KTH Biotechnology

Interrogation of Nucleic Acids by Parallel Threading

Erik Pettersson

Royal Institute of Technology
School of Biotechnology

Stockholm 2007

ISBN 978-91-7178-802-3

Royal Institute of Technology
School of Biotechnology
AlbaNova University Center
SE-106 91 Stockholm
Sweden

Printed at Universitetsservice US-AB
Drottning Kristinas väg 53B
SE-100 44 Stockholm
Sweden

Erik Pettersson (2007): Interrogation of Nucleic Acids by Parallel Threading
Department of Gene Technology, School of Biotechnology, Royal Institute of Technology (KTH), Stockholm, Sweden
ISBN 978-91-7178-802-3

ABSTRACT

Advancements in the field of biotechnology are expanding the scientific horizon and a promising era is envisioned with personalized medicine for improved health. The amount of genetic data is growing at an ever-escalating pace due to the availability of novel technologies that allow massively parallel sequencing and whole-genome genotyping, that are supported by the advancements in computer science and information technologies. As the amount of information stored in databases throughout the world is growing and our knowledge deepens, genetic signatures with significant importance are discovered. The surface of such a set in the data mining process may include causative- or marker single nucleotide polymorphisms (SNPs), revealing predisposition to disease, or gene expression signatures, profiling a pathological state. When targeting a reduced set of signatures in a large number of samples for diagnostic- or fine-mapping purposes, efficient interrogation and scoring require appropriate preparations. These needs are met by miniaturized and parallelized platforms that allow a low sample and template consumption.

This doctoral thesis describes an attempt to tackle some of these challenges by the design and implementation of a novel assay denoted Trinucleotide Threading (TnT). The method permits multiplex amplification of a medium size set of specific loci and was adapted to genotyping, gene expression profiling and digital allelotyping. Utilizing a reduced number of nucleotides permits specific amplification of targeted loci while preventing the generation of spurious amplification products. This method was applied to genotype 96 individuals for 75 SNPs. In addition, the accuracy of genotyping from minute amounts of genomic DNA was confirmed. This procedure was performed using a robotic workstation running custom-made scripts and a software tool was implemented to facilitate the assay design. Furthermore, a statistical model was derived from the molecular principles of the genotyping assay and an Expectation-Maximization algorithm was chosen to automatically call the generated genotypes.

The TnT approach was also adapted to profiling signature gene sets for the Swedish Human Protein Atlas Program. Here 18 protein epitope signature tags (PrESTs) were targeted in eight different cell lines employed in the program and the results demonstrated high concordance rates with real-time PCR approaches. Finally, an assay for digital estimation of allele frequencies in large cohorts was set up by combining the TnT approach with a second-generation sequencing system. Allelotyping was performed by targeting 147 polymorphic loci in a genomic pool of 462 individuals. Subsequent interrogation was carried out on a state-of-the-art massively parallelized Pyrosequencing instrument. The experiment generated more than 200,000 reads and with bioinformatic support, clonally amplified fragments and the corresponding sequence reads were converted to a precise set of allele frequencies.

Keywords: genotyping, multiplex amplification, trinucleotide threading, single nucleotide polymorphism, genotype calling, Expectation-Maximization, protein-epitope signature tag, expression profiling, Human Protein Atlas, pooled genomic DNA, Pyrosequencing, 454, allelotyping, association studies, bioinformatics

"...she's the most beautiful thing I've ever laid eyes on... Eureka! We have a thread!"

In a research facility at 59°21'13.35"N; 18°3'31.45"E. 19/12/2003. 11:51:26 GMT.

LIST OF PUBLICATIONS

This thesis is based on the papers listed below which will be referred to by their Roman numerals.

I. **Erik Pettersson**, Mats Lindskog, Joakim Lundeberg & Afshin Ahmadian. Tri-nucleotide Threading for parallel amplification of minute amounts of genomic DNA. *Nucleic Acids Res* **34**(**6**), e49 (2006).

II. Hedvig Norlén, **Erik Pettersson**, Afshin Ahmadian, Joakim Lundeberg & Rolf Sundberg. Classification of SNP genotypes by a Gaussian mixture model. *Manuscript.*

III. Pawel Zajac, **Erik Pettersson**, Marcus Gry, Joakim Lundeberg & Afshin Ahmadian. Expression profiling of signature gene sets with tri-nucleotide threading. *Genomics. In press.*

IV. **Erik Pettersson**, Pawel Zajac, Patrik L. Ståhl, Josefin A. Jacobsson, Robert Fredriksson, Claude Marcus, Helgi B. Schiöth, Joakim Lundeberg & Afshin Ahmadian. Allelotyping by massively parallel Pyrosequencing of SNP-carrying trinucleotide threads. *Human Mutation. In press.*

CONTENTS

INTERROGATION
OF NUCLEIC ACIDS
BY PARALLEL THREADING

PROLOGUE

FROM THE HMB ENDEAVOUR TO JAY FLATLEY'S iPHONE AND BACK TO THE TEMPLE OF APOLLO

Although exploration is part of human nature, the endeavors are clearly century dependent. Adventurers of today are somewhat different from Vasco da Gama and James Cook, even if some are still using sailing vessels (or more precisely a 100-foot sailing yacht) circumnavigating the globe on a quest to discover and explore brand new worlds[1-3]. Yet, while most such explorers of today are armed with a lab coat and a MacBook rather than a sloop, they are investigating the very same worlds – the ones of genes, genomes and those spawned thereof – transcriptomes, proteomes, metabolomes and phenomes. The curiosity and urge to explore are wired deep in our genes but the voyages for the 21st century are about looking inside ourselves, finding the very core of these traits and investigating the molecules of life. Grasping the beauty of the ocean, or simply holding this book and reading these lines – comprehending, remembering, imagining and feeling – all possible because of the billions of tasks carried out by the complex biomolecules operating in the various cells throughout our bodies. Allowing us to experience the sound of waves spraying high over the bows, the cold wind and smell of the ocean, the taste of salt on the lips and the warmth of the sun. Registered and transmitted by proteins acting as enzymes, receptors, transporters, signal peptides and structural elements. Translated and regulated by ribonucleic acids, transcribed from the blueprints in the cell nucleus – a seemingly endless series of nitric bases. DNA. A genome. A code for life. Humans, light-harvesting cyanobacteria, dragonflies, sea turtles or whales, all with unique genomes as representatives of some of the Earth's millions of different species. Pick up a handful of soil, and you'll have five to ten thousand species in your hand, most of which are still unknown.

The possibility of decoding genomes is changing the scientific horizon. The sailor with the 100-foot sloop envisions the production of clean and renewable energy by designing synthetic bacteria for hydrogen or ethanol production, thus lessening our need for fossil fuels[4]. Mining sequence data from microorganisms of the oceans provides us with knowledge of many novel genes, some which may be part of an artificial, custom made bacteria[5] in the service of humanity[6], softening the economical and environmental impact of the ever escalating climate change. Besides greentech and cleantech, venture capital firms are now focusing on technology investments preventing pandemics and bioterrorism[7]. Knowledge of the genetic sequence of the coronavirus causing SARS[8] or the H5N1-strain of Influenza A that causes avian flue[9] will be an aid in diagnostics, vaccine production and the design of new antiviral agents. Pathogens of all forms, HIV[10], tuberculosis[11] and malaria[12] are acquiring resistance to drugs at an alarming rate[13]. Knowledge of genetic compositions and metabolic pathways will decide upon suitable treatment and render rational design of novel drugs possible.

The closest genome to explore is our own, and we have much to gain by doing so. Since Genentech[14] pioneered cloning and manufacturing of human insulin in the beginning of the 1980's, many more genetically engineered drugs have hit the market as a consequence of biotechnological achievements and the knowledge of human genetic sequences. In 2001, the sequence of more or less the entire human genome saw the light of day at an estimated cost of $3 billion[15,16]. Since then, methods for investigating individual variations, in particular the millions of single nucleotide polymorphisms (SNPs), have emerged. The pace of change has been very rapid and the level of innovation in this field has exceeded all expectations. The cost for scoring a single SNP has dropped since the beginning of the new millennium with a factor of a thousand and thus a million SNPs throughout a person's genome can be analyzed for less than a $1,000[17,18]. This knowledge will work for us medically by pinpointing the combinations of

genetic variations in complex genetic traits such as cardiovascular disease, schizophrenia, diabetes, obesity and cancer and thereby impending an era of personalized and preventive medicine as well as pharmacogenomics. Awareness of individual genetic profiles will prevent the inappropriate prescription of drugs, which currently results in the death of hundreds of thousands of people each year in adverse drug reactions[19]. Besides diminishing such adverse events, the success rate in clinical trials may be improved and as a result, a larger number of specialized drugs will reach the pharmaceutical market, targeting the many non-responders to common drugs. Comprehending the genetic information, including the genome of mitochondria, may shed light on the aging process and how we can optimize life style, nutrition and elucidate the possibility of "Reprogramming your biochemistry for immortality"[20].

Nevertheless, there are other variations in the genome. Sequencing technologies are improving at the same rate as SNP scoring techniques and thereby reducing the costs by a factor of two or three each year, hence comparable to Moore's Law in electronics and computer science[21]. Currently, $1 million spent on engine development in Formula 1 would give an average gain of 4 ms of lap time[22]. As a contrast, the same amount of money will provide us with a resequence of an individual human genome – the sequence of 6 billion bases[23]. At this cost, we find ourselves at a point in history where several individual genomes are starting to appear, although unfortunately these "early adopters" may find it as useful as the first telephones. In 2007-2008, the price may drop to $100,000 and within a decade it will be possible to sequence a human genome for $1,000[21]. Several rewards have been proposed to facilitate these efforts. The private Archon X Prize in Genomics has announced a $10 million cash reward for the first team to sequence 100 human genomes in 10 days[24]. The $1,000 genome is a true challenge – but with a biotech industry filled with competitive and committed people – people like the majority owner of MerckSerono and winner of America's Cup in 2003 and 2007 – it is likely to happen sooner than we think.

By then, genotyping of SNPs may have been pushed into the consumer market. Genetic analysis has been commercial for quite a while, in the form of single gene or polymorphic analysis, revealing, for instance, whether you are more likely winning when participating in a marathon than when sprinting a hundred meters[25-28]. With the analysis of millions of markers at a time, as described above, genetics is transforming into genomics. One provider of these analyses, San Diego based Illumina[18], a company where the financials are as exciting as the science, is teaming up with Google and Genentech in the form of Mountain View based start-up 23andMe[29]. Genomics has become an information science. Before the end of this year, you can have your genome analyzed by Illumina and by logging on to the 23andMe-website browse through parts of your genetic makeup. Jay Flately, CEO of Illumina Inc., recently piloted the software and now has his genome on his new Apple iPhone[30]. As an "early adopter" (or "technology geek" as his wife put it) he can now browse his genome with his fingertips. 23andMe will be focusing on many aspects, including medical, but initially answer questions such as "Am I more related to my brother than to my sister?" and "Which are my distant relatives?". An educated guess is that the business model is similar to Google's – but instead of selling advertisements based on search preferences, it will be based on genetic profiles. Illumina estimates that in a few years, the market in consumer genotyping will hit $1 billion. Imagine the personalized adverts depending on earwax type, evening preference or novelty seeking.

One's destiny is a result of genes, environment, life style, behavior and luck. Although in some cases, knowledge of a few of your alleles might just save your life. There is a new meaning to the ancient inscription at the temple of Apollo. "Nosce te ipsum" – Know thyself.

INTRODUCTION

CHAPTER ONE

THE TLAs IN DNA

Dry earwax is due to a typical SNP. Drug metabolism may depend on a CNV. And there are MNPs, MSVs and STRs. All embedded in the DNA – deoxyribonucleic acid – a double helix formed by two polynucleotide strands twisted around each other, storing the genomic information. Each polymer, a string of the four nitric bases – Adenine (A), Guanine (G), Cytosine (C) and Thymine (T), form Watson and Crick base pairing complementarily to its counterpart, hence gaining stability and enabling replication[31]. The G-C formation in the coil is stronger than the A-T, with three and two hydrogen bonds respectively. These features make synthetic DNA an efficient and lightweight storage of information (1 g of dehydrated DNA stores data at a density 11 orders of magnitude greater than a present day DVD) as well as a possible tool for parallelized problem solving[32-34].

Since humans are diploid organisms, i.e. with genetic material from both parents, the three billion base pairs that constitute the genome actually are twice as many, distributed along 46 chromosomes (23 from each parent) and stored in the cell nucleus together with histones, known as chromatin. Each cell also contains a few thousand mitochondria, each carrying a separate genome of approximately 15 kb (kilo bases).

The genomic sequence is as cryptic as the complexity of the products it encodes. Regions of the genome encode for different ribonucleic acid polymers (RNAs), or transcripts, acting as templates for protein translation or with regulatory functions. There is an ongoing debate of what actually is a gene, and a recent definition suggests that a gene is "a union of genomic sequences encoding a coherent set of potentially overlapping functional products"[35]. This somewhat more vague definition has arisen due to the fact that the genome is full of overlapping transcripts as a result of overlapping genes and genes within genes[36]. There is also antisense transcription, believed to be a part of gene expression regulation, which seems to be a common regulatory mechanism in the genome[37,38]. Other transcripts encode microRNAs (miRNAs), small nucleolar RNAs (snoRNAs) and small interfering RNA (siRNA) and a large fraction of the RNAs encoded by the genome will not be translated into a protein[36]. Non-coding transcripts have also been identified in regions of the genome previously thought to be transcriptionally silent and one might suspect that RNAs are very active regulators of the cellular processes and that the transcriptome is far larger and much more complex than previously believed[39].

The genes that do encode proteins (approximately 23,000)[40] consist of exons (coding sequence or CDS) interspersed with non-coding introns (intervening sequence or IVS) which by splicing (i.e. combining different exons by the removal of introns in the mRNA template) renders different protein products. Approximately 50% of all protein coding genes are thought to give different such products. Taken into account the post-translational modifications the total number of possible proteins may reach 100k, in spite of the relatively low number of protein coding genes and the fact that such exons only constitute about 1.2% of the genome[16]. Novel processes, such as trans-splicing, has been reported in model organisms, where mRNA molecules originating from different parts of the genome are joined by ligation, thus further complicating the picture[41]. Similar as for the transcriptome, by specifying a certain cell type or tissue at a certain point in time, one can define a proteome. The flow from DNA to RNA into proteins is known as the Central Dogma, although it might benefit from some revision considering the recent findings in regulatory functions of the transcriptome.

A recent estimate of a 99.5% similarity between any two non-related humans indicates a greater diversity than previously reported[42,43]. Variations include macro-, mini-, and microsatellites of which the latter, also known as STRs (short tandem repeats) or SSRs (simple sequence repeats) are di-, tri-, or tetra repeats used in for instance genetic mapping and DNA fingerprinting in forensic investigations. The most abundant form of genetic variation is the single nucleotide polymorphism (SNP) and more than 10

million have been reported[40,43]. This polymorphism is a single base substitution where the minor allele must be present in a defined population at a frequency of at least 1%. SNPs are found everywhere in the genome, on average 1 in every 300 bp, thus giving rise to different phenotypes. SNPs have been linked to traits such as hair-, eye- and skin color, short term memory efficiency, earwax type (dry/wet), alcohol-dependence, prevalence to tobacco addiction (as well as protection to), effect of pravastatin treatment, drug metabolism and morningness-eveningness preference[42,44].

Other variations found in the genome include copy-number variations (CNVs) where, for instance, the number of CYP2D6-gene copies will affect drug metabolism[19]. Another example is the salivary amylase gene (AMY1), where the number of gene copies is correlated positively with the enzyme levels in the saliva, thus affecting the hydrolysis of starch[45]. More complex variations include multi-nucleotide polymorphisms (MNPs) and multi-site variations (MSVs) and since chunks of DNA are moving around there are insertions, deletions (indels) and inversions[43]. The first diploid human sequence indicates 4.1 million DNA variants (3.2 million SNPs) spanning a total of 12.3 million bases where 30% of the variations are previously unreported[42]. The availability of additional genomes in the near future will further our understanding of human genetic variation, but currently, a difference between any two humans in the range of 15-30 million bases is expected[46]. Finally, an important variation found within many cancer cells is the loss of heterozygosity (LOH) where the final functioning copy of a tumour-supressor gene (i.e. caretaker or gatekeeper) is lost through mutation or complete loss.

Genetic predisposition to disease comes in many forms and the genetic contribution to the trait may vary. In cases involving classical Mendelian traits, a variation in a single gene will suffice to cause disease, as in autosomal dominant disorders (Huntington's disease and Familial Hypertrophic Cardiomyopathy) or as in autosomal recessive disorders (Cystic Fibrosis). In these monogenic afflictions, the penetrance can be 100%. Although family history is one of the strongest risk factors for nearly all diseases, a phenotype cannot always be derived from a single locus. A multifactorial disease arises due to a set of causative alleles, each contributing to an increase in risk, in combination with environmental- and life style factors. By comparing affected individuals and controls, one may find genetic influences on disease and causative variations, notwithstanding the fact that examining all genetic differences in a large number of individuals currently is not possible. SNPs are then optimal as markers for identifying contributing loci to complex traits since they are distributed evenly across the genome at a high frequency. The block nature of the genome further facilitates the efforts, since an SNP at one locus often gives information about variants of an entire haplotype. Approximately 1 million SNPs in 270 individuals from four different populations were initially genotyped in the HapMap project[47], rendering haplotype blocks with SNPs in linkage disequilibrium interspersed with hot-spots of recombination. Recently, results from phase II of the project were published where an additional 2.1 million markers were genotyped in the same individuals[48]. Employing the haplotype map, the number of SNPs to be typed in an association study can be significantly reduced and this knowledge is now being used for whole-genome association studies in the quest for markers of disease[49]. The studies require large cohorts to reduce the risk of erroneous associations and recent findings include markers in prostate-, and colorectal cancer[50-53], childhood asthma[54] and type 2 diabetes[55,56]. Causative SNPs or other variations can then be identified by increasing the resolution at the locus of interest and later used in preventive diagnostics and treatment.

Equally intensive efforts are being made at investigating the transcriptome and proteome. It has been shown that a moderate number of transcripts are sufficient to correctly diagnose and characterize a cancer form[57] and by examining the proteome, for instance by generating antibodies to localize the proteins[58,59], one can find promising biomarkers for many forms of disease, such as prostate- and ovarian cancer.

CHAPTER TWO

THE NOBLE ART OF MULTIPLEXING

Imagine being as euphoric as Ross Geller, when holding an amber with an embedded, fossilized mosquito from the Jurassic period in the palm of your hand[60]. Go back 150 million years to the point of attack and there are cells in the blood, with genomic landscapes as wide as they are miniature, each a blueprint of your favorite dinosaur. But just as you need a magnifying glass to carefully study the mosquito in detail, you need amplifying aid to methodically study the genetic code of your dinosaur. One should never let go of hope, but the creation of a Jurassic Park will most likely remain fiction. Nevertheless, there are vast amounts of undegraded DNA in the world worth studying and that process would benefit from sophisticated amplification methods enabling efficient and accurate sequence investigation.

A top goal in genomics is the ability to investigate any genetic region of interest, preferably an entire genome, swiftly, accurately and at a low cost starting from minute amounts of material. Such scarce DNA samples may include micro dissections from tumors, samples from biobanks or crime scenes and the previously mentioned ancient DNA and paleogenomics[61-63]. There are methods for investigating genetic sequences at a single molecule level, but these are still at the proof-of-concept stage[64]. An example of such a method is the monitoring of nucleotide incorporation events in real-time by carrying out a reaction in a zero-mode waveguide[65], rendering an effective observation volume in the order of 10 zeptoliters (10^{-21}). Unfortunately, the practical and cost-effective use of these promising methods for high-throughput sequence analysis lies in the future, although the efforts are beginning to show results[66]. Instead, in the years to come, we are relying on methods exploiting biomolecular principles and enzymes for target amplification and preparation prior to downstream sequence analysis. There are a number of amplification strategies, which can be combined with different downstream detection platforms, and commercial endeavors have adopted different strategies for the amplification prior to sequence interrogation.

The replication and amplification of DNA templates to detectable levels can be performed targeting either specific loci, while reducing complexity, or whole genomes, preferably preserving the sequence representation. Multiplexing, i.e. specifically amplifying many regions in parallel, is challenging considering the large size of genomes, the ambition of constantly reducing cost and the often limited supply of starting material.

A CHAIN REACTION BY POLYMERIZATION

By designing a pair of locus-specific oligonucleotides, adding a thermostable polymerase and reagents while alternating the reaction temperature, a chain reaction can be achieved, exponentially amplifying the target while reducing complexity[67,68]. The polymerase chain reaction, or PCR, has evolved to its current state through a series of improvements and is now fully automated with thermocyclers adjusting reaction temperature for each cycle i.e. (1) denaturing of strands, (2) annealing of new oligonucleotides (primers) and (3) primer elongation. In theory, the number of fragments is doubled in each cycle, hence rendering an exponential amplification. Variants of the assay include nested PCR, for further reduction in complexity, long-range PCR, for fragments up to 40 kb, and quantitative PCR for assessing the number of starting molecules. There are also methods for whole genome amplification (WGA) relying on degenerate primers. The techniques include primer extension preamplification (PEP)[69] and degenerate

oligonucleotide primed PCR (DOP)[70]. However, there are drawbacks to PCR. The rivalry between free primers and elongated such (i.e. products), eventually hampering and halting the reaction, is one issue. Perhaps more critical is that additional loci require additional primer pairs and the theoretical number of spurious fragments increase with the addition of each new pair. This is due to the annealing of primers to non-specific sequences and the phenomenon restricts the multiplexing capability of a traditional PCR to only 10-20 fragments[71,72]. Also, a multiplex PCR requires the amplified products to be of similar length to prevent amplification bias. Approaches using chimeric primers, i.e. locus-specific primers fused with general tag-sequences, and dual amplification steps increase the multiplexing capability to approximately 100-plex[73,74]. By highly optimized primer design and selection of loci by computational aid, multiplexing at the level of 1,000 loci has been reported but with a significantly lower flexibility[75].

IMPROVING PERFORMANCE BY LIGATION

Addition of a DNA ligase to the molecular amplification toolbox gives rise to several new possibilities and has shown to be very useful in different approaches to multiplex amplification.

The oligonucleotide-ligation assay, OLA, pioneered the field by presenting the possibility of covalently joining two adjacently annealed oligonucleotides by ligation[76]. Labeling one of the oligonucleotides with biotin and the other with a fluorophore, ligated products can be captured and thereby indicating the presence of a gene or allele. First carried out on a PCR-amplified product, this principle of oligonucleotide ligation was later adapted to the Ligation Chain Reaction (LCR). The method, with probes complementary to both strands, utilizes only a ligase as catalytic enzyme for an exponential amplification of the locus[77].

The principle described above, by allowing probes to identify a correct sequence by ligation, followed by an amplification of the merged probes rather than the genomic sequence itself, has been further developed in several ways. Joining the two oligonucleotide probes by a sequence acting as a linker, i.e. forming one probe that upon ligation becomes circularized, is used in the Padlock assay[78,79]. The circularized nature prevents the probes from being degraded by exonucleases, hence facilitating degradation of unreacted probes and the removal of genomic DNA prior to amplification. By introducing PCR-motifs in the linker sequence, multiplex amplification at a high level can be accomplished. By adjusting the position of the oligonucleotides in respect to the sequence of detection, a gap fill of one or more nucleotides can be performed prior to ligation[80,81]. Regardless of approach, the methods require knowledge of the target sequence.

A method that does not require prior knowledge, and thereby can be regarded as unspecific, is the ligation mediated PCR (LMP, also called adaptor ligation). Initially, the use of one or more restriction enzymes will digest the genome. A double stranded adaptor oligonucleotide is then ligated to the fragments, thus enabling single-primed amplification as well as a reduction in complexity[82]. The major drawback of this approach is that regions too short or too long will not give rise to products and the method would categorize as semi-whole genome amplification. Combining this approach with the addition of locus-specific chimeric primers, carrying an additional, general PCR-tag has been demonstrated for increased specificity[83]. Combining the features of initial degradation with restriction enzymes followed by ligation and circularization have been used in the form of Selector probes, capable of isolating larger genomic fragments[84,85]. Selectors are oligonucleotide duplex constructs with single-stranded target-complementary end-sequences that are linked by a general sequence motif, hence forming circular constructs harboring the target sequence. The circular structure permits exonucleolysis for degradation of linear DNA and with general PCR motifs in the linker sequence; multiplex amplification with a general primer pair is possible[86].

A PROMINENT POLYMERASE

The DNA polymerase from the bacteriophage φ29 can synthesize long DNA chains without dissociating from the template. An additional prominent feature of the enzyme is effective strand displacement activity[87,88]. Considering the characteristics, it is very well suited for copying circular DNA fragments, using one primer, delivering a long DNA strand with multiple copies of the circle sequence, known as rolling circle amplification or RCA[89]. Circular constructs for multiplex amplification such as Padlocks and Selectors, can therefore be amplified using this approach but traditional RCA amplification is linear and not exponential as PCR. As a result, there is hyperbranched RCA (HRCA)[90], generating a network structure by introducing a complementary primer to the linear sequence initially copied. Another approach is the Circle-to-Circle amplification (C2CA), which gives 100-times higher concentration of amplified product than PCR[91].

The principle of HRCA has also been adapted to the amplification of linear, genomic DNA, i.e. whole genome amplification[92]. A multibranched amplification rendering a network of DNA structures from linear DNA is possible by the addition of random hexamers. This approach is termed multiple displacement amplification, MDA, and has higher yield in total amount of DNA formed than the PCR-based WGA methods described above[93]. MDA performs well when scoring single base variations and shows high-concordance rates[94,95]. In order to avoid allele dropout, high-levels of starting material is recommended[18].

However, the original sequence representation might be affected due to differences in primer efficiency and the variation in GC-content throughout the amplified sequence[92,96] and, furthermore sequence losses have also been reported[95,97]. In general MDA based WGA methods perform well and generate the highest amplification yield, the most complete genome coverage and introduce less bias as compared to the PCR-based WGA-methods mentioned above[98].

PARALLELIZATION BY SPATIAL SEPARATION

A different way of multiplexing is by parallelized solid-phase PCR of single fragments, thus clonally amplifying millions of non-interacting amplicons. The downstream interrogation is facilitated by the separation of features in space, which can be accomplished using a flat surface, a gel matrix or beads – either in wells or oil-separated water drops.

Unlike when digesting a genome with restriction endonucleases, shearing will provide randomly fragmented pieces of more or less similar length. By the addition of general adaptor sequences to the fragments, only one primer pair is required for amplification. Bridge amplification can be performed after immobilization of oligonucleotides complementary to the adaptor sequences on a surface[18,99,100]. Sheared and adaptor-ligated sample DNA fragments can be attached to the solid support and due to the dense lawn of adaptor-complementary sequences on the surface, each will anneal to a nearby primer. A double stranded bridge will form after elongation, and denaturing will free the two strands, both now fixed on the surface. Repeated cycles will form colony-like local clusters, each containing approximately 1,000 copies and with a diameter of about 1 μm. A similar approach can be performed in a polyacrylamide matrix. Thermal cycling of the gel will permit formation of PCR-colonies, or polonies, with 10^8 identical molecules from each single template, since the gel restricts the diffusion of fragments. Covalently fixating each colony to the matrix facilitates downstream interrogation[101]. Spatial separation may also include single fragment immobilization on beads and the use of a PicoTiterPlate, enabling hundreds of thousands of picoliter-scale polymerase chain reactions in parallel[102]. Clonal amplification may also be performed in an oil-aqueous emulsion where each bead is isolated in a water micro-reactor, favoring a 1:1 bead to fragment ratio. Bead clones, not carrying any DNA, or those formed by more than one fragment have to be identified and removed once the emulsion is broken and the beads collected[103,104].

A slightly different approach of amplification by spatial separation is MegaPlex PCR[105] where target-specific primer pairs are immobilized on a solid support. The physical separation of different primer pairs confines the formation of spurious products.

CHAPTER THREE

ON DISCRIMINATING ALLELES

Life on Earth. A biosphere of complex interactions among millions of species in symbiosis, allowed to evolve during 3.8 billion years. Each species with individual representatives carrying novel and unique copies of their code. Humanity constitutes merely a tiny fraction of all the living cells – six billion souls, each composed of trillions of cells, where in most there resides a six billion base-pair genome. The trillions of human cells in each individual are accompanied by at least 10-fold additional microbial cells, an example of our highly symbiotic world.

Even though a seemingly endless task, we have only sequenced a small fraction of the full set of valuable nucleotides. Humanity seems like a good place to start. But considering the cost of simply investigating one human reference genome, there is an obvious desire for shortcuts. Alternatives to sequencing exist, and are delivering valuable information, made possible due to the existence of the reference sequence of the human genome and the fact that we are 99.5% identical[42]. Bearing in mind that the most abundant form of mutation, and therefore also variation, are single base shifts, some bases in the genome can be considered as more informative than others. Selecting and scoring single bases of high importance, rather than aiming at reading an entire sequence could significantly reduce cost while still generating significant data on genomic composition. Such bases may be mutations or SNPs that are directly causative, or well-selected tag SNPs pinpointing haplotype blocks, hence ideally suited when conducting association studies in the quest of finding markers for complex disorders. The trick of targeting single bases could also give information on CNVs, methylation patterns and answer questions regarding LOH, since each cancer cell might have its own genetic signature. In short, once a variation and its surrounding sequence is known, the information can be utilized as a surrogate to sequencing by employing effective scoring methods. The principles rely on one or more allele-specific oligonucleotides (ASO), sometimes assisted by one or more enzymes, and depending on assay design, the interrogation may be parallelized – multiplexed.

Just as in the 1840s, there is a new gold rush at the coast of the Pacific Ocean. The global market for genotyping is soon expected to hit $1 billion and in spite the fact that genotyping development is conducted in research facilities throughout the world, the financial resources of two California-based contenders are constantly extending the frontiers of genotyping in terms of probe density and the number of polymorphisms interrogated simultaneously. For Santa Clara based Affymetrix[106] and San Diego located Illumina[18], arrays with an ever-escalating density of SNP-targeting probes are presented, offering cost-effective approaches for whole genome association studies. Today, their flagships are targeting more than 1 million polymorphisms on one single array.

THE CONCEPTS OF SINGLE BASE INTERROGATION

As in the protocols for enzyme mediated amplifications, the topic of the previous chapter, the complementarity- and hybridization properties of nucleic acids can also be used to target variable regions and distinguish between single-base variations.

The most direct approach when discriminating alleles is exploiting the difference in complementarity between a template (an amplified target sequence) and a matched- or mismatched allele-specific oligonucleotide probe. The single, allele-specific base difference between the probes alters their thermal properties and ability to hybridize to the template[107,108]. The allele-specific oligonucleotide hybridization assay (ASOH or simply ASH) can be implemented by allowing a labeled sample to

hybridize to an oligonucleotide array with matched and mismatched probes for all loci of interest and, following staining, measuring the intensities. The method is hard to optimize in a multiplex fashion since thermal properties are dependent on probe sequence and may differ substantially between probes targeting different loci[108-110]. Design of multiple probes for each locus has improved the assay performance[82].

Increased specificity and distinction between alleles is achieved by using enzymes. In the single nucleotide primer extension assay, also known as single-base extension (SBE) or minisequencing, a locus-specific primer with the 3'-end adjacent to the variation is annealed to the template[108]. A single base extension with labeled nucleotides will determine the identity of the variable base, thereby creating allele discrimination in the polymerase extension step. One reaction is sufficient when different dye labels are used on the four different nucleotides and spatial separation using microarrays permits high levels of multiplexing. The method demonstrates very good discrimination and robustness, 10-fold higher than for ASOH described above[109].

Enzymatic discrimination can also be achieved with the allele-specific extension assay (ASE)[111] or allele-specific ligation[76,112]. The ASE method requires one primer for each of the possible alleles, since the oligonucleotides are designed to match their respective allele variant. Hence, if the allele is present in the sample, a perfect match at the 3'-end base will allow extension, either with labeled nucleotides similar to minisequencing, or as part of a PCR. There are examples of situations where the allele discrimination by polymerase is not complete[111,113,114]. Methods tackling this problem have been suggested utilizing the difference in kinetics between the match- and mismatch primers. If an extension with a mismatched primer does take place, it is slightly delayed as compared to a matched primer extension, and the use of enzymes such as apyrase (AMASE) or proteinase (PrASE) improves assay performance and increases the specificity[115-117]. In the case of allele-specific ligation, there is a need for a secondary primer to which the allele-specific primer can be ligated[76,78].

DISCRIMINATING WHILE MULTIPLEXING

The methods of interrogation described above perform well, but are bottlenecked by the hampered multiplexing capabilities of the often required PCR amplification. Assays circumventing this problem have inverted the work flow by first utilizing specific probes for either a single-base extension (SBE) or an allele-specific extension (ASE) followed by the evolved OLA-principles[78] of ligation and amplification with a general PCR-primer pair. The multiplexing capacity could thereby be improved while retaining specific amplification, a reduction in complexity and a flexibility in choosing loci of interest.

The Molecular Inversion Probes (MIP) assay provided by Affymetrix combines the principle of circularized oligonucleotide ligation assay (Padlock probes) with a single-base extension step[80]. One probe is designed for each locus and is carrying, besides locus-specific sites, general PCR-primer sites, a unique address tag and two cleavage sites. Four SBE reactions are performed, one for each of the four nucleotides, and when a complementary allele is present, a polymerase-catalyzed gap-fill is carried out followed by ligation, rendering a circularized probe. Exonuclease treatment removes all linear DNA, including non-reacted probes, genomic DNA and any probes that might have been wrongfully joined, thus not circular. Utilizing one of the cleavage sites, the probes can be released from the genome and "inverted". The inverted and now linear probes can be amplified by two general PCR-primers, where one is labeled with a fluorophore. Prior to hybridization, using the final cleavage site, the address tag with fluorophore is released from the rest of the probe, minimizing the risk for unspecific hybridizations and thus lowering background signals. Since the vast majority of SNPs are biallelic, the two reactions that are not supposed to give a signal are used to assess background and signal-to-noise ratio. The arrays are manufactured using photolithographic-based techniques and the method allows multiplexing levels of up to 20,000 SNPs[106,118].

On the other hand, the Golden Gate assay by Illumina is a combination of the allele-specific primer extension and the oligonucleotide-ligation assay[81]. For biallelic SNPs, each locus requires three primers. Two primers are allele-specific and designed accordingly, each matching one of the two possible alleles at the 3'-end. Each of these two oligonucleotides also carries a general amplification tag at the 5'-end. There is also a locus-specific primer designed to anneal in close proximity to the SNP. This primer has a unique address tag and a general amplification tag at the 3'-end. Therefore, the total number of primers for the assay is 3n+3 where n is the number of targeted loci. Genomic DNA is attached to a solid support, primers are annealed and allele-specific extensions are performed followed by ligation. Amplification is then carried out with three general amplification primers, where two of them (each with a different fluorophore) target the regions on the allele-specific primers, thus distinguishing the alleles of each locus. The address tags will direct each amplicon to a particular bead type in a randomly ordered and decoded bead array and the signal intensities from the different fluorophores reveal the genotype for a particular locus[81,119]. Multiplexing has been carried out at 1,152- and 1,536-plex levels.

Both methods provide cost-effective and flexible genotyping of polymorphisms at high multiplexing levels offering a reduction in complexity. The conversion rate may vary since the flexibility in probe design for each locus is limited. The call rate and accuracy are in general high, but the methods require several hundred nanograms of genomic DNA.

PLATFORMS ENABLING WHOLE GENOME GENOTYPING

Evolution in the field of genotyping has recently allowed scoring of millions of single nucleotide markers on a single array. In achieving these high multiplexing levels, the methods are relying on different combinations of amplification and interrogation principles.

The Affymetrix Genome-Wide Human SNP GeneChip 6.0 has probes targeting 1.8 million genetic markers – 900k SNPs and just as many CNV loci. Amplification is performed with WGSA (Whole genome sampling analysis) and depending on FSP (fragment selection by PCR)[82]. 250-500 ng of genomic starting material is digested with a combination of restriction enzymes. Adaptor ligation to the generated fragments, followed by size selective PCR, enable a simultaneous amplification and approximately a 50-fold reduction in complexity[120]. Using the restriction enzymes XbaI and HindIII in a previously reported 100k assay, 300 Mb of the genome was represented[121].

The interrogation is performed with allele-specific hybridization (ASH). The GeneChip arrays are manufactured by a combination of photolithography and combinatorial chemistry and contain oligonucleotide probes of 25 bases in length. Only probes targeting SNPs that are anticipated to perform well in the assay, i.e. in a region amplified and without repetitive sequences flanking the SNP, are synthesized on the array. To reduce the problems associated with ASH, each polymorphism is targeted with 40-56 different probes by tiling, meaning that there are several slightly shifted probes targeting both the sense and antisense strand of the template[82]. An algorithm is then used for pattern recognition and genotype assignment. Despite this redundancy, there can be a problem distinguishing between heterozygous and homozygous SNP genotypes for a large fraction of the targeted SNPs[108]. Genotyping of large regions using GeneChip arrays has also been performed after Long-Range PCR, but cannot be categorized as multiplex amplification[122].

The Illumina Infinium assays have a different approach[18]. Starting from 250-750 ng of genomic input, whole-genome amplification is performed with multiple displacement amplification (MDA), thus amplifying without lowering complexity[17]. The product hybridized to the array is 50,000 more complex than with the Golden Gate method but performs at a 95% conversion rate[123]. The Infinium I assay, relying on allele-specific primer extension, and the latest Infinium II with single-base extension are both carried out on the randomly ordered BeadChip arrays using long 75-mer probes – 50 locus-specific bases for increased specificity and 25 for decoding purposes[119].

Although with features such as a single color on a single array, the cost of designing two probes for each locus and the sometimes unspecific extensions have led to the Infinium II assay with a dual-color, multi-layer labeling sandwich assay based on single-base extension. Designing for the four most common base-substitutions in the genome (A/G, A/C, C/T, G/T), the two-color detection scheme of Infinium II can detect 83.5% of all SNPs in one assay. It shows high call rate as well as accuracy and reproducibility[17]. The latest version, the Human1M BeadChip, has approximately 1 million markers for SNPs and CNVs in the human genome[18].

Although only representing 0.03% of the genome, the 1 million markers are informative and provide the means of locating significant regions in association studies prior to fine mapping with more cost effective methods such as MIP or the Golden Gate assay.

CHAPTER FOUR

ON PARALLELIZED SEQUENCING

November 2019. The City of Los Angeles. Harrison Ford is tracking down genetically manufactured humans, known as replicants, to the futuristic and melancholic tunes of Vangelis. The plot of 1982 cult classic movie Blade Runner – directed by Ridley Scott. 2019 lies far in the future but biotech- and sequencing technologies are currently redefining the genetical-, medical-, biological-, and biotechnological sciences. In a decade, will there be genetic designers such as J.F. Sebastian creating genetically engineered toys and artificial owls? At the current pace of development nothing is sure – and then everything is possible. In a not too distant future, the first artificial bacteria will be presented, a goal achievable due to progress in the field of synthetic biology. Knowledge is essential for the ability to create – and the advancements in sequencing are turning many dreams into reality.

Sanger sequencing has been one of the most influential innovations in biotech since it was first presented in 1977. A little more than 20 years later, a bioluminescence sequencing-by-synthesis approach saw the light of day[124]. Today, Pyrosequencing has evolved at 454 Life Sciences, generating a hundred million bases of raw sequence in just a few hours[103]. This throughput, although heavily refined and improved during the years, is something Sanger sequencing in its current form cannot easily match. However, during the last year, Illumina and Applied Biosystems (ABI) have introduced sequencing systems offering even higher throughput than the systems provided by 454, although there are other factors to sequencing than throughput worth considering.

These novel methods all rely on parallel, cyclic interrogation of sequences from spatially separated clonal amplicons. Although with shorter read-lengths and a slower sequence extraction from individual features as compared to the Sanger method, the parallelized process offers a much higher total throughput and reduces cost significantly by generating thousands of bases per second. By shearing the template and parallel sequencing of single fragments, over-sampling may provide improved coverage and the possibility of stitching together the original sequence while increasing total accuracy. Avoiding bacterial cloning may also provide data from sequences that cannot be contained in bacteria[125].

The systems from 454, Illumina and ABI are spawning a revolution. Two examples are *de novo* sequencing of microbial genomes and metagenomic studies of rainforests and oceans. Both these endeavors are generating novel genes that might prove beneficial for synthetic biology by providing new tools to the art of genetic engineering. As a result, novel enzymatic pathways in designed genomes[5] may be of aid in the production of petrol substitutes or provide systems for mopping up excessive carbon dioxide in the atmosphere[4,126-128].

The most captivating application in a nearer future is, however, probably the possibility of resequencing larger and larger fractions of human genomes at an ever-decreasing cost. The reference genome allows us to superimpose sequence reads permitting very cost-effective comparisons. Knowledge of an entire sequence rather than just single bases can give many answers, since there are more variations to the genome than just SNPs – for instance insertions, deletions, CNVs and large chromosomal rearrangements such as inversions etc. Sequencing could also identify rare mutations or provide efficient interrogation of highly variable regions such as genes for the human leukocyte antigen system (HLA), Cytochrome P450 or randomized B- and T-cell receptors. It could also be used for discovering causative variants in linkage disequilibrium with haplotype blocks[129]. With an exponential increase in sequencing efficiency and reduction in cost we are closing in on the $1,000 genome. These ultra-low-cost sequencing (ULCS) methods may still be a few years away, but when that time comes sequencing really can become a part of routine health care and further our understanding of aging and cancer by tracking mutations in pathways and at essential checkpoints.

Already today the high-throughput methods mentioned above, and further described below, are expanding our knowledge, also in the related fields of transcriptome- and proteome research. Gene expression analysis with whole-transcriptome sequencing is possible and furthermore, in proteome research, by sequencing DNA extracted by antibodies targeting DNA-binding proteins (ChIP-Seq), transcription factor binding sites and chromatin modifications can be investigated[130,131].

TERMINATING CHAINS

Since 1977, a total nucleic acid polymer of approximately 10^{11} bases has been determined with the Sanger sequencing method[129]. The technique relies on chain termination. By halting the elongation with a labeled, and thereby identifiable, dideoxyribonucleotide triphosphate (ddNTP), the length of the fragment can be utilized for interrogating the base identity of the terminating base[132]. In its current form, fluorescently labeled ddNTPs[133,134] are mixed with regular, non-labeled, non-terminating nucleotides in a cycle sequencing reaction[135,136] rendering elongation stops at all positions in the template. Capillary electrophoresis can then be applied for separating sequences by length and providing subsequent interrogation of the terminating base.

Initially at a high cost, refinements and automation have improved cost-effectiveness significantly. In 1985, $10 resulted in one single base, while the same amount of money rendered 10,000 bases 20 years later[21,129]. Current instruments deliver read-lengths of up to 1,000 bases, high raw-accuracy and allow for 384 samples to be sequenced in parallel generating 24 bases per instrument second. Projects of multiplexing and miniaturization in order to reduce reagent volumes, lower consumable costs and increase throughput are being pursued[137,138].

HYBRIDIZATION TO TILING ARRAYS

The concept of allele-specific hybridization (ASH), described in the previous chapter, can be adapted for resequencing purposes by expanding each probe set, targeting a specific position in the genome, to include interrogation of each of the four possible nucleotides[139]. A tiling array can be fabricated with probe sets targeting each position in the reference genome. Read-length is given by the probe length (often 25 bp) and base calling is performed by examining the signal intensities for the different probes of each set. Accuracy is an issue and is dependent on the ability of the assay to discriminate between exact matches and those with a single base difference. Performance may vary significantly due to different base compositions (different thermal annealing properties) of different regions, resulting in problems with false positives as well as with large inaccessible regions composed of repetitive sequence stretches[140,141].

The throughput is an obvious benefit, since all bases are interrogated simultaneously and the concept has been applied to resequencing the human chromosome 21 and HIV[140,141]. By representing all possible sequences for a given probe length, *de novo* sequencing can be performed and overlapping sequences used for sequence assembly[142].

FLASHING BEADS

The Genome Sequencer FLX by 454 Life Sciences[143] and Roche depends on an emulsion PCR followed by parallel and individual Pyrosequencing of the clonally amplified beads in a PicoTiterPlate. Amplified beads are enriched and distributed on the plate, where a well (44 μm in diameter) allows fixation of one bead (28 μm in diameter)[103]. Out of the 1.6 million wells, not all will contain a bead and not all of those

that do will give a useful sequence.

Pyrosequencing is a sequencing-by-synthesis method where a successful nucleotide incorporation event is detected as emitted photons[144]. Since the single-stranded DNA fragments on the beads have been amplified with general tags, a general primer is annealed permitting an elongation towards the bead. The emission of photons upon incorporation depends on a series of enzymatic steps. Incorporation of a nucleotide by a polymerase releases a diphosphate group (PPi), which catalyzed by ATP sulphurylase forms adenosine triphosphate (ATP) by the use of adenosine phosphosulphate (APS). Finally, the enzyme luciferase (together with D-luciferin and oxygen) can use the newly formed ATP to emit light. Another enzyme, apyrase, is used for degradation of unincorporated dNTPs as well as to stop the reaction by degrading ATP[144].

In the 454 system, the Pyrosequencing technology is adapted as follows. The enzymes luciferase and ATP sulphurylase are immobilized on smaller beads surrounding the larger amplicon carrying beads. All other reagents are supplied through a flow allowing reagents to diffuse to the templates in the PicoTiterPlate. Polymerase and one exclusive dNTP per cycle generate one or more incorporation events and the emitted light is proportional to the number of incorporated nucleotides. Photons are detected by a CCD-camera and after each round, apyrase is flowed through in order to degrade excess nucleotides. The washing procedure for the removal of by-products permits read-lengths of up to 250 bp. This limitation is due to negative frame shifts (incorporation of nucleotides in each cycle is not 100% complete) and positive frame shifts (the population of nucleotides that is not fully degraded by the apyrase and can therefore be incorporated after the next nucleotide) that eventually will generate high levels of noise. Approximately 400,000 wells will give one unique sequence of 250 bp, on average generating 100 million bases (Mb) in one single run. Whole-genome sequencing has been performed on bacterial genomes in single runs[103].

A REVERSIBLE TERMINATION

The Illumina 1G Genome Analyzer is relying on clonal bridge amplification on a flow cell surface generating 10 million single molecule clusters per square centimeter as described above. Sequencing is then carried out with fluorescently labeled nucleotides that are also reversible terminators. One base is incorporated and interrogated at a time since further elongation of the chain is prevented[125]. When all colonies are scanned at the end of a cycle and the base determined for each colony, the terminating fluorophores are cleaved off, which allows another round of nucleotide incorporation. The presence of and competition among all four nucleotides is claimed to reduce the chance of misincorporation. Incomplete incorporation of nucleotides (efficiency is not 100%) and insufficient removal of reverse terminators or fluorophores may be the explanation for the relatively short read-length of 35 bp. Although shorter read-lengths than the 454 system, the throughput is much higher and as of October 2007, 1.3 Gb are generated in each run, which takes approximately 3 days. The use of paired-end libraries will generate 2.6 Gb in a single run. The raw accuracy is said to be at 98.5 % and the consensus (3x coverage) at 99.99%. The cost per base is approximately 1% of the cost for Sanger sequencing[18,30].

LIGATING DEGENERATED PROBES

Strategies for sequencing-by-ligation have been presented in the form of Massively Parallel Signature Sequencing (MPSS) and Polony sequencing[104,145].

MPSS was demonstrated as signature sequencing of expression libraries of *in vitro* cloned microbeads, i.e. beads carrying multiple copies of a single DNA sequence[146]. Signature sequencing was carried out by restriction enzyme mediated exposure of four nucleotides in each cycle followed by ligation

of an interrogator probe. This process was repeated for 4-5 cycles, i.e. querying 16-20 bases in total. An overhang of four bases would require 256 different complementary probes and just as many fluorophores for immediate recognition. Instead, the use of 16 (4*4) probes, each with a unique decoder binding site, has enabled single dye detection.

Resequencing of a bacterial genome was used to demonstrate the Polony sequencing method[104]. A mate-paired library was clonally amplified with an emulsion PCR on 1 μm beads and subsequently immobilized in a polyacrylamide gel. Each DNA-carrying bead (polony) represented two 17-18 bp genomic sequences flanked by different universal sequences. Due to the nature of the mate-pair construction, the two genomic sequences were separated by approximately 1 kb in the genome. Sequencing by ligation could then be performed using degenerate nonamers, where each known nucleotide was associated with one of four fluorophores. By using four different anchor primers, degenerate sequencing by ligation could be performed from each end of the tags. 7 bases could be obtained when sequencing in the 5' to 3' direction and 6 bases from 3' to 5'. Ligated primers were removed after each round rendering information of 26 bp from each amplicon in a pattern of: 7 bp, a gap of 4-5 bases, 6 bp, then a gap of approximately 1 kb (mate-paired constructed) and then another 7 bp, a gap of 4-5 bp, followed finally by 6 bp.

These two methods have spawned the development of the commercial SOLiD-system (Sequencing by Oligonucleotide Ligation and Detection) from Applied Biosystems where clonal amplicons on 1 μm beads are generated by an emulsion PCR, either from fragments or mate-paired libraries. The beads are enriched, so that 80% of them will give signals, and attached on a glass surface forming a very high-density random array. Sequencing-by-ligation is performed by ligating 3'-degenerated and 5'-labeled probes to the amplicons and detecting the color. Accuracy is improved by implementing a two-base encoding system that leads to interrogation of each base twice. A sequencing run takes 8-10 days and the output is high, approximately 3 Gb per run given a read-length of 25-35 bases per clonally amplified bead[147].

THE ROAD AHEAD

Resequencing a human genome with the Sanger sequencing method would today cost approximately $10 million[42,125] while the 454 system enables a 10-fold reduction in cost and about 20-fold reduction in time[23]. Illumina claims to be able to sequence a human genome with the 1G Analyzer for $100,000[30]. Although this is a significant reduction in cost it is still too expensive for routinely sequencing human genomes at a larger scale.

Realization of the $1,000 genome requires novel approaches. While it is an ongoing process, it is very hard to overview the current progress since research reported today began many years ago and the most recent ideas are in the majority of cases confidential. But clearly, there is a lot of activity within this field. Visigen[148], who recently joined as a contestant in the Archon X Prize for genomics, is working on single-molecule sequencing with engineered polymerases and nucleotides rendering an expected throughput of 1 million bases per instrument second. Complete Genomics[149] is combining high-density DNA nanoarrays with ligation- and hybridization chemistry in the quest to design a platform that will have significantly higher throughput than today's systems. Intelligent Biosystems[150] are pursuing an array-based sequencing-by-synthesis approach[151] similar to the Illumina 1G system. Examples of other contenders are Pacific Biosciences[152] and Helicos[66], where the latter has an instrument claimed to produce over 600 Mb per day[66].

CHAPTER FIVE

DESIGNING & MINING

Our world is changing and it is changing fast. There was a time – not that long ago – when very few people were familiar with the concept of hyperlinks. In 1994, Netscape was incorporated and the awareness of the Web ignited a revolution. The man behind Netscape was a bit ahead of his time. He built a custom made yacht and equipped the ship with thousands of sensors, 60 km of fiberoptic cable and 24 Silicon Graphics workstations. Then, using a satellite link, the Hyperion could be controlled via the Internet from his living room in the heart of Silicon Valley, or from anywhere else on the planet allowing a connection to the Net[153].

Today, the world has adopted his visions. From less than 3000 web servers in 1994 to more than 100 million today, hosting billions of webpages[154] there are few things in society not influenced by the Internet. A laptop and a wireless network puts the world at our hands wherever we are, and the development and progress are escalating since we are constantly generating new tools that are helping technology evolve at an even higher rate.

Via the laptop we can even be conducting cancer research, whether be it from a hotel room or a beach. The exponential growth is not exclusive to the fields of computing and communication, it is also affecting all disciplines depending on them, i.e. sciences of information and knowledge, of which modern biology is certainly one[155]. The emerging field of bioinformatics for storage, manipulation, interpretation and visualization of the vast amount of data generated from the investigation of complex biological systems is essential to all medical and biological research of today. This interdisciplinary field, a mixture of computer science, statistics and biology has led to a myriad of databases, tools and web-server applications[156] with portals enabling easy navigation and access to such means. An example is the Bioinformatics Link Directory[157] provided by Nucleic Acids Research with links to 1,200 various bioinformatic- and life science tools[158].

The amount of biological data has grown tremendously after the leap from genetics to genomics. One of the first established databases, GenBank[159], which contains all publicly available nucleotide sequences, is growing exponentially and doubling every 18 months[160]. Miniaturization, parallelization and automation implemented in the form of increasing probe densities on arrays, novel sequencing systems and more samples will increase the demand for data storage and interpretation even further, a good example being the cyberinfrastructure[161] of the Global Ocean Sampling metagenomics project[1,2].

Bioinformatics and information technology are essential parts of the whole scientific pipeline, from upstream design of probes targeting genetic sequences, via the Laboratory Information and Management System (LIMS) tracking samples in the wet-lab process, to the conversion of measurements into useful data, mining, interpretation, visualization and the final storage of the results. Additionally, information technology is contributing to the progress of research by enabling efficient distribution of novel results to the research community[162].

DEPOSITORIES FOR GENES, GENOMES AND GENETIC VARIATIONS

It is not only the existing databases that are growing fast with the constant addition of sequence reads and novel genomes, new biological databases continue to pop up as new findings and technologies for high-throughput biology emerge. The Nucleic Acids Research update for 2007 lists 968 databases, 110 more than the previous year[163].

There are a number of interesting resources for finding information regarding genetic and genomic

data. The National Center for Biotechnology Information[164] is providing web-based tools and applications for data mining as well as maintaining a number of databases[162]. Besides the previously mentioned GenBank, an example is the NCBI Taxonomy database, growing at a rate of 2,900 new taxa a month, in total containing information about more than 240,000 named organisms[162]. The complete genomic sequence of 370 microbial genomes can be accessed via Entrez Genome, where 120 have been added over the past year and 20 belong to higher eukaryotic species, including homo sapiens[162]. Recently, two human individual genomes have been added, the ones of Craig Venter and James Watson.

Databases for genetic variation at NCBI include the dbMHC, which is a registry for variants of the major histocompatibility complex (MHC) also known as HLA (human leukocyte antigen). The database stores information about HLA alleles and genotyping kits which are of significant importance when predicting success in organ transplants and the susceptibility to infectious diseases[162]. Another important database of genetic variation is dbSNP which contains information on single nucleotide polymorphisms as well as insertions and deletions[165]. There are approximately 12 million human SNPs in the database and another 22 million from a variety of other species. During 2006, 14 million new variants were added[162]. For each variation, the database stores information regarding position in the genome, flanking sequences, validation status, population-specific allele frequencies and haplotype data. Functional variants can be matched to OMIM records – the Online Mendelian Inheritance in Man, where information regarding phenotype can be found[166]. Another important database for variations is the Human Cytochrome P450 (CYP) Allele Nomenclature Committee[167], which stores information regarding variations in the cytochrome genes with links to the dbSNP database and OMIM. Two additional examples are the SNP500Cancer database[168], which lists SNPs of importance to molecular epidemiology studies in cancer and the Human Obesity Gene Map[169], which lists markers of importance to obesity research[170].

Other important sites are the Ensembl Genome Browser[40] and the Genome Browser at University of California, Santa Cruz[171]. These are both exchanging data with NCBI on a regular basis and provide tools for efficient mining and extraction of data, one example being the Ensembl BioMart.

ALGORITHMS AND SOFTWARE TOOLS

Sequence analysis is an essential part of bioinformatic work, for instance when mapping reads to a reference genome, in a *de novo* assembly or when designing PCR primers. Whatever one's reason for sequence investigation might be there are a number of tools to do the job, many running as stand-alone applications or available as web-services. The most widely known bioinformatic algorithm is probably the Basic Local Alignment Search Tool (BLAST) which allows for local pairwise sequence alignments and comparisons[172]. BLAST is a heuristic method and faster than the optimal-alignment seeking Smith-Waterman algorithm (based on dynamic programming) but in the majority of cases they give similar results[173]. Alignments could also be performed globally, i.e. aligning sequences from beginning to end, and in multiple sequence alignments more than two sequences are compared[174]. An example of a bioinformatic toolbox is the EMBOSS package[175], an open-source collection of 150 applications including software for sequence alignments, nucleotide pattern analysis (for identification of CpG islands, repetitive regions and codon usage) and features for visualization of data.

Besides applications for comparing and analyzing data, the information needs to be stored and managed properly. Therefore there are standardized file formats for sequence data, for instance the FASTA-format containing a single description line (header), which is followed by the sequence. Genetic sequence data can also be represented in many other file formats, for instance XML[176] (eXtensible Markup Language). Larger sets of data may require a database for efficient management and there is a number of relational database systems available, for instance mySQL[177], Oracle[178] or PostgreSQL[179].

There are a few scripting or programming languages, providing open-source bioinformatic packages and APIs (Application Programming Interfaces) for stitching together custom made

bioinformatic tools for efficient automation. Four options are Java[180,181], Perl[182,183], Python[184,185] and Ruby[186,187]. The bio-packages include classes and modules for interacting with command-line based software, for instance BLAST or tools included in the EMBOSS package, for efficient querying, retrieving and parsing of output data. Client-server database support to the relational databases mentioned above is provided by most and some provide classes for accessing public databases such as Ensembl. The languages each have their own strengths and weaknesses, but Perl (Practical extraction and report language) is probably the most widely used due to its exceptional efficiency in handling text strings and the highly evolved BioPerl[182] package.

STATISTICS & MINING

Another aspect of bioinformatics is biostatistics. It is as important in validating base calls in sequence reads or in assessing genotype calls from array data[188,189] as in the final analysis of the many data points generated from expression- or SNP-typing arrays. Once the genotypes are established, statistical methods are required to mine the data from Genome-Wide Association studies (GWA) including thousands of patients and just as many controls[190] in the quest of locating statistically significant genetic markers.

Spreadsheets, as provided by Microsoft Excel, are not appropriate for handling large amounts of data efficiently. Statistical open-source software tools like the R-environment[191] are well suited for data mining and visualization in these cases. The package enables efficient calculation on arrays and matrices and provides an effective programming language with pre-written modules for statistical analysis. Many modules are adapted to biological assays, for instance those provided by the Bioconductor framework[192,193]. Other features of R and Bioconductor are access to Ensembl's BioMart, an integrated web-server and an extensive visualization environment[156].

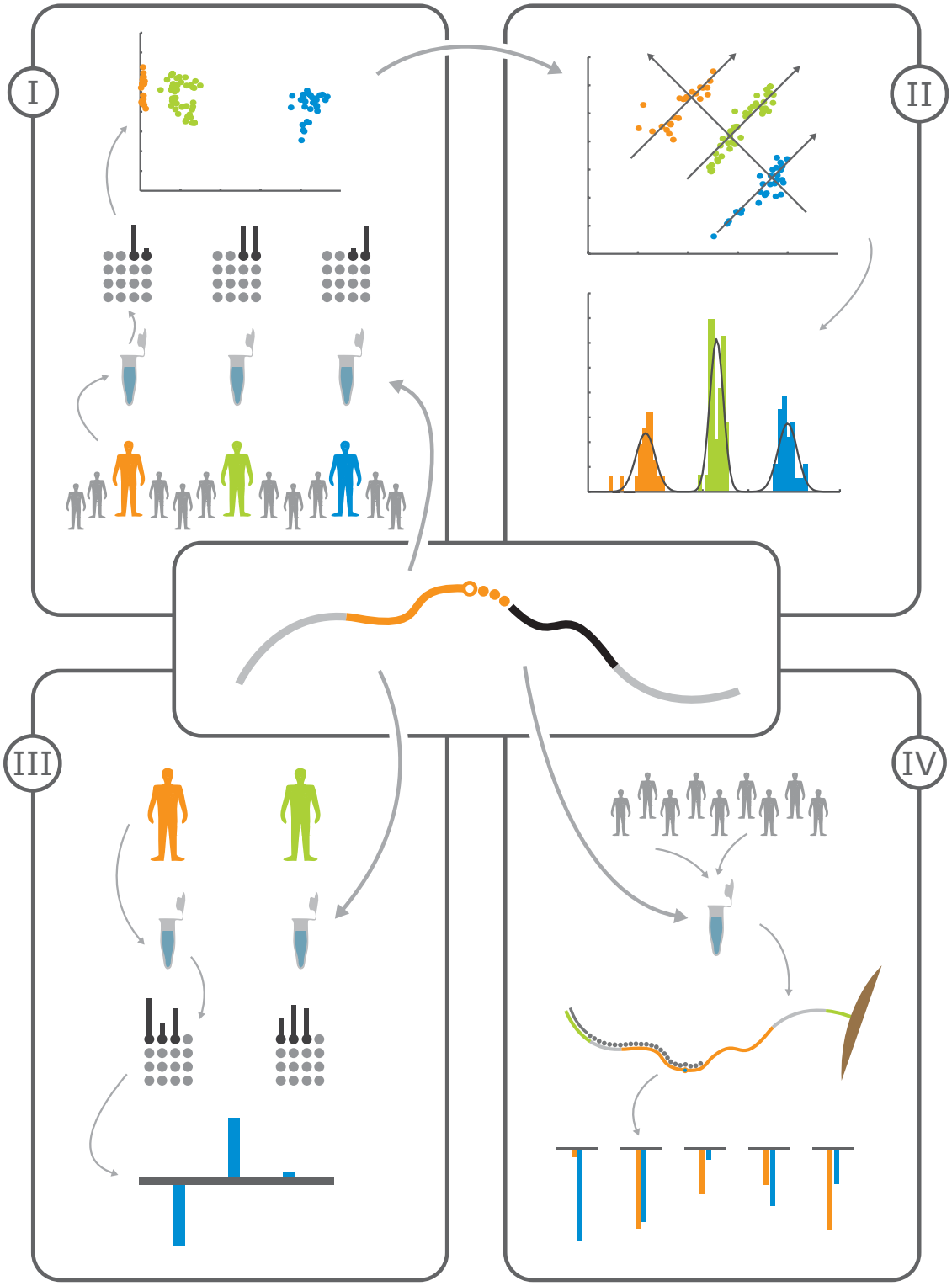
INVESTIGATIONS

CHAPTER SIX

INVESTIGATIONS

In contrast to the latest systems for whole-genome sequencing and genotyping, there is a need for versatile interrogation of moderate sets of selected genetic signatures in a large number of samples for cost effective diagnostic- or fine-mapping purposes. The aim of the work presented in this thesis was to facilitate parallel interrogation of such moderate nucleic acid signature sets by developing assays and algorithms with emphasis on improving flexibility and automation. An illustration of how the four papers, forming this thesis, are interconnected can be viewed on the following page.

A primer-based enzymatic assay, denoted Trinucleotide Threading (TnT), was developed to enable specific multiplex amplification of targeted loci from minute amounts of template while also reducing sample complexity. Analogous to threads in computer science, which represent parallel executions, the DNA threads formed in the Trinucleotide Threading assay provide simultaneous amplification of specific genomic regions for parallel downstream analysis. In addition to the biochemical technique, an algorithm facilitating assay design was implemented, resulting in a user-friendly software tool. The amplification method was applied to genotyping, expression profiling and digital allelotyping. As outlined in Paper I, primers were designed by the software to target 75 polymorphic loci in cancer related genes in a set of 96 individuals as a proof-of-concept. The objective of Paper II was to formulate a special 3-component mixture model for automated classification and genotype calling. The study was based on the data generated from the 96 individuals genotyped in Paper I. Furthermore, as described in Paper III, the Trinucleotide Threading assay was employed for expression profiling of 18 genes in eight cell lines and the data was compared with real-time PCR and genome-wide cDNA arrays. Finally, the TnT method was adapted to digital interrogation of allele frequencies in large cohorts using pooled genomic DNA and massively paralleled Pyrosequencing using the GS20 system from Roche/454 Life Sciences (Paper IV). A total of 147 threads querying polymorphisms in obesity and cancer related genes were analyzed in a cohort of 462 individuals in a proof-of-principle study.



Interrogation of Nucleic Acids by Parallel Threading (Paper I-IV). **(I)** Trinucleotide Threading for multiplex amplification of 75 SNPs in 96 individuals and subsequent array-based genotyping using PrASE. **(II)** Automated classification and genotype calling of the generated PrASE data by a 3-component mixture model derived from molecular principles and the employment of an EM-algorithm. **(III)** Trinucleotide Threading for multiplex gene expression profiling comparing the expression of 18 genes in eight different cell lines by array-based decoding. **(IV)** Trinucleotide Threading targeting 147 SNPs using a pooled set of genomic DNA from 462 individuals and massively parallelized Pyrosequencing enabling estimation of the allele frequencies in the cohort by counting individual reads.

MULTIPLEX AMPLIFICATION BY TRINUCLEOTIDE THREADING (PAPER I)

There is an ever-increasing interest for multiplex preparation of genetic material to facilitate effective scoring of moderate numbers of polymorphisms in large numbers of samples. Although there are millions of polymorphisms in the human genome[194], some are more informative than others and interrogation of an intermediary set of specifically selected causative or marker signatures is sufficient in many diagnostic- or forensic investigations. Paper I describes the development of the Trinucleotide Threading assay – a strategy for simultaneous preparation and amplification of targeted loci from minute amounts of genomic DNA. Altogether, the method relies on the following steps: (i) selection and filtering of target polymorphisms and software aided design of two probes for each locus; (ii) polymerase-mediated elongation with a set of three nucleotides, filling a gap designed to be between the probes and where the second probe is designed to start with the first occurrence of the fourth nucleotide downstream of the first probe; (iii) covalent joining of the 3'-end of the extended probe with the 5'-end of the second probe; (iv) repeating the extension-ligation steps to achieve a linear amplification; (v) capture of the ligated products (denoted DNA threads) resulting in complete removal of genomic DNA; (vi) amplification of the DNA threads by a pair of universal PCR primers and (vii) genotyping of the amplified fragments.

The Trinucleotide Threading method is based on the use of a reduced set of nucleotides permitting lower amounts of starting material, while keeping the number of spurious amplification products at a minimum. Specificity is achieved by allowing ligation of the trinucleotide-elongated products to a secondary primer (denoted thread-joining primer). This procedure is favoring the formation of specific DNA threads and linear amplification while preventing unspecific ligation and exponential amplification of spurious products.

The thread design for each locus requires that i) an extending probe must anneal just before the polymorphic site, ii) the extension with the three selected nucleotides must span the polymorphic region regardless of allele status and iii) the second thread-joining primer must anneal adjacent to the elongation stop. This must be fulfilled for all targeted loci in a defined reaction mixture of three dNTPs. Hence, the sequence flanking a SNP decides the extension length by the employed three dNTP-set and thus defines the position of the thread-joining primer. Different thread lengths may be achieved for each locus by exploiting the double stranded nature of DNA using either the sense- or antisense strand as template.

There are six classes of substitution polymorphisms K (G/T), M (C/A), Y (C/T), R (G/A), W (A/T) and S (C/G), where the two latter are intracomplementary. Furthermore, since there are four different dNTPs, a total of four different three-nucleotide combinations can be achieved (TCG, ACG, ATG, ATC). Consequently, considering the possibility of thread design on both strands, all four possible three-nucleotide combinations may be used for covering the polymorphisms K, M, Y and R since they are not intracomplementary. The two remaining, W and S, can also be extended on both strands, but only with two of the four possible three-nucleotide combinations. Since S requires dGTP and dCTP and W requires dATP and dTTP, the two possible combinations for each variability type do not overlap. As a consequence, two reactions are required to amplify all six classes of SNPs, but one single three-nucleotide mixture is sufficient to cover either (K, M, Y, R and W) or (K, M, Y, R and S).

To facilitate the assay design, a software tool was implemented in Java[180,181]. Given a FASTA-file of the SNP sequences and a window of pre-defined primer annealing temperatures, extensions are performed *in silico* and primers designed accordingly. For each SNP, the four possible extensions with different three-nucleotide sets are performed on both strands. A scoring system, favoring extensions of 5-7 bases is applied and since amplification of all six classes of SNPs is allowed, different pairs of three-nucleotide combinations are suggested. The rational for favoring extensions between 5 and 7 bases is to increase the overall assay specificity. Each extension is scored and allocated to a reaction mixture in a way to maximize the total score rather than balancing the number of SNPs in each of the two reactions. Universal amplification tags are added to the extension- and thread-joining primers, and the primers are adjusted so that the maximum difference in length between any two DNA threads is no more than 10%.

Threading is then performed *in vitro* using a few nanograms of genetic material in a thermocycler permitting multiple extension- and ligation rounds. Biotinylated extension primers allow for downstream immobilization on streptavidin-coated beads and an automated washing procedure removes genomic DNA and excess primers. Release[195] of the DNA threads from the beads and a final PCR amplification step using one universal primer is then performed allowing genotyping with any method of choice that is capable of multiplex analysis.

As a proof-of-concept, polymorphisms with relatively high minor allele frequency in the Caucasian population were identified and the sequences retrieved from dbSNP. Subsequently, the loci were screened for neighboring polymorphisms (<40 residues) and low-complexity regions. A final check against the human genome using BLAST was performed removing loci indicated to be located in duplicated regions. 88 candidate positions were subjected to *in silico* extensions. However, the software tool was designed to select the top 75 SNPs. The combination rendering the highest score was dCTP, dTTP and dGTP (with 39 SNPs) and dATP, dTTP and dGTP with the remaining 36 SNPs. All six classes of SNPs were represented among the assayed polymorphisms, which originated from 61 different genes distributed amongst all but four chromosomes. To avoid PCR bias from being introduced due to thread length differences, primers were designed to form DNA threads between 104 and 115 residues. In addition, the software designed primers for the subsequent genotyping of the amplified fragments. Different concentrations of genomic DNA from a panel of 10 Swedish individuals were employed to determine the limit of detection and reliable genotype calling. Furthermore, a total of 96 individuals of Caucasian origin were subjected to the threading reactions followed by amplification. Thereafter, genotyping was performed using protease mediated allele-specific extension (PrASE) with detection using in-house produced tag microarrays[117].

Out of the 75 assayed SNPs, 68 rendered partitioned clusters of homozygous and heterozygous genotypes. Three of the seven failed SNPs were considered as completely failed, after investigation by duplex amplification reactions harboring one performing and one failed thread. Therefore, the conversion rate was determined to be 94% (68/72) with a call rate of 98%. The assay was also shown to exhibit good sensitivity with concordant results for the 68 approved SNPs in 10 individuals when starting from 1, 10 or 200 ng of genomic DNA, respectively.

AUTOMATED GENOTYPE CLASSIFICATION BY A FINITE MIXTURE MODEL (PAPER II)

Based on the results from Paper I, a three-component mixture model was derived from molecular principles to automatically classify and call the genotypes from the same dataset. The genotyping procedure by PrASE in Paper I was performed with two allele-specific primers differing at their 3'-ends and with different 5'-end signature sequences, complementary to tags on generic tag arrays[117]. A positive signal obtained from a spot on the array corresponds to the presence of the allele in the sample. However, in a defined set of polymorphisms, different SNPs will exhibit individual patterns due to different magnitudes in signals from the two spots and thus requiring manual inspection of each cluster diagram for calling of the genotypes. This stems from the fact that primers have their own unique characteristics. Hence, an algorithm that allows for automated estimation of the different normal components would facilitate the genotype calling significantly. Approaches for genotype classifications have been reported previously[196-198], but they are based on different genotyping methods. These undertakings are also more or less non-parametric in the sense that no parameter relationships and no distributions are specified which also means that no structure is expected in the data[118,199,200]. In Paper II we demonstrate that such structures exist for the PrASE assay and that a model derived from molecular principles can be incorporated in the statistical analysis. Much of the same type of associated classification method has recently been proposed for Affymetrix SNP data[201] and the Illumina bead array[189] but without explicit motivations.

As mentioned, the signal intensities from the spots for a particular polymorphism are not strictly binary. These differences can be attributed to sequence context (hybridization and extension properties) and the quality of the synthetic primers. Instead of representing data as Allelic Fraction (AF) (Paper I), the genotype clouds are visualized by plotting the signal intensities from the two spots as log(x) vs. log(y). This renders three distinct bands, spatially separated in the 135°-direction, representing the members of homo- and heterozygous genotypes. In addition, modeling of the signal contribution for the PrASE assay on a molecular level relies on the fact that the allele-specific primers differ only at their 3'-ends and that they bind both to the designated target allele and to the second allele. Binding of a primer to its designated allele and the other allele is denoted match and mismatch, respectively. But the difference in efficiency between the primers towards their specific target alleles will accordingly affect the mismatch contribution in a similar relationship and we can introduce a specificity factor Θ which is a quotient describing the relationship between an allele-specific primer's match and mismatch. From this follows that the two homozygous bands should be on equal distance from the heterozygous band in the middle, which in turn is slightly shifted in the 45°-direction. This effect originates from the fact that both alleles are present in a heterozygous sample and thus mismatch contribution is observed from both primers. The shift of the heterozygous cluster is of relevance when only two variants are reported in the sample.

The interest is primarily in variation in the 135°-direction. Thus, classification of the data, based on the model described above, can be performed by viewing the clusters as three univariate normal distributions with mean, variance and mixing proportions unknown. These parameters can be estimated with an Expectation-Maximization (EM) algorithm. The EM algorithm is an iterative method for computing the maximum likelihood (ML) estimates, when standard likelihood maximization is numerically difficult or impossible because of incomplete data, i.e. lack of membership knowledge that characterizes the mixtures[202,203]. Assuming appropriate knowledge of the complete data and adjusting this knowledge in a two-step iterative method, an expectation E-step alternating with a maximization M-step, converges the maximum likelihood.

Out of the 75 assayed SNPs (Paper I) five were excluded due to low signal intensities (SNP 16, 34, 40, 70 and 71). The algorithm was applied to all 70 data sets with different starting values for the unknown parameters depending on the number of mixture components. The algorithm assesses the mean, variance and also determines the probability that a particular data point belongs to a certain component. Each data set has to be investigated for one, two and three components. However, assessing the number of components in finite mixture models is difficult[204]. Therefore, four information criteria were used for indicating the correctness of the assessment. In 65 of the 70 cases, all methods gave concordant results. The distribution was 51 cases with three components, 12 cases with two components and 2 cases with one component. At most, 5 out of 70 SNPs were ambiguous with respect to the component number in the sense that different assessment criteria gave discordant answers. Among these five were the additional two SNPs (52 and 60), which had been classified as failed by manual inspection in Paper I.

TRINUCLEOTIDE THREADING FOR EXPRESSION PROFILING OF SIGNATURE GENE SETS (PAPER III)

In Paper III, the Trinucleotide Threading procedure was adapted for gene expression analysis. Since research has shown that the majority of genes in global expression analysis exhibit only minor fold changes in pathological states, careful selection of a representative diagnostic set of variably expressed genes can suffice for accurate profiling. Interrogation of such intermediate signature gene sets conveying the most relevant information (10-100 genes), has lately been shown to be a powerful predictor in severity, progression and metastasis in different cancer forms as well as the clinical outcome of different disorders [57,205,206].

Methods dominating the field of gene expression, such as genome-wide cDNA and

oligonucleotide microarrays and techniques based on real-time PCR, are not suited for expression profiling of a moderate set of genes in a large number of samples. In recent years, a few bead-based systems for analysis of relevant sets of signature genes have been described. Some of these include the cDNA-mediated, annealing, selection, extension and ligation (DASL)[207] and its RNA based predecessor (RASL)[208] that uses randomly ordered bead arrays[119]. Methods utilizing microsphere-based suspension array technology are the ligation-mediated amplification approach[209] and multiplex branched DNA (bDNA) assay[210]. Still, these methods may have inherent problems relating to specificity.

Paper III describes a proof-of-concept strategy for studying signature gene sets by profiling 18 genes – 15 randomly chosen sample genes and three housekeeping genes in eight different cell lines using TnT. The threads were designed to target gene sequences corresponding to protein epitope signature tags (PrESTs), utilized in the Swedish Human Protein Atlas (HPA) program for antibody-based proteomics[58]. Each PrEST is carefully designed to represent a unique region of a protein and corresponds to approximately 100-150 amino acids. The 8 different cancer cell lines investigated were also employed in the HPA program.

To facilitate the TnT approach, a software tool for primer design was developed. The script identifies suitable 10 to 12 bp gaps, composed of a chosen trinucleotide set, and designs the two TnT probes so as to flank these gaps. Consequently, the output is a list of all possible threading regions and the primers associated with them. Allowing only a short span of approved lengths will minimize the risk of introducing length-bias in the exponential amplification. Improved specificity for the signature profiling is achieved in several steps. First, the use of three nucleotides prevents long extensions and thereby also the formation of spurious ligation products. In addition, since the 10-12 unique elongated bases are required for proper annealing of the detection probes, the unlikely event of unspecific thread formation by misligation does not lead to a reduction of specificity, thus retaining a high level of discrimination.

The assay included parallel and linear DNA thread formation targeting PrEST regions in cDNA formed by reverse transcription primed either by oligo dT or locus-specific primers. Both strategies were investigated in all cell lines. After exponential amplification, detection was performed utilizing the specific extension primers followed by hybridization to generic tag arrays.

A normalization/fold change scheme of the data, with respect to all three housekeeping genes, was devised and the expression data generated by the TnT-approach was validated by real-time PCR over all 15 genes. Furthermore, a comparison with genome-wide cDNA arrays was performed. An initial examination of the comparison revealed that one gene exhibited considerable fold change discrepancies between real-time PCR and TnT and consequently, this gene was omitted in further analysis. This PrEST also displayed differing fold changes in the array vs. real-time PCR comparisons.

The results from the Trinucleotide Threading assay and real-time PCR generally agreed both with respect to mode and to approximate extent of differential expression, although the TnT approach slightly compresses the data, but less so than cDNA arrays. The average Pearson correlation coefficient across the eight cell lines was 0.84 and the different priming strategies for cDNA generation showed concordant results. The genome-wide array data, on the other hand, exhibited a rather poor correlation with real-time PCR. The results demonstrated that the Trinucleotide Threading approach is a viable alternative for multiplex analysis of moderate gene sets in a high-throughput fashion attaining a high level of discrimination while not forgoing flexibility.

DIGITAL ALLELOTYPING BY TRINUCLEOTIDE THREADING AND MASSIVELY PARALLEL PYROSEQUENCING (PAPER IV)

In Paper IV, the possibility of digital and reliable estimation of allele frequencies in large cohorts was investigated by combining the Trinucleotide Threading assay with a second-generation sequencing

system. Even though the cost per assayed polymorphism is low when employing whole-genome arrays for association studies, the investigation of thousands of individuals for obtaining sufficient power generates a high total cost. Therefore, methods for interrogating allele frequencies by pooling genomic DNA have been proposed. Although significant progress has been made in this direction, accuracy is still an issue due to the fact that array formats provide relative quantifications with large deviations[211].

Emerging second-generation sequencing technologies are providing interrogation of entire sequence-reads in a vast number of parallel reactions and have considerably reduced the cost per base. The idea of extracting targeted alleles from a pool of genomic DNA by forming Trinucleotide Threads followed by digital examination of the pooled cohort using a GS20 instrument (Roche/454 Life Sciences) is therefore appealing.

In this study, 147 single nucleotide polymorphisms were targeted for amplification. Initially, a set of 75 SNPs was designed to target relevant polymorphisms in genes related to obesity and tested in an array-based genotyping procedure as described in Paper I. A total of 88 individuals were genotyped at a conversion rate of 88% and a call rate of 99%. Signal intensities varied considerably between different polymorphisms and three loci gave unspecific signals and were therefore removed after an initial round of genotyping, reducing the active set of threads to 72. The low conversion rate was not surprising since many of the polymorphisms exhibited low GC-content and repetitive regions. In addition, the 75 SNPs from Paper I targeting polymorphisms in cancer related genes were added to increase the total number of targeted loci. Primers for the entire cancer set were included in the assay, despite a known conversion rate of 94% (see Paper I).

As a consequence, a total of 147 threads targeting single nucleotide polymorphisms in a pool of 462 genomes was subjected to TnT amplification followed by massively parallel Pyrosequencing. An array-based PrASE detection of the pooled genomic DNA verifying the thread population prior to sequencing showed 2.6 times stronger signals from the obesity set as compared to the cancer set, likely due to purification of the corresponding probes by PAGE. Approximately 204k reads passed the GS20 filters and were subjected to downstream analysis. A total of 130k reads (63,6%) could immediately be assigned to a thread without tolerating a single sequencing- or polymerase introduced error. Further investigation of the thread population using a more flexible algorithm employing BLAST rendered the identification of 177k (86,6%) in total. Reads that could not be identified may have arisen from microreactors in the emulsion PCR harboring more than one DNA fragment but still passing the GS20 filters. Of the 147 assayed SNPs, 126 gave a high number of reads and were approved (86,3%). An interesting result was that approximately 11k reads (5,6%) could be identified as an illegitimate DNA thread, formed by the extension primer of SNP 16 and the thread-joining primer of SNP 31 in the obesity set with an extension of one base. As predicted in Paper I, this did not affect the outcome of SNP 16 and 31 in the array-based genotyping due to the discriminating design of the threads and genotyping primers.

The number of reads from each locus varied significantly, as well as the number of reads from the two different sets. As predicted by the array experiments, considerably more reads originated from the obesity set (136k or 82%) but the cancer set had less variability among the different DNA threads which may be attributed to a more homogenous GC-content and no low-complexity regions. As expected, the SNPs considered as nonfunctioning in the cancer set showed no or very few sequence counts, except for one (SNP 52).

Since the material in the pool is homogenous and the GS20 system allows for counting reads originating from single threads, the reads from the sequencing run will be expected to follow a binomial distribution. The number of reads for each DNA thread and allele in the pool was calculated to be 1.5. In most cases, this significantly exceeds the number needed for allele frequency determination of a homogenous DNA pool, which indicates scale-up possibilities with regard to the size of the cohort and/or the number of assayed SNPs.

The results indicate that combining Trinucleotide Threading with massively parallel sequencing

provides a flexible platform for association studies by offering the possibility of selecting polymorphisms as well as of adjusting the size of the investigated cohort. Although the total cost is relatively high considering the cost for one sequencing run, pooling of individuals will give a low total cost for the assay as compared to traditional genotyping methods. The obvious drawback is the lack of individual genotype data, but considering the reduction in time, workload and cost while enabling comparisons of cohorts of thousands of individuals, for instance when verifying results obtained in whole-genome association studies, the approach is highly interesting and valuable.

CHAPTER SEVEN

FORWARD LOOKING STATEMENTS

The research papers founding this thesis are describing novel assays and algorithms that all relate to the development of the Trinucleotide Threading method and the facilitation of multiplex amplification and parallel interrogation of nucleic acid signatures. The latter polymers constitute the very foundation for life and the demand for genetic analysis will likely be even more prominent in the future. Therefore, under certain circumstances, the strategies presented in this thesis may provide versatile and useful tools for flexible analysis of a vast number of DNA signatures as our understanding deepens and our need increases.

The investigations in Paper I illustrate the flexibility of the method and also indicate further possibilities of the assay. The successful amplification of 126 DNA threads in Paper IV indicates a considerable increase in the multiplexing capabilities while reducing the amount of reagents needed as well as the time and the cost for the assay.

Since the assay permits low levels of starting material, down to 100 cells, it may be of interest when only low amounts of template are available, i.e. in forensics or when studying loss of heterozygosity. Not only single nucleotide polymorphisms can be queried as the method has been successfully shown to amplify short tandem repeats (STRs) (unpublished data) which may be combined with branch migration[212] for efficient array-identification.

Facilities with an established platform and typing routine may also be able to increase throughput by employing the Trinucleotide Threading approach. Genotyping (Paper I) and expression profiling (Paper III) for diagnostic purposes can be adapted to the Luminex platform[213] which enables high throughput sampling of many individuals at a fairly high multiplexing level. Since the encoded beads of the Luminex system nowadays are magnetic, preparation could be further enhanced with automation using magnet equipped robots[195].

Further work can be done on the algorithms and software tools (Paper I-IV). In terms of thread design, an interesting feature would be the selection of one single reaction tube that will enable parallel amplification of five of the six different types of polymorphisms. In addition, investigation of primer-dimer and illegitimate thread formation can be implemented at the *in silico* design stage. Further improvements of the filters in Paper IV and the design of threads for expression analysis (Paper III) could be integrated with the original design tool. This implementation is preferably done with a web-based interface, offering interactions with dbSNP, OMIM and dbMHC for selection, filtering and subsequent piping to the thread design tool which would offer threads and detection primers to an assay of choice, for instance PrASE or minisequencing. An implementation of the algorithm for genotype calling (Paper II) could also be a useful tool for the software package.

As sequencing cost drops with the introduction of the next generation sequencing systems, we are moving away from analogous arrays towards digital interrogation by sequencing biological systems. Paper IV illustrates the concept of using Trinucleotide Threading for digital allelotyping by massively parallel Pyrosequencing of single thread clones. Although individual genotypes cannot be extracted, the approach provides a cost effective way for accurate and sensitive allelotyping of large cohorts. In these assays, throughput is more important than read length as long as the polymorphisms can be identified and systems like the ABI SOLiD and the Illumina 1G offer millions of reads in parallel as compared to the 400k offered by the 454 system. Additionally, by tagging threads, different cohorts exhibiting different phenotypic traits can be analyzed simultaneously, thus introducing increased flexibility to the system. Regardless of approach – Trinucleotide Threading, the Golden Gate or Molecular Inversion Probes – the combination of multiplex amplification and parallel sequencing may provide cost-effective, accurate and sensitive digital allelotyping of thousands of markers in thousands of individuals.

ABBREVIATIONS

A	Adenine
ABI	Applied Biosystems
AF	Allelic Fraction
AMASE	Apyrase mediated allele-specific extension
AMY1	salivary Amylase gene
API	Application Programming Interface
APS	Adenosine phosphosulphate
ASE	Allele-specific extension
ASH	Allele-specific hybridization
ASO	Allele-specific oligonucleotide
ASOH	Allele-specific oligonucleotide hybridization
ATP	Adenosine triphosphate
bDNA	branched DNA
BLAST	Basic Local Alignment Search Tool
bp	base pairs
C	Cytosine
C2CA	Circle-to-circle amplification
CCD	Charge coupled device
cDNA	complementary DNA
CDS	coding sequence
CEO	Chief Executive Officer
ChIP-Seq	Chromatin Immunoprecipitation sequencing
CNV	Copy number variation
CpG	Cytosine-phosphate-Guanine
CYP	Cytochrome P450
DASL	cDNA-mediated, annealing, selection, extension and ligation
dATP	2'-deoxyadenosine triphosphate
db	database
dCTP	2'-deoxycytidine triphosphate
ddNTP	2',3'-dideoxynucleotide triphosphate
dGTP	2'-deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dNTP	2'-deoxynucleotide triphosphate
DOP	degenerate oligonucleotide primed PCR
dTTP	2'-deoxythymidine triphosphate
DVD	digital versatile disc
EM	Expectation-Maximization
FSP	fragment selection by PCR
G	Guanine
Gb	giga bases
GS20	Genome Sequencer 20
GWA	Genome-Wide Association study
H5N1	Hemagglutinin5 Neuraminidase1
HIV	Human Immunodeficiency virus
HLA	Human leukocyte antigen
HMB	His Majesty's Bark
HPA	Human Protein Atlas
HRCA	hyperbranched rolling circle amplification
Indel	Insertion/deletion
IVS	intervening sequence
K	Guanine or Thymine

kb	kilo bases
LCR	Ligation Chain Reaction
LIMS	Laboratory Information and Management System
LMP	ligation mediated PCR
LOH	loss of heterozygosity
M	Adenine or Cytosine
Mb	mega bases
MDA	multiple displacement amplification
MHC	major histocompatibility complex
MIP	Molecular Inversion Probes
miRNA	microRNA
ML	maximum likelihood
MNP	multi-nucleotide polymorphism
MPSS	Massively Parallel Signature Sequencing
mRNA	messenger RNA
MSV	multi-site variation
NCBI	National Center for Biotechnology Information
OLA	oligonucleotide-ligation assay
OMIM	Online Mendelian Inheritance in Man
PAGE	Polyacrylamide Gel Electrophoresis
PCR	polymerase chain reaction
PEP	primer extension preamplification
Perl	Practical Extraction and Report Language
PPi	pyrophosphate
PrASE	protease mediated allele-specific extension
R	Guanine or Adenine
RASL	RNA-mediated, annealing, selection, extension and ligation
RCA	rolling circle amplification
RNA	ribonucleic acid
S	Cytosine or Guanine
SARS	Severe Acute Respiratory Syndrome
SBE	single-base extension
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
SNP	single nucleotide polymorphism
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
SQL	Structured Query Language
SSR	simple sequence repeat
STR	short tandem repeat
T	Thymine
TLA	Three letter acronym
TnT	Trinucleotide Threading
ULCS	ultra-low-cost sequencing
W	Adenine or Thymine
WGA	whole genome amplification
WGSA	Whole genome sampling analysis
XML	eXtensible Markup Language
Y	Cytosine or Thymine

EPILOGUE

SOMETIMES OBJECTS ARE CLOSER THAN THEY APPEAR

I bland är väsentliga saker betydligt närmare än vad man tror och det gäller inte bara i en ytterbackspegel på passagerarsidan. Det finns även tillfällen när till och med smånördiga, och till vetenskapen hängivna, forskarstuderande börjar tro på högre makter. Som till exempel när man stiger upp klockan 02:30 en morgon och flyger till bilmässan i Frankfurt för att 84 timmar senare upptäcka att mässans huvudattraktion – en sann ikon för vetenskap och konst – står parkerad framför ens egen port. En av väldigt få enheter i Sverige som, om den hade stått parkerad där 84 timmar tidigare, eventuellt hade gjort resan betydligt kortare. Men nu gjorde den inte det, utan stod där först flera dagar senare. Vid min gata. När jag passerar förbi. Varför. Tolkningar av sådana här sammanträffanden är väldigt svåra för en doktorand med en mycket liten hjärna som besväras av långa sekvenser. Men när nu flitens lykta lyser i det lilla biblioteket denna tidiga söndagsmorgon i slutet av oktober tror jag mig komma till insikt.

Hur långt man än reser för att uppleva äventyr och glädje så finns de båda betydligt närmare än man kanske tror. Och allt medan blåmesen hackar på det kalla fönsterblecket och vill komma in till mig och min MacBook inser jag att av alla mina resor, drömmar och exkursioner har det bästa äventyret och den största glädjen hela tiden funnits mitt framför mig. Science under Kaliforniens sol eller vid Long Islands kust har på det hela taget svårt att jämföra sig med äventyret vid Brunnsvikens strand. Allt sedan den första tråden skapades i ett tidsrum långt, långt borta har jag haft ett obeskrivligt roligt företag med många av de moment som jag älskar – teknologi, biokemi, programmering och robotar i en underbar miljö, full av frihet och skaparglädje tillsammans med helt fantastiska människor.

Alla har vi vår Akilleshäl. Jag har minst två. Men det jag snarast tänker på nu är att jag kanske inte har så lätt för att uttrycka min uppskattning alla gånger. Jag har haft många varma och minnesvärda möten med människor som bidragit till harmoniska och glada dagar, månader och år och att denna avhandling blivit till. För det är jag evigt tacksam. Min största tacksamhet och uppskattning vill jag rikta till min sensei och läromästare **Afshin**, för att du lotsade mig igenom ett examensarbete av högriskkaraktär som ledde vidare till flera år av fördjupad forskning. För ditt fantastiska handledarskap fyllt av engagemang, kreativa idéer och effektiv problemlösning och för en avslappnad atmosfär där din dörr alltid står öppen. Jag är obeskrivligt glad för att jag fick förtroendet och friheten att genomföra projekt i linje med mina intressen och även om **någon** har kallat våra möten laid back har det resultatmässigt varit en succé att blanda vetenskap med diskussioner om allt mellan himmel och jord. Jag vill även rikta ett stort tack till **Joakim**, min andra handledare, för att jag fick börja som doktorand och för dessa år av oändlig entusiasm och visdom. Du har lärt mig vad det innebär att vara effektiv och att alltid se det positiva i forskningsresultat. Inte minst är jag tacksam för samtal om vindsurfing och för att du bjuder på en Red Bull eller en Delicatoboll när det verkligen behövs. **Mathias Uhlén** vill jag tacka för en bra utbildning, för att vara en aldrig sinande källa till inspiration och för att visa att det går att reducera barriärer och gränser – såväl vetenskapliga som geografiska. Även ett tack för din strålande energi och för din förmåga att leverera motiverande superlativ likt **Steve Jobs** samt för ett rekommendationsbrev som gjorde att jag fick uppleva några minst lika intensiva höstveckor som nu, men på andra sidan ett stort hav. Ett tack till professorerna **Per-Åke**, **Stefan** och **Sophia** för bra undervisning, värdefulla råd och en bild av hur mitt drömkontor skulle se ut. Även ett tack till **Karl Hult** för bra samarbete, många intressanta presentationer och prat om långfärdsskridskor samt till **Gunnel Dalhammar** för givande insikter och för ditt bidrag till utbildningen. Vidare vill jag tacka alla PIs, inklusive **Patrik Samuelsson**, **Peter Savolainen**, **Jacob Odeberg**, **Peter Nilsson**, **Jenny Ottosson** och **Henrik Wernérus**, för bra kurser och hjälp med allehanda ting samt för det mycket varierade och spännande forskningsklimatet på *"Stureplan"*. **Pawel**, en av få som jag faktiskt litar mer på än mig själv och **Patrik**, effektiviteten personifierad – utan er hade jag inte varit där jag är idag. Tack för allt samarbete och för allt roligt vi har

haft, inte bara innanför labbets väggar. **TnT** is truly powered by **PnP**. Ett stort tack till **Hedvig** och **Rolf** för att ha visat mig en cool värld av statistisk modellering och till **Josefin** för vidgade forskningsvyer. Till denna grupp av essentiella *partners in crime* hör även **någon** mer, ingen mindre än **Mats Lindskog**. Tack för allt samarbete av bioinformatisk och datalogisk natur, men även för luncher, möten på ett annat Stureplan och prat om båtar och skärgården. Jag vill tacka mina föregångare. **Emilie** för roliga stunder i, såväl som utanför, labbet och **Max** – tack, inte minst för burgarna i Silicon Valley och räksmörgåsarna i Sandhamn.

I mitt arbete har jag fått mycket praktisk hjälp från många håll. Jag vill speciellt tacka **Kicki**, **Torbjörn**, **Annica**, **Mia**, **Anders Holmberg**, **Henrik Uhlén** och **Anna Eidefors** för hjälp med bl.a. MBSer, skriptdesign och pyroapparater. Ett stort tack även till **Annelie**, **Mårten** och **Ronald** för hjälp med printning av mina slides och till **Bahram** för hjälp med sekvenering. Vidare vill jag visa min uppskattning till alla mina härliga rumskamrater, gamla som nya. **Nina Bandmann** för att du verkligen försökte få mig att dricka mindre Cola och äta mer frukt. **Ason** inte minst för vetenskapliga idéer, vare sig de kläcks framför datorn i rummet eller bland bergslejon i Yosemite och **Jochen**, mästaren i effektivt labbande med Excelsorteringar och innovativa pappers- och platthållare. **Anna Westring**, inte minst för att du tagit hand om **Greta** åt mig och **Rebecca** för expressionshjälp och Lago di Como. **Cilla**, inte minst för gott samarbete under grundutbildningens labbar och **Seiji Shibasaki** inte minst för att ha visat mig OS X på japanska. **My** vill jag tacka, inte minst för min fina sax. Jag använder den varje dag. Nästan. Sist men inte minst, mästarens mästare – **Daniel Klevebring** för allehanda kluriga programmeringslösningar, kritiskt tänkande, effektiv problemlösning och alla dina trevliga fester.

Luncherna skulle inte blivit desamma utan **Christian Natanaelsson** och **Mattias Oskarsson**. Skönt att det finns flera som inte orkar laga mat så ofta (läs nästan aldrig). Som tur är har det också funnits folk som verkligen förstår värdet av teknikprylar. Ett tack till **Jesper**, även för löpning och bad i Brunnsviken, **Eric Björkvall** för dykningstips och all machjälp samt **Carl Hamsten**, **Erik Malm** och **Danko**. I nära anslutning finns bioinformatiken och där vill jag tacka **Lisa** för värdefull hjälp, **Jorge** och **Malin** (även för Stoffel) samt **Linn**, **Per Unneberg** och **Michael Strömberg**.

Gene Technology har många andra härliga personligheter som bidragit på olika sätt **Cecilia**, **Esther**, **Marcus**, **Johan Lindberg** och **Valtteri**. Även nyare stjärnor som **Julia Sandberg, Henrik Stranneheim** och **Johanna Hasmats** vill jag tacka, inte minst för trevlig biljard och givande diskussioner.

Ibland händer det att man gör ett *guest appearance* i något proteinlabb, till exempel Sörgården. Där vill jag gärna tacka **John** och **Nina Kronqvist** för att snällt låna mig utrymme och **Johan Rockberg** för lån av pipetter – men också för att ha lärt mig konsten "hur man landar ett tänk" och avancerad matlagning. **Erik Vernet** för allehanda inspirerande diskussioner om segling och business. Även ett tack till **Hanna**, **Caroline** och **Maria** för trevliga år på grundutbildningen och lika trevliga på forskarnivå. Tack till **Anna Sköllermo** för bl.a. hiphop, San Diego samt assistans i att ta den parallella trådningen till expressionsvärlden.

Min tid på strukturbiologi var mycket givande och lärorik där jag lärde känna **Jacob Dogan**, **Per-Åke Löfdahl**, **Inger** och **Holger.** Även ett tack till **Anders Hamberg** för trevliga luncher och **Sacha Akhras** för alla år av kemi, från gymnasiets början till nu, minst 10 år senare. Tack också för skjutsen till flygplatsen i San Francisco. **Reza**, for your incredible hospitality and for showing me the heart of L.A. Crusing in downtown under the Californian sun to the tunes of ABBA is the meaning of life.

Jag vill även tacka alla andra trevliga människor på labbet för roliga jul-, vår- och disputationsfester med mera, däribland **Johanna Steen**, **Cajsa**, **Torun**, **Emma Lundberg**, **Jenny Fall**, **Paula**, **Tove**, **Basia**, **Mikaela**, **Sara, Sebastian Grimm** och **Anna Konrad**. I övrigt vill jag tacka **Jojje** för en Simpsonsglass och en energidryck.

Min värld utanför labbet har faktiskt bidragit mer än man kan tro till denna avhandling. Jag vill därför rikta ett tack till följande. **Fredrik** inte minst för Falkenberg, Formentor och vindsurfing. **Stefan Lindberg**, inte minst för en nedfryst bakterie, immaterialrätt, sushiwraps vid Nybrokajen och

mantarockor utanför Azorerna. **Alex Flores**, inte minst för all träning, bastu i Umeå och många framgångsrika projekt. Kapten **Ställberg**, inte minst för vandring i Mount Aspiring, bitande blåmesar vid Landsort och träning tillsammans med **Greger**. Mina vänner sedan grundskolan; **Carl Ljungmark** för 20 år av upptåg, från Narmer och Egypten framför Super Mario Kart, via otaliga timmar Karate till kraschade radiostyrda helikoptrar. Tack även till **Jonas**, **Daniel**, **Mounir**, **Joel** och **Björn**.

Carl Drougge för att illustrera vad äga passion för något verkligen innebär. Installera Debian på en PowerBook G3 och skriva drivrutiner för ethernetkort till Amiga är inte bara grymt. Det är helt otroligt. **Palle Derkert**, inte minst för design, bilmässor, Fjällbacka och "thinking out of the box". **Mathias Montin** för framgångsrika kemilabbar under gymnasiet och **Kristian Wikström** för alla badmintonmatcher. **Eric Wahlforss** och **Martin** för min hittills korta, men framgångsrika, karriär på Handelshögskolan och alla vänner på **Tatsue Karate Dojo** som jag inte ser så ofta längre, men vi hade det riktigt roligt på alla träningsläger i skärgården. **Knut Gezelius** för att ha introducerat mig till konceptet "Work hard, play hard". Det kan faktiskt vara lika roligt på labbet en julafton som att dyka med havssköldpaddor utanför Påskön. Tråkigt att vi inte kan ses oftare, men det är lika roligt varje gång. En halvtimmes middag vid Centralen, en Mojito i hamnen vid St. Barths eller en Long Island Ice Tea på en nattklubb i Moskva. **Marie**, tack för att du gav mitt liv en ny dimension.

Avslutningsvis vill jag rikta mitt allra varmaste tack till min familj för den omtanke, stöd och hjälp jag får i alla väder. Den där **Efva** och **Far** som alltid finns där för mig. Tack för alla resor till landet medan jag har författat den här boken. Tack till den lilla glada **tanten** och till **moster** samt **farbror** som också är ett stort stöd och hjälp. Och till den där lilla, lilla **människan** som krävde ett hedersomnämnande. Alternativt stå först. Alternativt stå i extra stor text. Nu hamnar den här istället. Sist. Med vanligt text. Men det är kanske lite av ett hedersomnämnande det också... Pet! ;)

Den här avhandlingen börjar närma sig sitt slut. Tusentals spetsar senare, och antagligen lika många Colaburkar, kan jag konstatera att drömmar ibland går i uppfyllelse. Det blir inte coolare än så här och utan er – några av de finaste människorna i världen – hade det inte gått. Jag är sagolikt glad för de här åren och det finns nog ingenting jag ångrar. Bortsett från att jag borde hållit bättre ordning vid min skrivplats. Och kanske att jag borde ätit lunch i lunchrummet vid mer än ett tillfälle. Och kanske, kanske borde jag inte handlat i automaten i korridoren så ofta...

Erik Pettersson
Brotorp i oktober 2007

REFERENCES

1. Rusch, D.B. et al. The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**, e77 (2007).
2. Yooseph, S. et al. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol* **5**, e16 (2007).
3. http://www.sorcererIIexpedition.org.
4. http://www.syntheticgenomics.com.
5. http://www.biobricks.org/.
6. Hutchison, C.A. et al. Global transposon mutagenesis and a minimal Mycoplasma genome. *Science, New York, N.Y* **286**, 2165-2169 (1999).
7. http://www.kpcb.com.
8. Zust, R. et al. Coronavirus non-structural protein 1 is a major pathogenicity factor: implications for the rational design of coronavirus vaccines. *PLoS pathogens* **3**, e109 (2007).
9. Yen, H.L. et al. Neuraminidase Inhibitor Resistant Recombinant A/Vietnam/1203/04 (H5N1) Influenza Viruses Retain Their Replication Efficiency and Pathogenicity in Vitro and in Vivo. *J Virol* (2007).
10. Pan, C., Kim, J., Chen, L., Wang, Q. & Lee, C. The HIV positive selection mutation database. *Nucleic acids research* **35**, D371-375 (2007).
11. Wells, C.D. et al. HIV infection and multidrug-resistant tuberculosis: the perfect storm. *The Journal of infectious diseases* **196 Suppl 1**, S86-107 (2007).
12. Parola, P. et al. Antimalarial Drug Susceptibility and Point Mutations Associated with Drug Resistance in 248 Plasmodium falciparum Isolates Imported from Comoros to Marseille, France in 2004 2006. *Am J Trop Med Hyg* **77**, 431-437 (2007).
13. http://www.who.int/drugresistance/en/.
14. http://www.genentech.com.
15. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
16. Venter, J.C. et al. The sequence of the human genome. *Science, New York, N.Y* **291**, 1304-1351 (2001).
17. Steemers, F.J. et al. Whole-genome genotyping with the single-base extension assay. *Nature methods* **3**, 31-33 (2006).
18. http://www.illumina.com.
19. Ingelman-Sundberg, M. Pharmacogenetics of cytochrome P450 and its applications in drug therapy: the past, present and future. *Trends in pharmacological sciences* **25**, 193-200 (2004).
20. http://www.kurzweilai.net/articles/art0648.html.
21. Service, R.F. Gene sequencing. The race for the $1000 genome. *Science, New York, N.Y* **311**, 1544-1546 (2006).
22. Trabesinger, A. Formula 1 racing: power games. *Nature* **447**, 900-903 (2007).
23. http://www.iht.com/articles/2007/06/01/america/dna.php.
24. http://www.genomics.xprize.org.
25. Santiago, C. et al. ACTN3 genotype in professional soccer players. *Br J Sports Med* (2007).
26. Lucia, A. et al. ACTN3 genotype in professional endurance cyclists. *International journal of sports medicine* **27**, 880-884 (2006).
27. Yang, N. et al. ACTN3 genotype is associated with human elite athletic performance. *American journal of human genetics* **73**, 627-631 (2003).
28. http://www.gtg.com.au.

29. http://www.23andme.com.
30. Illumina Analyst Day, September 15th 2007, Mandarin Oriental, New York, NY.
31. Watson, J.D. & Crick, F.H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
32. Reif, J.H. Computing. Successes and challenges. *Science, New York, N.Y* **296**, 478-479 (2002).
33. Adleman, L.M. Molecular computation of solutions to combinatorial problems. *Science, New York, N.Y* **266**, 1021-1024 (1994).
34. Braich, R.S., Chelyapov, N., Johnson, C., Rothemund, P.W. & Adleman, L. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science, New York, N.Y* **296**, 499-502 (2002).
35. Gerstein, M.B. et al. What is a gene, post-ENCODE? History and updated definition. *Genome research* **17**, 669-681 (2007).
36. Pearson, H. Genetics: what is a gene? *Nature* **441**, 398-401 (2006).
37. Yelin, R. et al. Widespread occurrence of antisense transcription in the human genome. *Nature biotechnology* **21**, 379-386 (2003).
38. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research* **34**, D140-144 (2006).
39. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799-816 (2007).
40. http://www.ensembl.org.
41. http://www.wormbook.org.
42. Levy, S. et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* **5**, e254 (2007).
43. Feuk, L., Carson, A.R. & Scherer, S.W. Structural variation in the human genome. *Nature reviews* **7**, 85-97 (2006).
44. Chasman, D.I. et al. Pharmacogenetic study of statin therapy and cholesterol reduction. *Jama* **291**, 2821-2827 (2004).
45. Perry, G.H. et al. Diet and the evolution of human amylase gene copy number variation. *Nature genetics* (2007).
46. Church, G.M. Genomes for all. *Scientific American* **294**, 46-54 (2006).
47. http://www.hapmap.org/.
48. Frazer, K.A. et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
49. Gibbs, J.R. & Singleton, A. Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. *PLoS genetics* **2**, e150 (2006).
50. Haiman, C.A. et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature genetics* **39**, 638-644 (2007).
51. Gudmundsson, J. et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nature genetics* **39**, 977-983 (2007).
52. Gudmundsson, J. et al. Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature genetics* **39**, 631-637 (2007).
53. Tomlinson, I. et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988 (2007).
54. Moffatt, M.F. et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470-473 (2007).
55. Steinthorsdottir, V. et al. A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nature genetics* **39**, 770-775 (2007).
56. Sladek, R. et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881-885 (2007).
57. Lu, Y. et al. A gene expression signature predicts survival of patients with stage I non-small cell lung cancer. *PLoS medicine* **3**, e467 (2006).

58. Uhlen, M. et al. A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **4**, 1920-1932 (2005).
59. Nilsson, B.L., Soellner, M.B. & Raines, R.T. Chemical synthesis of proteins. *Annual review of biophysics and biomolecular structure* **34**, 91-118 (2005).
60. Poinar, G.O., Jr. & Danforth, B.N. A fossil bee from Early Cretaceous Burmese amber. *Science, New York, N.Y* **314**, 614 (2006).
61. Noonan, J.P. et al. Sequencing and analysis of Neanderthal genomic DNA. *Science, New York, N.Y* **314**, 1113-1118 (2006).
62. Green, R.E. et al. Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330-336 (2006).
63. Poinar, H.N. et al. Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science, New York, N.Y* **311**, 392-394 (2006).
64. Braslavsky, I., Hebert, B., Kartalov, E. & Quake, S.R. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 3960-3964 (2003).
65. Levene, M.J. et al. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science, New York, N.Y* **299**, 682-686 (2003).
66. http://www.helicosbio.com.
67. Saiki, R.K. et al. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science, New York, N.Y* **230**, 1350-1354 (1985).
68. Mullis, K. et al. Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology* **51 Pt 1**, 263-273 (1986).
69. Zhang, L. et al. Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 5847-5851 (1992).
70. Telenius, H. et al. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718-725 (1992).
71. Landegren, U. & Nilsson, M. Locked on target: strategies for future gene diagnostics. *Annals of medicine* **29**, 585-590 (1997).
72. Nilsson, M. et al. Making ends meet in genetic analysis using padlock probes. *Human mutation* **19**, 410-415 (2002).
73. Shuber, A.P., Grondin, V.J. & Klinger, K.W. A simplified procedure for developing multiplex PCRs. *Genome research* **5**, 488-493 (1995).
74. Brownie, J. et al. The elimination of primer-dimer accumulation in PCR. *Nucleic acids research* **25**, 3235-3241 (1997).
75. Wang, H.Y. et al. A genotyping system capable of simultaneously analyzing >1000 single nucleotide polymorphisms in a haploid genome. *Genome research* **15**, 276-283 (2005).
76. Landegren, U., Kaiser, R., Sanders, J. & Hood, L. A ligase-mediated gene detection technique. *Science, New York, N.Y* **241**, 1077-1080 (1988).
77. Barany, F. Genetic disease detection and DNA amplification using cloned thermostable ligase. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 189-193 (1991).
78. Nilsson, M. et al. Padlock probes: circularizing oligonucleotides for localized DNA detection. *Science, New York, N.Y* **265**, 2085-2088 (1994).
79. Baner, J. et al. Parallel gene analysis with allele-specific padlock probes and tag microarrays. *Nucleic acids research* **31**, e103 (2003).
80. Hardenbol, P. et al. Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature biotechnology* **21**, 673-678 (2003).
81. Fan, J.B. et al. Highly parallel SNP genotyping. *Cold Spring Harbor symposia on quantitative biology* **68**, 69-78 (2003).

82. Kennedy, G.C. et al. Large-scale genotyping of complex DNA. *Nature biotechnology* **21**, 1233-1237 (2003).
83. Shapero, M.H. et al. MARA: a novel approach for highly multiplexed locus-specific SNP genotyping using high-density DNA oligonucleotide arrays. *Nucleic acids research* **32**, e181 (2004).
84. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic acids research* **33**, e71 (2005).
85. Stenberg, J., Dahl, F., Landegren, U. & Nilsson, M. PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. *Nucleic acids research* **33**, e72 (2005).
86. Dahl, F. et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 9387-9392 (2007).
87. Blanco, L. et al. Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *The Journal of biological chemistry* **264**, 8935-8940 (1989).
88. Esteban, J.A., Salas, M. & Blanco, L. Fidelity of phi 29 DNA polymerase. Comparison between protein-primed initiation and DNA polymerization. *The Journal of biological chemistry* **268**, 2719-2726 (1993).
89. Fire, A. & Xu, S.Q. Rolling replication of short DNA circles. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 4641-4645 (1995).
90. Lizardi, P.M. et al. Mutation detection and single-molecule counting using isothermal rolling-circle amplification. *Nature genetics* **19**, 225-232 (1998).
91. Dahl, F. et al. Circle-to-circle amplification for precise and sensitive DNA analysis. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4548-4553 (2004).
92. Dean, F.B. et al. Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5261-5266 (2002).
93. Pinard, R. et al. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC genomics* **7**, 216 (2006).
94. Lovmar, L., Fredriksson, M., Liljedahl, U., Sigurdsson, S. & Syvanen, A.C. Quantitative evaluation by minisequencing and microarrays reveals accurate multiplexed SNP genotyping of whole genome amplified DNA. *Nucleic acids research* **31**, e129 (2003).
95. Paez, J.G. et al. Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic acids research* **32**, e71 (2004).
96. Hosono, S. et al. Unbiased whole-genome amplification directly from clinical samples. *Genome research* **13**, 954-964 (2003).
97. Lage, J.M. et al. Whole genome analysis of genetic alterations in small DNA samples using hyperbranched strand displacement amplification and array-CGH. *Genome research* **13**, 294-307 (2003).
98. Lovmar, L. & Syvanen, A.C. Multiple displacement amplification to create a long-lasting source of DNA for genetic studies. *Human mutation* **27**, 603-614 (2006).
99. U.S. Patent 5,641,658.
100. Adessi, C. et al. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research* **28**, E87 (2000).
101. Mitra, R.D. & Church, G.M. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic acids research* **27**, e34 (1999).
102. Leamon, J.H. et al. A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**, 3769-3777 (2003).
103. Margulies, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

104. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science, New York, N.Y* **309**, 1728-1732 (2005).
105. Meuzelaar, L.S., Lancaster, O., Pasche, J.P., Kopal, G. & Brookes, A.J. MegaPlex PCR: a strategy for multiplex amplification. *Nature methods* **4**, 835-837 (2007).
106. http://www.affymetrix.com.
107. Wallace, R.B. et al. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic acids research* **6**, 3543-3557 (1979).
108. Syvanen, A.C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature reviews* **2**, 930-942 (2001).
109. Syvanen, A.C. From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human mutation* **13**, 1-10 (1999).
110. Hacia, J.G. Resequencing and mutational analysis using oligonucleotide microarrays. *Nature genetics* **21**, 42-47 (1999).
111. Newton, C.R. et al. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic acids research* **17**, 2503-2516 (1989).
112. Alves, A.M. & Carr, F.J. Dot blot detection of point mutations with adjacently hybridising synthetic oligonucleotide probes. *Nucleic acids research* **16**, 8723 (1988).
113. Kwok, S. et al. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic acids research* **18**, 999-1005 (1990).
114. Ayyadevara, S., Thaden, J.J. & Shmookler Reis, R.J. Discrimination of primer 3'-nucleotide mismatch by taq DNA polymerase during polymerase chain reaction. *Analytical biochemistry* **284**, 11-18 (2000).
115. Ahmadian, A., Gharizadeh, B., O'Meara, D., Odeberg, J. & Lundeberg, J. Genotyping by apyrase-mediated allele-specific extension. *Nucleic acids research* **29**, E121 (2001).
116. O'Meara, D., Ahmadian, A., Odeberg, J. & Lundeberg, J. SNP typing by apyrase-mediated allele-specific primer extension on DNA microarrays. *Nucleic acids research* **30**, e75 (2002).
117. Hultin, E., Kaller, M., Ahmadian, A. & Lundeberg, J. Competitive enzymatic reaction to control allele-specific extensions. *Nucleic acids research* **33**, e48 (2005).
118. Hardenbol, P. et al. Highly multiplexed molecular inversion probe genotyping: over 10,000 targeted SNPs genotyped in a single tube assay. *Genome research* **15**, 269-275 (2005).
119. Gunderson, K.L. et al. Decoding randomly ordered DNA arrays. *Genome research* **14**, 870-877 (2004).
120. Matsuzaki, H. et al. Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. *Genome research* **14**, 414-425 (2004).
121. Matsuzaki, H. et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nature methods* **1**, 109-111 (2004).
122. Hinds, D.A. et al. Whole-genome patterns of common DNA variation in three human populations. *Science, New York, N.Y* **307**, 1072-1079 (2005).
123. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature genetics* **37**, 549-554 (2005).
124. Nyren, P. The history of pyrosequencing. *Methods in molecular biology (Clifton, N.J* **373**, 1-14 (2007).
125. Bentley, D.R. Whole-genome re-sequencing. *Current opinion in genetics & development* **16**, 545-552 (2006).
126. http://www.ls9.com.
127. http://www.amyrisbiotechnologies.com.
128. http://www.aurorabiofuels.com.
129. Shendure, J., Mitra, R.D., Varma, C. & Church, G.M. Advanced sequencing technologies: methods and goals. *Nature reviews* **5**, 335-344 (2004).

130. Barski, A. et al. High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837 (2007).
131. Mikkelsen, T.S. et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560 (2007).
132. Sanger, F., Nicklen, S. & Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**, 5463-5467 (1977).
133. Smith, L.M. et al. Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).
134. Prober, J.M. et al. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science, New York, N.Y* **238**, 336-341 (1987).
135. Innis, M.A., Myambo, K.B., Gelfand, D.H. & Brow, M.A. DNA sequencing with Thermus aquaticus DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 9436-9440 (1988).
136. Carothers, A.M., Urlaub, G., Mucha, J., Grunberger, D. & Chasin, L.A. Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. *BioTechniques* **7**, 494-496, 498-499 (1989).
137. Emrich, C.A., Tian, H., Medintz, I.L. & Mathies, R.A. Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Analytical chemistry* **74**, 5076-5083 (2002).
138. Koutny, L. et al. Eight hundred-base sequencing in a microfabricated electrophoretic device. *Analytical chemistry* **72**, 3388-3391 (2000).
139. Cutler, D.J. et al. High-throughput variation detection and genotyping using microarrays. *Genome research* **11**, 1913-1925 (2001).
140. Lipshutz, R.J. et al. Using oligonucleotide probe arrays to access genetic diversity. *BioTechniques* **19**, 442-447 (1995).
141. Patil, N. et al. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science, New York, N.Y* **294**, 1719-1723 (2001).
142. Drmanac, R., Labat, I., Brukner, I. & Crkvenjakov, R. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics* **4**, 114-128 (1989).
143. http://www.454.com.
144. Ronaghi, M., Uhlen, M. & Nyren, P. A sequencing method based on real-time pyrophosphate. *Science, New York, N.Y* **281**, 363, 365 (1998).
145. Brenner, S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology* **18**, 630-634 (2000).
146. Brenner, S. et al. In vitro cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1665-1670 (2000).
147. http://solid.appliedbiosystems.com.
148. http://www.visigenbio.com.
149. http://www.completegenomics.com.
150. http://www.intelligentbiosystems.com/.
151. Ju, J. et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 19635-19640 (2006).
152. http://www.pacificbiosciences.com.
153. Lewis, M. The New New Thing: A Silicon Valley Story.
154. http://www.zakon.org/robert/internet/timeline.
155. http://www.kurzweilai.net/meme/frame.html?main=/articles/art0184.html.
156. Moore, J.H. Bioinformatics. *Journal of cellular physiology* **213**, 365-369 (2007).

157. http://bioinformatics.ca/links_directory.
158. Fox, J.A., McMillan, S. & Ouellette, B.F. Conducting research on the web: 2007 update for the bioinformatics links directory. *Nucleic acids research* **35**, W3-5 (2007).
159. http://www.ncbi.nlm.nih.gov/Genbank/.
160. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. & Wheeler, D.L. GenBank. *Nucleic acids research* **35**, D21-25 (2007).
161. http://camera.calit2.net/.
162. Wheeler, D.L. et al. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **35**, D5-12 (2007).
163. Galperin, M.Y. The Molecular Biology Database Collection: 2007 update. *Nucleic acids research* **35**, D3-4 (2007).
164. http://www.ncbi.nlm.nih.gov.
165. Sherry, S.T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308-311 (2001).
166. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A. & McKusick, V.A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* **33**, D514-517 (2005).
167. http://www.cypalleles.ki.se.
168. http://snp500cancer.nci.nih.gov.
169. http://obesitygene.pbrc.edu.
170. Rankinen, T. et al. The human obesity gene map: the 2005 update. *Obesity (Silver Spring, Md* **14**, 529-644 (2006).
171. http://genome.ucsc.edu.
172. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403-410 (1990).
173. Smith, T.F. & Waterman, M.S. Identification of common molecular subsequences. *Journal of molecular biology* **147**, 195-197 (1981).
174. Needleman, S.B. & Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* **48**, 443-453 (1970).
175. http://emboss.sourceforge.net.
176. http://www.w3.org/XML/.
177. http://www.mysql.com.
178. http://www.oracle.com.
179. http://www.postgresql.org.
180. http://java.sun.com.
181. http://biojava.org.
182. http://www.perl.org.
183. http://www.bioperl.org.
184. http://www.python.org.
185. http://biopython.org.
186. http://www.ruby-lang.org.
187. http://bioruby.com.
188. Hua, J. et al. SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. *Bioinformatics* **23**, 57-63 (2007).
189. Teo, Y.Y. et al. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* (2007).
190. Couzin, J. & Kaiser, J. Genome-wide association. Closing the net on common disease genes. *Science, New York, N.Y* **316**, 820-822 (2007).
191. http://www.r-project.org.

192. http://www.bioconductor.org.
193. Reimers, M. & Carey, V.J. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods in enzymology* **411**, 119-134 (2006).
194. Syvanen, A.C. Toward genome-wide SNP genotyping. *Nature genetics* **37 Suppl**, S5-10 (2005).
195. Holmberg, A. et al. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *Electrophoresis* **26**, 501-510 (2005).
196. Moorhead, M. et al. Optimal genotype determination in highly multiplexed SNP data. *Eur J Hum Genet* **14**, 207-215 (2006).
197. Xiao, Y., Segal, M.R., Yang, Y.H. & Yeh, R.F. A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics* **23**, 1459-1467 (2007).
198. Plagnol, V., Cooper, J.D., Todd, J.A. & Clayton, D.G. A method to address differential bias in genotyping in large-scale association studies. *PLoS genetics* **3**, e74 (2007).
199. Lovmar, L., Ahlford, A., Jonsson, M. & Syvanen, A.C. Silhouette scores for assessment of SNP genotype clusters. *BMC genomics* **6**, 35 (2005).
200. Callegaro, A. et al. Algorithm for automatic genotype calling of single nucleotide polymorphisms using the full course of TaqMan real-time data. *Nucleic acids research* **34**, e56 (2006).
201. Carvalho, B., Bengtsson, H., Speed, T.P. & Irizarry, R.A. Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics (Oxford, England)* **8**, 485-499 (2007).
202. Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **Series B, 39(1):1-38** (1977).
203. McLachlan, G.J. & Thriyambakam, K. The EM algorithm and extensions. *Wiley Interscience* (1997).
204. McLachlan, G.J. & Peel, D. Finite Mixture Models. *Wiley Interscience* (2000).
205. van 't Veer, L.J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).
206. Bull, T.M. et al. Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *American journal of respiratory and critical care medicine* **170**, 911-919 (2004).
207. Fan, J.B. et al. A versatile assay for high-throughput gene expression profiling on universal array matrices. *Genome research* **14**, 878-885 (2004).
208. Yeakley, J.M. et al. Profiling alternative splicing on fiber-optic arrays. *Nature biotechnology* **20**, 353-358 (2002).
209. Peck, D. et al. A method for high-throughput gene expression signature analysis. *Genome biology* **7**, R61 (2006).
210. Flagella, M. et al. A multiplex branched DNA assay for parallel quantitative gene expression profiling. *Analytical biochemistry* **352**, 50-60 (2006).
211. Macgregor, S. Most pooling variation in array-based DNA pooling is attributable to array error rather than pool construction error. *Eur J Hum Genet* **15**, 501-504 (2007).
212. Pourmand, N. et al. Branch migration displacement assay with automated heuristic analysis for discrete DNA length measurement using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6146-6151 (2007).
213. http://www.luminexcorp.com/.