1999, **59**(13):3157-3165.

267. Lind GE, Raiborg C, Danielsen SA, Rognum TO, Thiis-Evensen E, Hoff G, Nesbakken A, Stenmark H, Lothe RA: **SPG20, a novel biomarker for early detection of colorectal cancer, encodes a regulator of cytokinesis**. *Oncogene* 2011, **30**(37):3967-3978.

268. Shao C, Sun W, Tan M, Glazer CA, Bhan S, Zhong X, Fakhry C, Sharma R, Westra WH, Hoque MO *et al*: **Integrated, genome-wide screening for hypomethylated oncogenes in salivary gland adenoid cystic carcinoma**. *Clinical Cancer Research* 2011, **17**(13):4320-4330.

269. Chen LH, Kuo W-H, Tsai M-H, Chen P-C, Hsiao CK, Chuang EY, Chang L-Y, Hsieh F-J, Lai L-C, Chang K-J: **Identification of prognostic genes for recurrent risk prediction in triple negative breast cancer patients in Taiwan**. *PLoS ONE* 2011, **6**(11):e28222.

270. Pils D, Horak P, Gleiss A, Sax C, Fabjani G, Moebus VJ, Zielinski C, Reinthaller A, Zeillinger R, Krainer M: **Five genes from chromosomal**

band 8p22 are significantly down-regulated in ovarian carcinoma. *Cancer* 2005, **104**(11):2417-2429.

271. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, Garnica A, Cheung SW, Beaudet AL: **Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster**. *Nature Genetics* 2008, **40**(6):719-721.

272. Bazeley PS, Shepelev V, Talebizadeh Z, Butler MG, Fedorova L, Filatov V, Fedorov A: **snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions**. *Gene* 2008, **408**(1-2):172-179.

273. Anastasiadou C, Malousi A, Maglaveras N, Kouidou S: **Human epigenome data reveal increased CpG methylation in alternatively spliced sites and putative exonic splicing enhancers**. *DNA and Cell Biology* 2011, **30**(5):267-275.

274. Elsarraj H, Stecklein S, Valdez K, Behbod F: **Emerging functions of microRNA-146a/b in development and breast cancer**. *Journal of Mammary Gland Biology and Neoplasia* 2012, **17**(1):79-87.

275. Li Y, VandenBoom TG, Wang Z, Kong D, Ali S, Philip PA, Sarkar FH: **miR-146a suppresses invasion of pancreatic cancer cells**. *Cancer Research* 2010, **70**(4):1486-1495.

276. Karakatsanis A, Papaconstantinou I, Gazouli M, Lyberopoulou A, Polymeneas G, Voros D: **Expression of microRNAs, miR-21, miR-31, miR-122, miR-145, miR-146a, miR-200c, miR-221, miR-222, and miR-223 in patients with hepatocellular carcinoma or intrahepatic cholangiocarcinoma and its prognostic significance**. *Molecular Carcinogenesis* 2011:n/a-n/a.

277. Paik JH, Jang J-Y, Jeon YK, Kim WY, Kim TM, Heo DS, Kim C-W: **MicroRNA-146a downregulates NFκB activity via targeting TRAF6 and functions as a tumor suppressor having strong prognostic implications in NK/T cell lymphoma**. *Clinical Cancer Research* 2011, **17**(14):4761-4771.

278. Fang MZ, Chen D, Sun Y, Jin Z, Christman JK, Yang CS: **Reversal of hypermethylation and reactivation of p16INK4a, RARbeta, and MGMT genes by genistein and other isoflavones from soy**. *Clinical Cancer Research* 2005, **11**(19):7033-7041.

279. Qin W, Zhu W, Shi H, Hewett JE, Ruhlen RL, MacDonald RS, Rottinghaus GE, Chen Y-C, Sauter ER: **Soy isoflavones have an antiestrogenic effect and alter mammary promoter hypermethylation in healthy premenopausal women**. *Nutrition and Cancer* 2009, **61**(2):238-244.

280. Kim Y-H, Lee HC, Kim S-Y, Yeom YI, Ryu KJ, Min B-H, Kim D-H, Son HJ, Rhee P-L, Kim JJ *et al*: **Epigenomic analysis of aberrantly methylated genes in colorectal cancer identifies genes commonly affected by epigenetic alterations**. *Annals of Surgical Oncology* 2011, **18**(8):2338-2347.

281. Ahlquist T, Lind G, Costa V, Meling G, Vatn M, Hoff G, Rognum T, Skotheim R, Thiis-Evensen E, Lothe R: **Gene methylation profiles of normal mucosa, and benign and malignant colorectal tumors identify early onset markers**. *Molecular Cancer* 2008, **7**(1):94.

282. Jacinto FV, Ballestar E, Ropero S, Esteller M: **Discovery of epigenetically silenced genes by methylated DNA immunoprecipitation in colon cancer cells**. *Cancer Research* 2007, **67**(24):11481-11486.

283. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S: **CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing**. *Nature*, **479**(7371):74-79.

284. Song J, Shi L, Li D, Sun Y, Niu Y, Chen Z, Luo H, Pang X, Sun Z, Liu C *et al*: **Extensive Pyrosequencing Reveals Frequent Intra-Genomic Variations of Internal Transcribed Spacer Regions of Nuclear Ribosomal DNA**. *PLoS ONE* 2012, **7**(8):e43971.

285. van der Walt EM, Smuts I, Taylor RW, Elson JL, Turnbull DM, Louw R, van der Westhuizen FH: **Characterization of mtDNA variation in a cohort of South African paediatric patients with mitochondrial disease**. *Eur J Hum Genet* 2012, **20**(22258525):650-656.

286. De Beuf K, De Schrijver J, Thas O, Van Criekinge W, Irizarry R, Clement L: **Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model**. *BMC Bioinformatics* 2012, **13**(1):303.

287. **Galaxy Website** [http://main.g2.bx.psu.edu/]

288. Taylor J, Schenck I, Blankenberg D, Nekrutenko A: **Using galaxy to perform large-scale interactive data analyses**. *Current Protocols in Bioinformatics* 2007, **Chapter 10**.

289. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL: **Searching for SNPs with cloud computing**. *Genome Biology* 2009, **10**(11).

Nawoord

Het schrijven van een doctoraatsthesis is het sluitstuk van een lange emotionele rollercoaster. Grote vreugde wanneer de resultaten veelbelovend zijn, een artikel aanvaard wordt, er deelgenomen kan worden aan een congres op een exotische locatie. Deze vreugdemomenten worden echter afgewisseld met frustratie wanneer resultaten niet zijn wat ze zouden moeten zijn en ontgoocheling wanneer artikels afgekeurd worden. Gelukkig kan de uiteindelijke afloop van de rollercoaster positief beschouwd worden. Ik had in dit stukje dan ook de mensen willen bedanken die dit alles mogelijk gemaakt hebben.

Ten eerste wil ik mijn promotor, **Wim Van Criekinge,** bedanken die dit alles mogelijk gemaakt heeft. Hij was het die mij tijdens het behalen van mijn diploma bio-ingenieur de wereld van de bio-informatica leerde ontdekken. Al tijdens mijn master thesis gaf hij me alle vertrouwen om naar hartenlust onderzoek te doen en stuurde hij enkel bij waar nodig. Een doctoraat was het logische gevolg van deze vlotte samenwerking. Ook hier liet hij me vrij mijn ding doen, maar op de cruciale punten stond hij er om bij te sturen en nieuwe ideeën aan te geven.

Ook wil ik mijn co-promotor **Sofie Bekaert** bedanken. In het bijzonder omdat zij één van de drijvende krachten was bij het opzetten van de **NXTGNT** sequencing faciliteit. Zonder deze sequencing faciliteit was mijn doctoraat nooit mogelijk geweest.

Toen ik mijn doctoraat begon, zaten we nog met een beperkte groep op het gelijkvloers van blok B op de Coupure. Sindsdien zijn we gegroeid, gemigreerd naar blok A en van vakgroep veranderd. Ik wil alle collega's van de vroegere vakgroep BW14 bedanken en de nieuwe collega's van BW10. In het bijzonder de collega's waar ik over de jaren nauw mee samengewerkt heb: **Leander**, **Maté**, **Tom**, **Evi**, **Katrien**, **Geert**, **Tim**, **Klaas**, **Gerben**, **Simon**, **Jeroen**, **Alexander**, **Selin** en **Sandra**. Verder wil ik ook iedereen van de administratie bedanken voor de vlotte afhandeling van alle administratieve formaliteiten. Tot voor de migratie naar BW10 was dat vooral het werk van **Sofie De Schynkel** en **Fien De Block**, momenteel wordt dat werk vlot geregeld door **Timpe Vogelaers**. **Daisy Flamez** wil ik bedanken voor de vlotte afhandeling van alles wat betreft patent filing, intellectueel eigendom en valorisatie, wat niet altijd evident is binnen Biobix.

Verder wil ik nog **Martijn Devisscher** en de Directie Informatie- en Communicatietechnologie (**DICT**) van de Universiteit Gent bedanken voor de hulp

bij het draaiende houden en terug draaiende krijgen van de server infrastructuur na een zoveelste 'upgrade' die nadien niet zo'n vooruitgang bleek te zijn als voorheen gedacht.

Onderzoek doen is multidisciplinair werken, dus moet er over de grenzen (van vakgroepen, faculteiten en universiteiten) heen samengewerkt worden. Ik wil dan ook enkele mensen bedanken waarmee ik samengewerkt heb de laatste jaren.

Ik wil de vele mensen van het **Centrum Medische Genetica Gent** (CMG) bedanken voor de intensieve samenwerking in het begin van mijn doctoraat. Ik wil specifiek **Jan Hellemans, Jo Vandesompele, Kim de Leeneer** en **Bram De Wilde** bedanken voor de verschillende sequencing experimenten die we samen geanalyseerd hebben. **Jean-Pierre Renard, Sarah De Keulenaer** en **Hendrik Van de Voorde** wil ik bedanken voor de vlotte samenwerking op het meer technische gebied van de sequencing experiment.

Verder wil ik de mensen van het **VUMC Amsterdam** bedanken voor de vlotte samenwerking van de RNA-sequencing experimenten. Bedankt aan **Linda Bosch, Beatriz Carvalho** en **Gerrit Meijer**.

Ook heb ik over de jaren heen aan enkele leuke zijprojecten kunnen meewerken. Ik wil **Danny Geelen** en **Nico Storme** van de vakgroep Plantaardige Productie bedanken voor het interessante Arabidopsis sequencing project. **Dagmar Obbels, Elie Verleyen** en **Wim Vyverman** wil ik vermelden voor het zeer interessante project 'sequencen van bacteriële species van Antarctica'. In dit project waren ook, toen nog studenten, **Pieter-Jan Volders** en **Frederiek-Maarten Kerckhof** nauw betrokken. Beide zijn ondertussen bezig met hun eigen doctoraatsonderzoek.

Als laatste wil ik **Kristof De Beuf, Lieven Clement** en **Olivier Thas** bedanken voor de productieve samenwerking die we gehad hebben bij het ontwikkelen van de verschillende tools ter verbetering van de pyrosequencing basecalling algoritmes.

Al deze professionele contacten en vlotte samenwerkingen zouden uiteraard niet mogelijk geweest zijn zonder steun op persoonlijk gebied. Ik wil vrienden, familie en ouders bedanken die mij zijn blijven steunen over de jaren heen. Ook wil ik hen bedanken voor het vrijwillig ondergaan van de gedwongen lessen biologie en genetica tijdens de pogingen om uit te leggen waar ik nu juist mee bezig was.

Curriculum vitae

General information

Name	Joachim De Schrijver
Address	Mariakerkelaan 338 8400 Oostende Belgium
Mobile	+32-(0)497-717724
E-mail	jds@joachimdeschrijver.be
Date of birth	01 July 1985
Place of birth	Brugge, Belgium

Working experience

2012-2013	**Ghent University – Lab for Computational Genomics and Bioinformatics** *Doctoral Researcher*
2009 - 2012	**NXTGNT sequencing facility** *Responsible for bioinformatics and general IT support*
2008 - 2012	**Ghent University – Lab for Computational Genomics and Bioinformatics** *Doctoral Researcher - Special Research Fund (BOF) scholarship*
2007	**BASF Plant Science – CropDesign** *Summer internship at the Data-Analysis and Bioinformatics department*
2003 – 2007	**DAIKIN Europe NV** *Student job in different departments*

Education

2012 - 2013 Ghent University: MSc in Economics – Financial Institutions and Markets

2010 - 2012 Ghent University: Preparatory courses for Msc in Economics
Graduated *magna cum laude*

2005 - 2008 Ghent University: MSc in Bioscience Engineering – Biotechnology
Graduated *magna cum laude*

2003 - 2005 Ghent University: BSc in Bioscience Engineering
Graduated *magna cum laude*

1997 - 2003 Onze Lieve Vrouwecollege Oostende: Mathematics – Science (8h)

Additional courses

2011 Introduction to C++ – STN Bruges
2009 Introduction to Banking, Brokerage and Insurance – Hogeschool Ghent
2009 Database Technology – Faculty of Engineering, Ghent University
2009 Internet Applications – Faculty of Engineering, Ghent University
2009 Effective Time Management – Doctoral Schools, Ghent University
2008 Operating Systems: Linux – PCVO Ghent
2007 Web-scripting (PHP/MySQL) – PCVO Ghent

Publications

Joachim De Schrijver et al. (2012), **Identifying genomic variation in a large number of samples using 3D-pooling and next-generation sequencing**, in preparation

Joachim De Schrijver et al. (2012), **Advancing the Variant Identification Pipeline**, in preparation

Geert Trooskens, Tim De Meyer, Veeck Jurgen, Simon Denil, Jean-Pierre Renard, Pierre Dehan, Joachim De Schrijver, Gerben Menschaert, Wim Van Criekinge (2012), **A map of the human methylome**, submitted

Joachim De Schrijver, Geert Trooskens, Linda J.W. Bosch, Pierre Dehan, Beatriz Carvalho, Leander Van Neste, Gerrit Meijer, Steve Baylin, Wim Van Criekinge (2012), **Genome-wide total RNA and MBD-sequencing in HCT116 and DKO cells as a global re-expression model**, submitted

Nico De Storme, Joachim De Schrijver, Wim Van Criekinge, Vera Wewer, Peter Dörman, Danny Geelen Danny (2012), **Sexual polyploidization through pre-meiotic cytokinesis defects in a callose synthase and sterol biosynthesis mutants**, submitted

Kristof De Beuf, Joachim De Schrijver, Rafael A. Irizarry, Olivier Thas, Wim Van Criekinge, Lieven Clement (2012), **Improved homopolymer base-calling of 454 pyrosequencing data using a weighted Poisson hurdle model**, BMC Bioinformatics 13:303

Sarah De Keulenaer, Jan Hellemans, Steve Lefever, Jean-Pierre Renard, Joachim De Schrijver, Hendrik Van de Voorde, Mohammad Amin Tabatabaiefar, Filip Van Nieuwerburgh, Daisy Flamez, Filip Pattyn, Bieke Scharlaken, Dieter Deforce, Sofie Bekaert, Wim Van Criekinge, Jo Vandesompele, Guy Van Camp, Paul Coucke (2012), **Molecular diagnostics for congenital hearing loss including 15 deafness genes using a next generation sequencing platform**, BMC Medical Genomics 5:17

Kim De Leeneer, Joachim De Schrijver, Lieven Clement, Machteld Baetens, Steve Lefever, Sarah De Keulenaer, Kathleen Claes, Filip Pattyn, Bram De Wilde, Paul Coucke, Jo Vandesompele, Jan Hellemans (2011), **Practical tools to implement massive parallel pyrosequencing of PCR products in next generation molecular diagnostics**, PloS ONE 6(9): e25531

Machteld Baetens, Lut Van Laer, Kim De Leeneer, Jan Hellemans, Joachim De Schrijver, Hendrik Van De Voorde, Marjolijn Renard, Bjorn Menten, Wim Van Criekinge, Julie De Backer, Anne De Paepe, Bart Loeys, Paul Coucke (2011), **Applying Massive Parallel Sequencing to Molecular Diagnosis of Marfan and Loeys-Dietz Syndrome**, Human Mutation 32: 1053 - 1062

Kim De Leeneer, Jan Hellemans, Joachim De Schrijver, Machteld Baetens, Wim Van Criekinge, Anne De Paepe, Paul Coucke, Kathleen Claes (2011), **Multiplexed amplicon sequencing of the breast cancer genes BRCA1&2: challenges, opportunities and limitations**, Human Mutation 32: 335 – 344

Michael Vyverman, Joachim De Schrijver, Wim Van Criekinge, Peter Dawyndt, Veerle Fack (2011), **Accurate long read mapping using enhanced suffix arrays**, Proceedings of the international conference on Bioinformatics Models, Methods and Algorithms 2011

Joachim De Schrijver, Kim De Leeneer, Steve Lefever, Nick Sabbe, Filip Pattyn, Filip Van Nieuwerburgh, Paul Coucke, Dieter Deforce, Jo Vandesompele, Sofie Bekaert, Jan Hellemans, Wim Van Criekinge (2010), **Analysing 454 amplicon resequencing experiments using the modular and database oriented Variant Identification Pipeline**, BMC Bioinformatics 11: 269

Acknowledgements

Michael Vyverman, Bernard De Baets, Veerle Fack, Peter Dawyndt (2012), **Prospects and limitations of full-text index structures in genome analysis**, Nucleic Acids Research (e-pub ahead of print)

Conference participations

Oral presentation (O) / Poster presentation (P) / Participation (*)

- MRP Bioinformatics - From Nucleotides to Networks (N2N): 2012 Event; Zwijnaarde, Belgium; 5 September, 2012 (P)
- Intelligent Systems for Molecular Biology (ISMB) 2012; Long Beach, CA, USA; 15-17 July, 2012 (P)
- Ghent University (Faculty of Medicine and Health Sciences) Science Day 2012; Ghent, Belgium; 14 March, 2012 (O)
- VIB-BGI Genomics Meeting; Leuven, Belgium; 15 February, 2012 (*)
- Benelux Bioinformatics Conference (BBC) 2011; Luxembourg, Luxembourg; 12-13 December, 2011 (P)
- Epigenomics-Next-Gen Sequencing; Brussels, Belgium; 16 November, 2011 (*)
- WOUD - Bioinformatics: Tools in Research; Ghent, Belgium; 28 September, 2011 (*)
- MRP Bioinformatics - From Nucleotides to Networks (N2N): Kick-Off Event; Zwijnaarde, Belgium; 4 May, 2011 (P)
- European Conference on Computational Biology (ECCB) 2010; Ghent, Belgium; 26-29 September 2010 (P)
- Third Belgium-Dutch-Luxembourg Next-Generation Sequencing Users Meeting; Nijmegen, Holland; 6 July, 2010 (*)
- Illumina's Sequencing Simplified Seminar; Brussels, Belgium; 16 June, 2010 (*)
- CHI's XGEN 2010, Applying NeXt GENeration Genomic Technologies for Now Generation Discoveries; San Diego, CA, USA; 15-19 March, 2010 (*)
- Advances in Genomics Symposium 2010: Next Generation Sequencing; Ghent, Belgium; 29 January, 2010 (P)
- Molecular Biotechnology Ugent: Kick-Off Event; Ghent, Belgium; 22 January, 2010 (O)
- Epiphamy: 3rd Bi-annual Symposium on DNA Methylation and Cancer; Aachen, Germany; 9 October, 2009 (*)

- Roche Molecular Biology Days: High Throughput Sequencing; Vilvoorde, Belgium; 8 October, 2009 (*)
- Second Belgium-Dutch-Luxembourg Next-Generation Sequencing Users Meeting; Utrecht, Holland; 7 July, 2009 (O)
- NXT-GNT SEEquencing Genes: Kick-Off Event; Ghent, Belgium; 8 May, 2009 (*)
- CHI's Next-Generation Sequencing 2009; San Diego, CA, USA; 17-19 March, 2009 (P)

Awards & honors

- European Conference on Computational Biology (ECCB) 2010 Benelux Fellowship Award
- Bioperl Programming Challenge (BPC) 2007

Appendix

Appendix File 5.1: Detailed overview of the reports generated by the VIP pipeline.

Appendix File 5.2: Overview of the parameters reported by VIP for each variant

Appendix File 5.3: Example demonstrating the variant conversion algorithm of VIP enabling deletions and insertions to be detected.

Appendix File 5.4: Example output of the VIP alignment visualizer.

Appendix File 5.5: Comparison of the variants detected by VIP and AVA.

Appendix File 6.1: Schematic overview of the integration of the Variant Identification Pipeline and the Variant Interpretation Pipeline.

Appendix File 6.2: Overview and description of all the parameters analysed by the Variant Interpretation Pipeline.

Appendix File 6.3: Example output of the Variant Interpretation Pipeline.

Appendix File 7.1: Example sample/pool input file for SNP-CUB3.

Appendix File 7.2: Validation dataset analysis using SNP-CUB3.

Appendix File 8.1: Complete overview of the Gene Ontology (GO) terms enriched in the set of re-expressed genes obtained by comparing HCT116 and DKO, but not considering methylation.

Appendix File 8.2: Complete overview of the Gene Ontology (GO) terms enriched in the set of re-expressed genes obtained by comparing HCT116 and DKO, and considering promoter methylation.

Appendix File 8.3: Overview of the significantly upregulated pathways in DKO (global re-expressoin).

Appendix File 8.4: List of all the re-expressed genes obtained by comparing HCT116 and DKO, and considering promoter methylation.

Appendix File 8.5: List all the re-expressed snoRNAs in DKO compared to HCT116.

Appendix File 8.6: Overview of the *TLR2* gene region in HCT116 and DKO.

Appendix File 8.7: Overview of the expression and methylation status of the top re-expressed genes and genes present in the PI3K/AKT/mTOR, MAPK/ERK, P53 and WNT pathways (67 genes in total) in colorectal cancer cell lines.

Appendix File 8.8: Overview of the expression values and methylation status of the genes of the WNT pathway, mTOR pathway, MAPK pathway and p53 pathway in HCT116, DKO and the additional cell lines.

Appendix File 8.9: Overview of the *TOM1L1* and *COX11* region. Demonstrating the advantages of directional RNA-sequencing over conventional RNA-sequencing.

Appendix File 5.1: Detailed overview of the reports generated by the VIP pipeline.

Coverage analysis per MID: Gives a report showing the abundance of each sample/MID in the experiment. Can be of interest to see whether samples were pooled in an orderly fashion.

Coverage analysis per amplicon: Gives a report showing the abundance of each amplicon in the experiment. Can be of interets to see whether all PCR experiments performed well.

Coverage analysis per MID/amplicon: Gives a report showing the abundance of each amplicon per MID in the total experiment. When either 'Coverage analysis per MID' or 'Coverage analysis per amplicon' shows an aberration, one can look up whether it is a sample of amplicon specific problem. For example, low coverage for a specific amplicon for only one of the many MIDs indicates a lab error, whereas low coverage for a specific amplicon for all MIDs indicates a global PCR problem.

Short sequence analysis: Gives an overview per amplicon of the detected 'short sequences'. A high number of 'short sequences' for a certain amplicon indicates that the PCR is producing primer-dimers rather than full lenght products.

Length distribution: Gives an overview of the length distribution of the produced sequence reads. Is of interest to assess the overall sequencing quality of an experiment.

Quality score analysis: Gives an overview of the average quality score per position. Is of interest to assess the overall sequencing quality of an experiment. Rapidly declining quality scores generally indicates poor sequencing quality.

Coverage analysis per single base: Reports coverage for each amplicon on an individual nucleotide resolution. Offers a more detailed view than 'Coverage analysis per amplicon'. Allows large deletions or duplications to be detected.

Variation analysis: Reports variation (single nucleotide variation (SNV), deletions (DEL), or insertions (INS) per amplicon and per MID.

Appendix File 5.2: Overview of the parameters reported by VIP for each variant

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile5.2.xlsx.

Appendix File 5.3: Example demonstrating the variant conversion algorithm of VIP enabling deletions and insertions to be detected.

Imagine a reference sequence ACTGACTGAAAATGTGT and a sample sequence with a 2bp deletion: ACTGACTGAATGTGT

The sample sequence is sequenced in both the forward and reverse direction. Due to sequencing errors different sequences are obtained. After mapping them back onto the reference sequence, different variation is reported (column 2). Variation is converted and processed by the VIP (column 3 and 4).

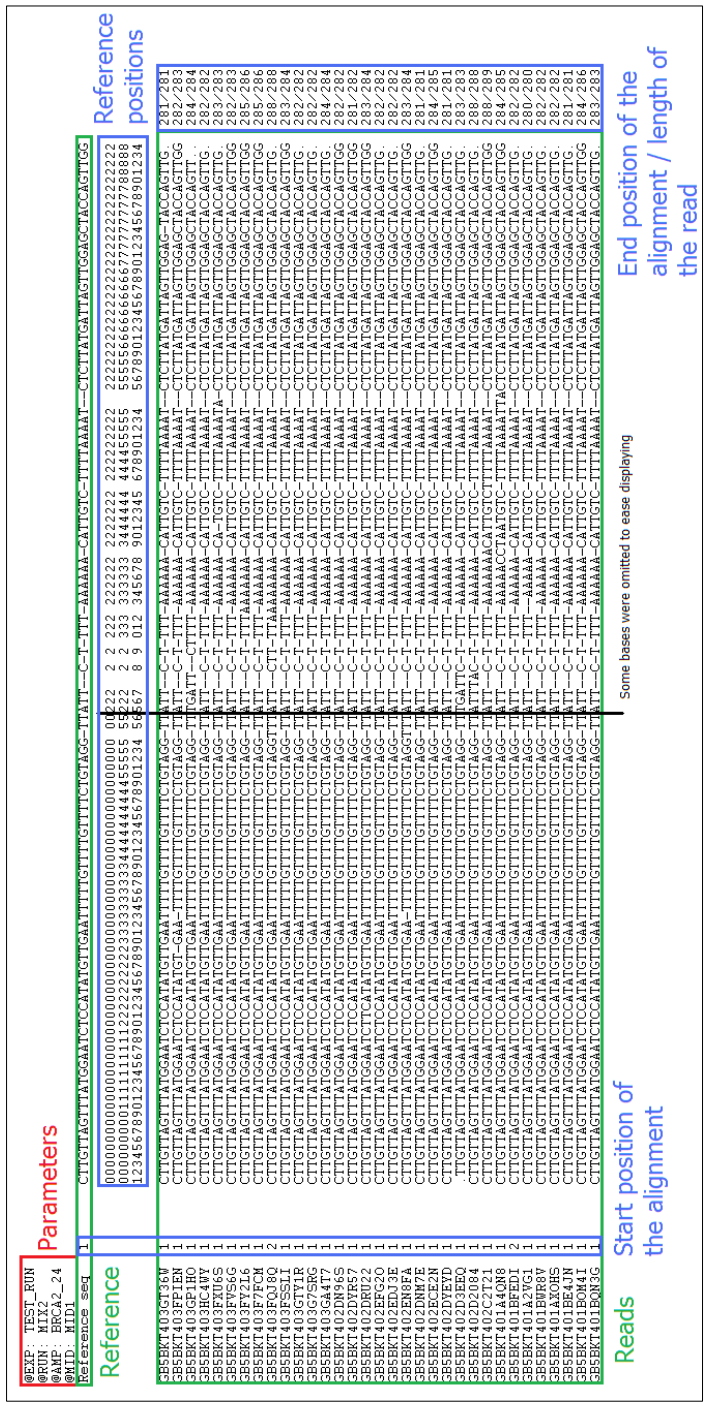
Sequence	Reported variation	Variation converted by VIP	Variation processed by VIP
ACTGACTGAATGTGT	AA deletion at pos 11	AA deletion at pos. 9	DEL9A; DEL10A
ACTGACTGAAATGTGT	A deletion at pos 9	A deletion at pos. 9	DEL9A;
ACTGACTGAATGTGT	AA deletion at pos 9	AA deletion at pos. 9	DEL9A; DEL10A
ACTGACTGAATGTGT	AA deletion at pos 10	AA deletion at pos. 9	DEL9A; DEL10A
ACTGACTGAATGTGT	AA deletion at pos 9	AA deletion at pos. 9	DEL9A; DEL10A
ACTGACTGATGTGT	AAA deletion at pos 9	AAA deletion at pos. 9	DEL9A; DEL10A; DEL11A
ACACATTTCAGTCAGT	T deletion at pos 12	A deletion at pos. 9	DEL9A;
ACACATTCAGTCAGT	TT deletion at pos 11	AA deletion at pos. 9	DEL9A; DEL10A
ACACATTCAGTCAGT	TT deletion at pos 11	AA deletion at pos. 9	DEL9A; DEL10A
ACACATTCAGTCAGT	TT deletion at pos 11	AA deletion at pos. 9	DEL9A; DEL10A
ACACATCAGTCAGT	TTT deletion at pos 10	AAA deletion at pos. 9	DEL9A; DEL10A; DEL11A
ACACATTCAGTCAGT	TT deletion at pos 9	AA deletion at pos. 9	DEL9A; DEL10A

In the variant reporting step, these variants are aggregated and reported to the end-user.

Position 09: Deletion A, coverage 12, absolute frequency 12
Position 10: Deletion A, coverage 12, absolute frequency 10
Position 11, Deletion A, coverage 12, absolute frequency 2

Using this data, it becomes obvious that the most probably deletion is a 2 bp deletion at position 9-10. Aggregating the variation as reported by BLAT would have made it impossible to come to this conclusion.

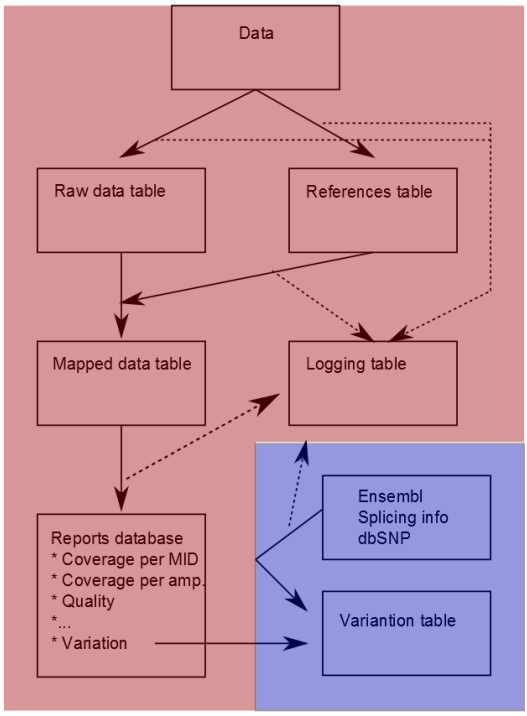
Appendix File 5.4: Example output of the VIP alignment visualizer.



Appendix File 5.5: Comparison of the variants detected by VIP and AVA.

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile5.5.xls.

Appendix File 6.1: Schematic overview of the integration of the Variant Identification Pipeline (red) and the Variant Interpretation Pipeline (blue).



Appendix File 6.2: Overview and description of all the parameters analysed by the Variant Interpretation Pipeline.

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile6.2.xlsx.

Appendix File 6.3: Example output of the Variant Interpretation Pipeline

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile6.3.xlsx.

Appendix File 7.1: Example sample/pool input file for SNP-CUB3.

```
### USE THE FOLLOWING FORMAT ###
### Sample<tab>Pool1,Pool2,Pool3 ###
### Sample and Pool should be integers (numbers) ###
1       1,4,7
2       1,5,7
3       1,6,7
4       2,4,7
5       2,5,7
6       2,6,7
7       3,4,7
8       3,5,7
9       3,6,7
10      1,4,8
11      1,5,8
12      1,6,8
13      2,4,8
14      2,5,8
15      2,6,8
16      3,4,8
17      3,5,8
18      3,6,8
19      1,4,9
20      1,5,9
21      1,6,9
22      2,4,9
23      2,5,9
24      2,6,9
25      3,4,9
26      3,5,9
27      3,6,9
```

Appendix File 7.2: Validation dataset analysis using SNP-CUB3.

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile7.2.xls.

Appendix File 8.1: Complete overview of the Gene Ontology (GO) terms enriched in the set of re-expressed genes obtained by comparing HCT116 and DKO, but not considering methylation.

This Word file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.1.doc.

Appendix File 8.2: Complete overview of the Gene Ontology (GO) terms enriched in the set of re-expressed genes obtained by comparing HCT116 and DKO, and considering promoter methylation.

This Word file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.2.doc.

Appendix File 8.3: Overview of the significantly upregulated pathways in DKO (global re-expression).

KEGG pathways

Pathway	Enrichment Score	Norm. Enrich. Score	p-value
HEMATOPOIETIC CELL LINEAGE	0.62	1.61	0.002
TOLL LIKE RECEPTOR SIGNALING PATHWAY	0.58	1.56	0.001
LEISHMANIA INFECTION	0.62	1.55	0.002
GRAFT VERSUS HOST DISEASE	0.67	1.52	0.018
NOD LIKE RECEPTOR SIGNALING PATHWAY	0.60	1.47	0.017
PRIMARY IMMUNODEFICIENCY	0.66	1.46	0.025
CHEMOKINE SIGNALING PATHWAY	0.51	1.44	0.003
CELL ADHESION MOLECULES CAMS	0.51	1.40	0.011
RETINOL METABOLISM	0.53	1.35	0.035
JAK STAT SIGNALING PATHWAY	0.47	1.32	0.025
FOCAL ADHESION	0.45	1.29	0.022
CYTOKINE CYTOKINE RECEPTOR INTERACTION	0.44	1.28	0.014

Reactome pathways

Pathway	Enrichment Score	Norm. Enrich. Score	p-value
CHEMOK. RECEPTORS BIND CHEMOKINES	0.76	1.83	0.000
COMPLEMENT CASCADES	0.77	1.82	0.000
CLASSICAL ANTIBODY MEDIATED COMPLEMENT ACTIVATION	0.79	1.78	0.000
INITIAL TRIGGERING OF COMPLEMENT	0.72	1.67	0.002
IMMUNOREGULATORY INTERACTIONS BETWEEN LYMPHOID AND NON LYMHOID	0.63	1.66	0.000
INNATE IMMUNITY SIGNALING	0.54	1.48	0.001
SIGNALING IN IMMUNE SYSTEM	0.48	1.44	0.000
PD1 SIGNALING	0.67	1.43	0.037
G PROTEIN BETA GAMMA SIGNALING	0.66	1.44	0.043
HEMOSTASIS	0.45	1.30	0.015
FORMATION OF PLATELET PLUG	0.45	1.27	0.039

Biocarta pathways

Pathway	Description	Enrichment Score	Norm. Enrich. Score	p-value
LAIR	Cells and Molecules involved in local acute inflammatory response	0.91	1.76	0.000
IL1R	Signal transduction through IL-1 receptor	0.75	1.67	0.002
INFLAM	Cytokines and inflammatory Response	0.71	1.56	0.005
COMP	Complement pathway	0.78	1.54	0.012
NTHI	NFkB activation by nontypeable hemophilus influenza	0.70	1.50	0.015
TOLL	Toll-like receptor pathway	0.65	1.48	0.016
IL22BP	IL-22 soluble receptor signaling pathway	0.76	1.47	0.041
IL10	IL-10 anti-inflammatory signaling pathway	0.74	1.47	0.031
IL2	IL-2 signaling pathway	0.70	1.45	0.40
NKFB	NF-kB signaling pathway	0.70	1.45	0.043
STEM	Regulation of hematopoiesis by cytokines	0.76	1.44	0.046
CD40	CD40L signaling pathway	0.74	1.43	0.047

Appendix File 8.4: List of all the re-expressed genes obtained by comparing HCT116 and DKO, and considering promoter methylation.

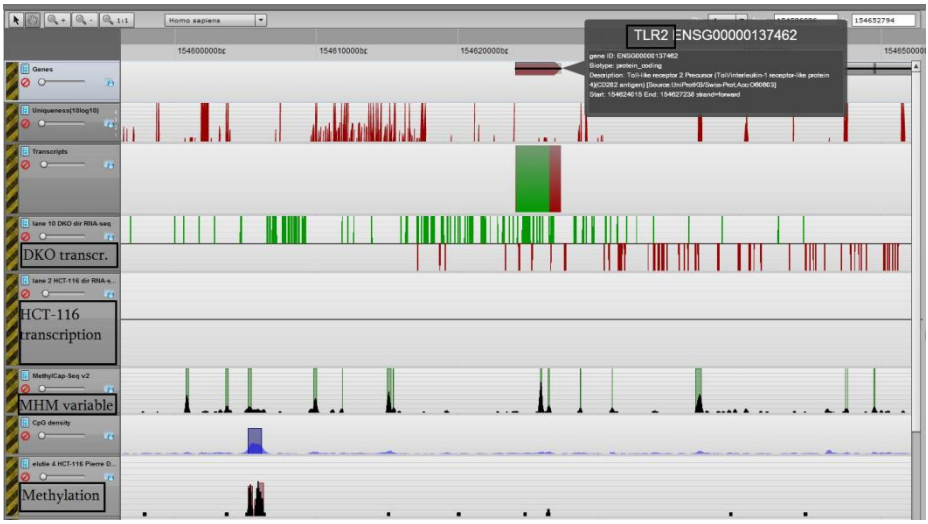
This Word file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.4.doc.

Appendix File 8.5: List all the re-expressed snoRNAs in DKO compared to HCT116.

Transcript id	Gene id	Gene name	HCT116	DKO
236180	ENSG00000223213	SNORD81	0.00	0.3871
236313	ENSG00000209418	SNORD64	0.00	83.8269
236457	ENSG00000202275	SNORD51	0.00	0.1264
238174	ENSG00000201009	SNORD46	0.00	0.5293
237413	ENSG00000200706	SNORD45	0.00	0.5056
236403	ENSG00000212302	SNORD41	0.00	0.4898
238802	ENSG00000239112	SNORD123	0.00	14.6387
235066	ENSG00000207245	SNORD116	0.00	191.4170
165390	ENSG00000224078	SNORD115-13	0.00	0.3397
165407	ENSG00000224078	SNORD115-13	0.00	0.2291
165361	ENSG00000224078	SNORD115-13	0.00	0.1422
165233	ENSG00000224078	SNORD115-13	0.00	0.1027
235791	ENSG00000212428	SNORD115	0.00	0.6241
237163	ENSG00000202191	SNORD113	0.00	2.8677
233434	ENSG00000239014	SNORD108	0.00	0.6715
234909	ENSG00000238810	SNORD107	0.00	19.5604
237625	ENSG00000200130	SNORA63	0.00	0.2844
234373	ENSG00000201516	SNORA51	0.00	0.6320
234101	ENSG00000207502	SNORA42	0.00	0.4266
232588	ENSG00000199571	SNORA22	0.00	0.9796
232977	ENSG00000212293	SNORA16	0.02	0.2528
232807	ENSG00000221245	SNORA11	0.00	0.4503

Normalized expression values are shown for HCT116 and DKO.

Appendix File 8.6: Overview of the *TLR2* gene region in HCT116 and DKO.



Appendix File 8.7: Overview of the expression and methylation status of the top re-expressed genes and genes present in the PI3K/AKT/mTOR, MAPK/ERK, P53 and WNT pathways (67 genes in total) in colorectal cancer cell lines.

This Word file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.7.doc.

Appendix File 8.8: Overview of the expression values and methylation status of the genes of the WNT pathway, mTOR pathway, MAPK pathway and p53 pathway in HCT116, DKO and the additional cell lines.

This Excel file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.8.xls.

Appendix File 8.9: Overview of the *TOM1L1* and *COX11* region. Demonstrating the advantages of directional RNA-sequencing over conventional RNA-sequencing.

This Word file is available online from the following location: http://athos.ugent.be/joachim_ds/PhD/AppendixFile8.9.doc.