APPLICATIONS OF NEXT GENERATION SEQUENCING IN (EPI)GENETICS

Michal Mokrý

ISBN: 978-94-6182-040-2

Layout & printing: Off Page, Amsterdam

APPLICATIONS OF NEXT GENERATION SEQUENCING IN (EPI)GENETICS

Toepassing van next-generation sequencing in (epi)genetica (met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op woensdag 23 november 2011 des ochtends te 10.30 uur

door

Michal Mokrý
geboren op 7 december 1982 te Košice, Slowakije

Promotor: Prof.dr. E. P. J. G. Cuppen

The research described in this thesis was performed at the Hubrecht Institute of the Royal Academy of Arts and Sciences (KNAW), within the framework of the Graduate School of Cancer Genomics and Developmental Biology (CGDB) in Utrecht.

Printing of this thesis was financially supported by the J.E. Jurriaanse stichting.

CONTENTS

1

DNA SEQUENCING YESTERDAY, TODAY, AND TOMORROW

ABSTRACT

Technological developments in DNA sequencing are an excellent example of how major advances in scientific techniques can lead to numerous conceptual discoveries in the life sciences. Starting in the 1970s with the development of DNA sequencing by chain termination methods, through the introduction of fluorescently labeled terminators and progress in sequencing automation in the late 1980s and 1990s, to the introduction of next-generation sequencers in the mid 2000s, these new approaches consistently opened novel and often unexpected horizons for life scientists. In this introduction, I present the major technological developments in DNA sequencing, discuss current possibilities and future challenges and opportunities, and outline the scope of this thesis.

Early Sanger sequencing era

Since the isolation of "nuclein," later named deoxyribonucleic acid, or DNA, by Friedrich Miescher in 1869 during his experiments with leukocyte nuclei (1), more than a century passed before the first practical techniques were developed to determine the sequential order of nucleotide bases – adenine, cytosine, guanine, and thymine- in DNA molecule. In the 1970s, two mainstream sequencing principles were used: the first used partial breaking of terminally-labeled DNA molecules at specific bases (2) and the second – now commonly known as Sanger sequencing, used dideoxynucleotides (ddNTPs) as terminators of DNA synthesis (3). In both methods, the resulting DNA fragments were separated by electrophoresis, and relative position of the fragmented bands determined the sequence of the DNA sample. Due to its efficiency and more versatile chemistry, the Sanger method became the method of choice in later years. Shortly after the introduction of the Sanger method, the genome of the DNA-based organism phi X 174 phage was completely sequenced (4) and showed, for the first time, the way genes and regulatory elements are organized in a simple genome.

Fluorescently labeled ddNTPs and the automation of Sanger sequencing

Though the sequencing of tiny viral genomes or the coding sequences of entire genes was within the limits of the original Sanger sequencing method, larger genomes like those of bacteria or eukaryotes were out of reach due to relatively high costs per sequenced base and difficulties in automation and parallelization of the sequencing process. The sequencing of each DNA fragment required four independent termination reactions for each base with radiolabeled ddNTPs, separation of the DNA fragments on long polyacrylamide gels, and visualization of radioactive signal with an estimated throughput of 5000 bases per week. These burdens were eliminated by the introduction of fluorescently-labeled primers and later ddNTPs, which could be combined in a single reaction with automated electrophoresis for fragment separation, resulting in a dramatic decrease in sequencing costs. The first commercially available sequencer using this chemistry was the Applied Biosystems (current Life Technologies) Model 370, introduced in 1985, which allowed parallel sequencing of up to 16 DNA samples with a length of 300 base pairs, delivering throughput of up to 5000 base pairs per run. In the next decade, further innovation allowed longer sequencing reads with an improved and simplified termination reaction and even higher parallelization of the fragment separation and detection. These advances led to sequencers capable of processing 96 samples at a time for up to 500 base pairs, like the ABI PRISM Model 377 Automated Sequencer introduced in 1995. These rates were still not economical for sequencing large mammalian genomes; however, it allowed the sequencing of viral genomes as the first example of independently living organisms (5) and even the sequencing of the first smaller eukaryotic genomes (6). Increased availability of sequencing also streamlined the discoveries of novel genes through the high throughput sequencing of complementary DNA clones (7).

In the early 2000s, automated capillary-based sequencers allowed affordable throughput of half a million bases per day per instrument. Using these machines, two independent groups, the International Human Genome Sequencing Consortium and Celera Genomics, were able to publish the first draft of the human genome (8,9) in February 2001. Shortly after, the first drafts of many other larger plant and vertebrate genomes, including rice (10,11), mouse (12), chimpanzee (13), rat (14), chicken (15), and dog (16) were published, giving insight into the organization and evolution of complex

genomes. Though de novo sequencing or re-sequencing of full genomes was still expensive and possible only in large sequencing centers or consortia, relatively large sequencing projects, including amplicon re-sequencing (17), expressed sequence tag sequencing (18), or high throughput discovery of novel small RNAs (19), was within reach for smaller laboratories and institutes.

The birth of next-generation sequencing

While automated sequencers that could handle 384 samples at a time and complex robotic setups for sample management were pushing the price per sequenced base down, a major burden on the capacity and economy of Sanger technology was the need to treat and sequence each sample individually. Therefore, a logical place for improvement in DNA sequencing was the parallelized sequencing of thousands or even millions of individual DNA fragments. The first commercially available next-generation sequencer was introduced in 2005 by 454 Life Sciences and allowed highly parallelized pyrosequencing of hundreds of thousands of individual DNA fragments, clonally amplified on the surface of beads and immobilized in a picotiter plate (20).

As the result of the affordable throughput of tens of mega bases per sequencing run and read length of hundreds of bases – long enough to perform de-novo assembly – the rate at which new genomes were being published was accelerated, including genomes of already extinct mammals (21) and hominine (22). From a simplified historical perspective, the sequencing throughput of the 454 system was an intermediate between the Sanger sequencing and the next-generation sequencers (NGS) introduced later in 2006 (Solexa by Illumina) and 2007 (SOLiD by ABI). Both systems were characterized by ultra high throughput with tens of millions DNA fragments sequenced per run with a read length of 25-35 base pairs at the time of commercial release. Despite the lower throughput and higher cost per base, the 454 system is still widely used for de-novo sequencing and the assembly of genomes (23), RNA sequencing (24), and amplicon re-sequencing (25) thanks to its relatively long read lengths (up to 1000 base pairs on recent models).

Next-generation sequencing today

The throughput of NGS has continually increased so that now it is routine to achieve a billion sequenced tags per sequencing run, as well as longer sequencing reads on the newest models, resulting in hundreds of billions of raw bases produced per sequencing run, which typically takes about a week of analysis time (Table 1). This opens numerous possibilities and allows applications that were unthinkable a decade ago for an affordable price (Table 2). Since its introduction in 2005, the numbers of publications mentioning NGS related terms have increased dramatically every year (Figure 1) and this popularity resembles the boom in microarray technology of the late 1990s. In the initial years, when the major obstacle for NGS was the complexity of NGS technique itself researchers mostly struggled with the development and application of new tools. Nowadays, NGS is a standard analytical process and has become the method of choice in many scientific fields, where it outcompetes standard methods. For example, the cost of sequencing a chromatin immunoprecipitation (ChIP) sample is estimated between 200-1000 Euros and delivers a genome-wide high resolution map of protein-DNA interaction events. This price is comparable to that of quantitative polymerase chain reaction (qPCR) analysis, which allows the assaying of only several binding events and requires *a priori* knowledge of the locations of protein-DNA interactions. As another practical example, the re-sequencing of the coding portion of

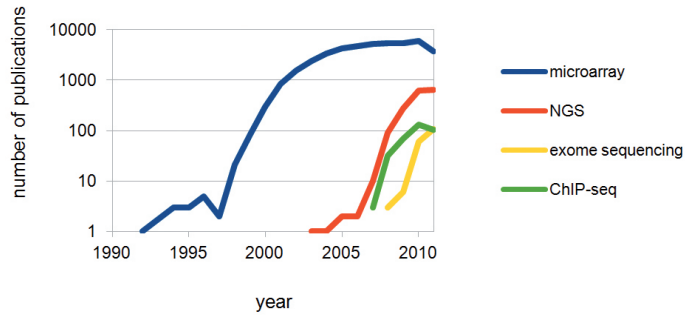


Figure 1. Number of publications in PubMed database with specific term in title or abstract. For year 2011, status on July 1 is depicted.

2000 genes in 200 individuals using Sanger sequencing would require an investment of several million Euros. The same task can be accomplished for approximately 30,000 Euros when the most recent sequencers and multiplexing technologies (26) are used.

Major applications of NGS

Whole genome sequencing

Though the sequencing of full genomes is still relatively expensive (Table 2), it continues to become more accessible for research and diagnostic labs while providing a comprehensive list of gnomic variants present in individual samples and opening enormous possibilities for diagnostics and treatment in clinical practice (27). Although our present ability to interpret the future impact of genomic variations and find those responsible for diseases is limited, identifying these elements has already helped us understand the complexity of many pathological states like cancer (28) and congenital diseases (29), even with the possibility for successful intervention and diagnosis (30,31).

Due to the relatively high false discovery rates of current techniques, especially those assessing structural variants, overlap between even slightly modified techniques is relatively low (32). Even more striking are the widespread batch effects (33), as found in the 1000 genomes project, a project where an emphasis on technical reproducibility has been stressed. Therefore, our current focus should be to find more standard and reproducible techniques of variant discovery that allow for meaningful comparisons of datasets produced by different centers (32).

It has become clear that within a few years, sequencing costs will not be the major bottleneck to widespread use of NGS in diagnostics. We can also expect that a technique that detects genomic variants from single nucleotides to whole chromosome events in a single experiment is on the near horizon. However, the implementation of NGS techniques into routine clinical and diagnostic practice may be a more complicated task (27), as unsolved issues about data interpretation, predictive value, and an array of ethical and legal issues together with misinformation and prejudgments among the general public and clinicians may delay the day when we enter a true personal genomics era.

Targeted re-sequencing

Despite the cost of sequencing whole genomes continuing to drop, the re-sequencing of large genomes to a depth that is sufficient for reliable variant calling is still relatively expensive (Table 2). In addition, with current knowledge we are not able to understand and interpret most of the variants in non-coding regions found by whole genome sequencing. Therefore,

Table 1: Overview of major commercially available NGS platforms.

NGS platform	Sequencing principle	Read length (fragment/mate pair/paired-end)
Roche/454 GS FLX Titanium	Pyrosequencing – based on release of pyrophosphate after nucleotide incorporation, with subsequent light emission, which is proportional to the number of incorporations. The order and intensity of light peaks corresponding to different nucleotides encodes the DNA sequence.	Up to 1000/NA/NA
Illumina/Solexa HiSeq 2000	Reversible terminator sequencing – relies on incorporation of single fluorescently labelled nucleotide, followed by washing and signal detection. After imaging, fluorescent dye is cleaved, nucleotide unblocked, and synthesized strand becomes available for next cycle of nucleotide incorporation.	100/100/100x100
Life/ABi SOLiD 5500xl System	Sequencing by ligation – uses specific hybridization of fluorescently labelled hybridization probes to primed template. Non-ligated probes are washed away and fluorescence signals, each specific to 4 of 16 combinations of dinucleotides, are detected. After set number of ligation cycles, sequencing primer is changed, allowing for sequencing of dinucleotides in subsequent reading frame.	75/60/75x35
Helicos HeliScope	Reversible terminator sequencing on single molecule template; modifications in chemistry allow for direct sequencing of both DNA and RNA.	25-55/25-55/25-55
Life/ Ion Torrent	Semiconductor technology – incorporation of specific nucleotide is coupled with pH change which is detected by semiconductor chip.	100/NA/NA
Pacific Biosciences PacBio	Four-color real time sequencing – based on direct optical observation of fluorescence signal after incorporation of single labelled nucleotide by DNA polymerase, which is immobilized at the bottom of 100 nm wide pore, allowing emitted light to travel only directly to the detector.	1000/NA/NA

Table 2: Major applications of next-generation sequencing.

Application example	Throughput needed (millions of 75-bases-long reads per sample)	Estimated costs per sample[3] (July 2011, €)	Discussed or used in this thesis
Whole genome re-sequencing	1,000[1]	10,000	Chapters 4-5
De-novo sequencing	2,3000[2]	25,000	
Targeted re-sequencing of whole human exome	100-200	1000-2000	Chapter 5
Targeted re-sequencing of selected genes (2000 genes)	10	150	Chapters 2-5
Transcriptome sequencing	10-50	300-600	Chapter 7
Small RNA sequencing	10-50	300-600	
ChIP-seq	5-100	200-1,000	Chapters 6-9
DNA methylation analysis	10-200	200-1500	
Metagenomics	5-500	200-4000	

[1]25× average coverage, human genome.
[2]Extrapolated from mate pair data used in sequencing of panda genome (83).
[3]Including sequencing chemicals and library preparation, without maintenance costs, personnel and other indirect costs.

Run time (days)	Max raw throughput per run (GB)	Advantages	Disadvantages	Template immobilization/clonal amplification
1	0.7	Long read lengths, short run times	Low throughput, high cost per base and run, high error rates in homopolymer repeats	Sequencing beads/emulsion PCR
2-11	600 (two flow cells)	High throughput per run, low costs per base, scalability	High cost per run, long run times	Glass surface/bridge amplification PCR
4-15	150 (two flow cells)	High throughput per run, low costs per base, sample multiplexing, scalability, intrinsic error detection using colorspace	High cost per run, long run times, lack of computational methods correctly treating colorspace data, shorter read lengths	Paramagnetic sequencing beads/emulsion PCR
8	21-35	Less biased representation of templates, direct RNA sequencing, scalability	High error rate, shorter reads, higher machine costs than other NGS platforms	Glass surface/single molecule sequencer
0.1	0.01	Very short run times, low cost per run	Low throughput, high cost per base	Sequencing beads/emulsion PCR
0.05	0.04	Very long read length, low cost per run, very short run times	High error rate, low throughput, high cost per base	Immobilised polymerase/single molecule sequencer

the re-sequencing of specifically the coding parts of the genome, particular group of genes, or genomic regions will remain the more reliable alternative in the upcoming years for most type of studies focused on discovering genomic variants (34). At the moment, there are numerous techniques for selective enrichment of targeted regions that are compatible with the throughput of NGS technologies. The most widely used simple hybridization-based techniques rely on specific hybridization of targeted regions to complementary probes, which are printed and immobilized on micro-array slides (35-38). Another widely used alternative is solution-based hybridization techniques with free biotinylated RNA or DNA probes (39). These approaches allow freedom for custom selection of targeted regions with a capacity ranging from several genes to whole exomes. Principles like those using microfluidics based multiplex-PCR technology (RainDance) (40) or rolling circle amplification of endonuclease-digested DNA mediated by specifically hybridized probes (Selector probes) (41) lack the capacity to enrich for a large number of genes or whole exomes. However, thanks to straightforward and highly efficient enrichment procedures and simple parallelization, such principles are suitable for routine diagnostics of a limited number of genes.

At this time, all enrichment techniques can be seen as temporary solutions until the cost for whole genome sequencing becomes comparable to the cost of actual enrichment procedures or clear benefits of whole genome sequencing from the point of data interpretation are shown. From an economical standpoint, with the current pace of price reduction for sequencing and after the introduction of techniques for multiplexing of enrichment procedures (26), which dramatically reduce enrichment costs, this switch may not happen for 3-5 years. However, the academic reasons for whole genome sequencing (more reliable variant calling, elimination of enrichment biases, and the possibility of detecting structural rearrangements) may force us to do it in less time.

Detection of structural variation

Variations among individual human genomes range from differences in single nucleotides to whole chromosomes. In recent years, it has become clear that this variation is higher than initially expected with an increasing proportion of structural variations (SVs) responsible for this variability compared to single nucleotide events (32). Structural variation can be characterized as events affecting >50 base pairs (32) including insertions, deletions, transversions, duplications, and translocations where often each single variation is a combination of two or more types (e.g., transversion is often coupled with deletions or insertions at the breakpoints, etc.). Though widely used karyotyping and arrayCGH analysis can be used to discover large SVs (in the case of arrayCGH only unbalanced SVs), the vast majority of smaller or balanced SVs are out of the detection limit of those techniques. The introduction of NGS allows more reliable detection of SVs with single nucleotide resolution and, through paired-end or mate-pair techniques along with approaches using long reads, we can now precisely map even complex SVs (29,42). Although NGS techniques provide revolutionary insights into SVs, they suffer from high false negative detection rates and low overlap between different sub-methods (32). This becomes even more problematic when analyzing mixed cell populations like tumor samples. In addition, there is no available method and analysis algorithm that would address an assortment of different SV types in a single assay. In the future, methods based on sequencing several sequencing paired-end and mate-pair libraries from the same sample with incrementing insert length coupled with longer read lengths and analysis software able to integrate information from different insert lengths may be a solution to this obstacle.

Metagenomics

Metagenomics allows culture-free identification of microorganisms in various types of samples, ranging from various environmental specimens to human microflora, by identification and quantification of their genetic signature through sequencing (43). The most widely used technique relies on sequencing a hyper-variable section of the 16S rDNA subunit, which is amplified using universal sets of primers that are cloned and sequenced using a classical Sanger-based sequencer or 454 sequencing. However, 16S-based metagenomics often lacks precision and reproducibility, depending on the part of the hyper-variable region used, and does not allow for closer sub-strain identification of bacteria as is within a reach when genome sequencing techniques are used (44). In the future, when reliable computational approaches for identification of different bacterial species through whole genome sequencing of mixed bacterial populations become available, metagenomics may be widely used not only in research, but also in clinical diagnostics, delivering a more comprehensive picture of the microorganisms present in a sample.

RNA sequencing

Genome-wide quantitative and qualitative analysis of transcriptomes has given valuable biological insight on the regulation of gene expression (45). Before the NGS era, the genome-wide analysis of transcriptomes relied on hybridization techniques and although micro-arrays are still the most frequently used technique, they deliver only a limited view of the transcriptome complexity and dynamics and are biased towards *a priori* knowledge of expression patterns. On the other hand, RNA-sequencing techniques provide a more complex strand-specific view of transcribed regions, with virtually unlimited dynamic range (46), and allow for hypothesis-free insights of transcriptional mechanisms. Shortly after the introduction of RNA-seq techniques, several fundamental mechanisms, including the wide spread existence of anti-sense transcription (47-49), the high frequency of RNA-editing (50), and the high dynamics of alternative splicing (51) were discovered, thanks to the complete and unbiased information encoded in RNA-seq datasets. In the future, proper integration of RNA-sequencing data and information from proteome analysis, ChIP-seq, and direct transcription rate measurements will provide a complete view of the mechanisms of gene expression; however, the development of effective methods for integrative data analysis and interpretation will remain a major challenge. In addition, long sequencing times and a lack of user-friendly computational tools that allow "non-experts" in bioinformatics to process, visualize, and analyze RNA-seq data in a way that is similar to micro-array experiments are barring the way for wider and more routine use of RNA-seq.

The sequencing of small non-coding RNAs is a particular area of RNA sequencing, due to the different approaches and strategies used in sample preparation (ineffective random priming and short nature) and data analysis, where the impact of NGS technologies has been exceptional (45), particularly for the discovery of novel small RNA genes and classes and understanding the underling biology (52,53). Although small RNA sequencing can provide an accurate view of the presence of different small RNAs and their dynamics, due to the nature and biases of sample preparation, it will remain challenging to extract information about absolute abundances of specific small RNAs (54).

ChIP-seq

Genome-wide identification of protein-DNA interactions is an essential tool in our aim to understand the mechanisms of transcriptional regulation. The most widely used technique for this purpose is ChIP, in which cross-linked DNA-protein complexes are immunoprecipitated using antibody raised against the protein of interest resulting in a chromatin sample with overrepresented DNA fragments coming from a specific region of DNA-protein interaction (55). Before the introduction of NGS, specific enrichment was measured by qPCR, allowing examination of only a few pre-selected candidate binding regions with limited resolution and requiring *a priori* knowledge of the genomic location of regulatory elements. Later, with dedicated micro-arrays it became possible to perform genome-wide analysis of ChIP samples (ChIP-chip) (56); however, extremely costly experiments that needed dozens of densely tiled micro-array slides (57) delivered only low resolution maps (approximately 100 base pairs) of binding regions without reliable quantitative information about the strength of binding sites due to the limited dynamic range of micro-arrays (58). After the introduction of NGS, ChIP samples can now be sequenced directly, resulting in high resolution (virtually single base), accurate, and quantitative genome-wide maps of protein-DNA interactions for only a fraction of the cost compared to whole genome ChIP-chip experiments (59-61). With the availability

of antibodies against a wide range of regulatory proteins and histone marks, databases of protein-DNA interactions for many different elements in various model systems and under multiple experimental conditions can now be built. Combined with information about RNA and protein levels, this integrative analysis may reveal novel mechanisms in transcriptional regulation. The biggest challenges in the coming years will be the ability to perform proper ChIP-seq on small quantities of (primary) material and the development of models for data analysis and the integration of different datasets (58).

Third-generation sequencing technologies

Naturally, the development of current NGS technologies is not the ultimate goal, and even within the bounds of its limitations there is room for novel innovations. A major common focus in the so-called third-generation sequencing (TGS) technologies is the development of a device for the direct sequencing of single DNA or RNA molecules with long read length, without the need for any extensive sample preparation, which is often a source of technical artifacts, and with the capacity to deliver results in few hours for relatively low costs per run. All known technologies with the potential to bring commercial TGS instruments to market rely on real time sequencing of single DNA molecules; however, they differ in the basic principles of signal detection. Pacific Biosciences uses direct optical observation of fluorescence signal after incorporation of a single labeled nucleotide by DNA polymerase, which is immobilized at the bottom of a 100-nm-wide pore, allowing emitted light to travel only directly to the detector. One limitation of this technology is the relatively small capacity of photo sensors, allowing sequencing of only up to 75,000 DNA molecules at a time. Another is the potential for photo bleaching of the single polymerase molecule (62,63). Other technologies do not use optics for detection, such as Oxford Nanopore technology (64), which relies on changes of electric current running through a protein nanopore after the translocation of a cleaved nucleotide, which is electronically detected on a chip. In another method developed by IBM, a single-stranded DNA is electrically pulled through a DNA transistor (65), where each base differently modulates the electric current in the semiconductor pore. Although it is difficult to predict which technology will become dominant and will most closely meet the expectations, we can already anticipate that thanks to the low costs per run, the available bench-top form, short run times, and simple sample preparation, TGS will allow for simple everyday use in research and diagnostics laboratories for applications like ChIP-seq, RNA-seq, or targeted re-sequencing where short turnaround times might be more important than extremely high coverage. Indeed, just as NGS did not fully replace Sanger sequencing, these new technologies will be complementary to current NGS due to the expected lower throughput and higher cost per base of TGS.

Future technical challenges in NGS

Current NGS sequencing data suffer from relatively high error rates compared to Sanger sequencing (66). The major cause of the error lies in fundamental base calling; however, the misalignments to reference genomes and the mistakes caused by DNA polymerases are also source of errors. Unfortunately, these errors are often systematic and occur more often in certain nucleotide combinations; therefore, this must be considered when analyzing and interpreting NGS data, as in studies of RNA editing (50) or large re-sequencing projects (67) with relatively low sequencing coverage per individual. Although this impacts all sequencing applications, it mostly affects variant calling from the sequencing data. Future emphasis has to be put on finding ways to decrease

the error rates of NGS technologies and in finding better computational tools for alignment and variant calling to minimize false discovery rates. In addition, it is important to implement widely accepted standards for variant calling to allow for reliable comparisons of datasets produced in different places.

Although the throughput of NGS sequencers has increased dramatically, the length of sequencing fragments, which can be sequenced, did not follow the same trend. Currently, the most used technologies can deliver 75-150 base reads. However, for certain applications, like whole genome assembly, RNA sequencing, and metagenomics, longer reads would be beneficial. The current nature of NGS chemistry does not allow for a tremendous increase in read-length (Solexa/Illumina, SOLiD/Life Technologies) or is not high throughput (454/Roche). Other alternatives, like those from Pacific Biosciences (62) or nanopore-based sequencing (64), may be capable of longer sequencing reads; however, the throughput of these technologies will probably be low compared to current NGS instruments with higher error rates.

With the unprecedented drop in sequencing costs, NGS machines are capable of producing enormous amounts of data, putting higher demands each year on the computational power for data analysis, storage, and sharing. Though, according the Moore's law, computer power, storage, or network capacity doubles every 2 years without increasing price, the drop in sequencing costs is much faster (Figure 2), resulting in a relative increase in the cost of the computational power necessary for data handling. This makes expenses for computing power a significant part of the budget necessary to run an NGS laboratory, with an increasing trend (34). For example, a sequencing run of a human individual, delivering 1 billion sequencing reads, needs approximately 6,000 central processing unit (CPU) hours just to map the results against a human reference genome, which is only the beginning of the computational analysis. With the current price of CPU hours from large computer clouds (provided by Amazon, Microsoft, or Google), this would equal approximately $600, not including the data transfer, just for this initial computational task, while outsourcing the computational power.

Therefore, in the coming years it will be necessary to find better computational solutions to address the most commonly used and computationally intensive tasks related to NGS data analysis. One possibility is to use graphical processing units, which are optimized to run multiple processes in parallel, for the alignment of sequencing reads to reference genomes. This has already been implemented in the newest versions of mapping algorithms like SOAP (68). Another possibility is to use grid computing by using the free capacity of personal computers within the networks of an institution (69). Problems with the network capacities for sharing large datasets can possibly be solved by more compact and flexible data formats, with the maximal compression rate containing only the information necessary for the desired analysis.

Future biological challenges in NGS

Though the technical limitations mentioned above may require substantial efforts to solve, NGS brings even more complex analytical and biological challenges.

The first concerns correct data interpretation. With the first whole genome sequencing and exome studies, it became clear that even a healthy individual carries an unexpectedly high number of potentially damaging mutations altering the functions of affected genes, together with numerous structural variations without clear phenotypical correlations (70-72). This finding is more striking from the perspective of similar analyses of patients with specific diseases, which

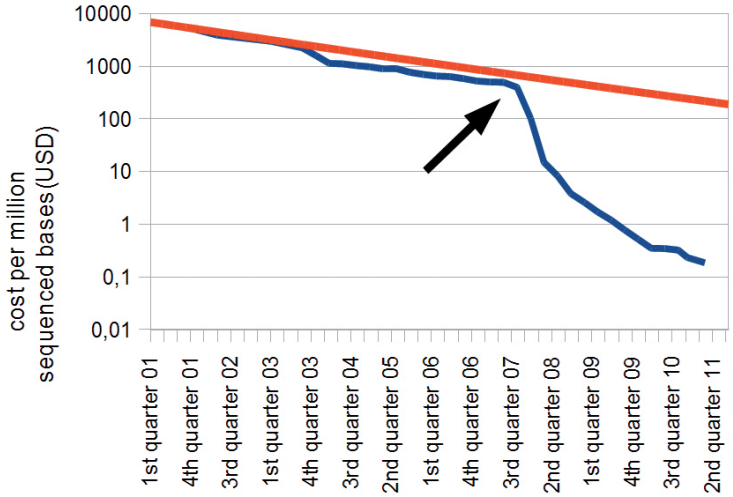


Figure 2. Dramatic drop of sequencing price after introduction of NGS (blue line) in 2007 depicted by black arrow (84). Red line simulates drop of the cost as it follows Moore's law.

would probably identify similar mutations as causative ones, while the true cause may lie elsewhere, such as a mutated regulatory element or missed genomic rearrangement. Therefore, careful validation of NGS results using model systems or more comprehensive analysis tools with standards requiring multiple levels of evidence is needed. Similar problems may arise after cancer genomes are sequenced for diagnostic or personalized treatment purposes. The high heterogeneity of tumors and their genetic differences from metastases (73), along with overwhelming numbers of mutations possibly altering protein functions may cause difficulties in deciding which targeted treatment has the highest chance for success. In addition, due to the high number treatment options and high genetic variability even within a single tumor type, it might be necessary to change the way clinical trials and drug approval procedures are performed, since the best treatment combinations may be unique for each patient, preventing the testing of particular drug combinations in a classical clinical trial.

The second challenge concerns variability and reproducibility of results generated by NGS. The many striking examples of batch effects (33) with subsequent catastrophic impact on conclusions derived from those datasets (32) might be only the tip of an iceberg. Maximal care must be taken to critically verify the quality of datasets, especially of those collected from different sources specifically for the purpose of integrative analysis (74). Widely accepted standards for data processing, normalization, and analysis must be established for each NGS application. However, due to the differing natures of biological samples and various biological questions, it may be problematic to find a "single standard fits all" solution. For example, demands on starting material, necessary depth of sequencing, and subsequent analysis will be different for ChIP-seq experiments using a high-quality antibody against a widely spread histone mark compared ChIP-seq of a transiently induced, tagged transcription factor. Other demands on sequence coverage, false positive, and false negative rates vary for diagnostic whole genome sequencing studies

focused on the discovery of genomic variants in patient with genetic diseases compared to the same experiment performed with the aim of lineage tracing of the tumor and its metastases. In both cases, datasets produced by different standards might be both valid and sufficient to answer a particular biological question; however, uncritical integration and use in single analyses might lead to incorrect conclusions (74).

The third challenge is proper data integration from different type of experiments necessary for a wider understanding of complex biological systems (74). Numerous datasets including ChIP-seq, whole genome sequencing, RNA-sequencing, or protein interaction screens provide unique information about biological systems on their own; however, their integration would provide a more complex picture. For example, the dynamics of ChIP-sequencing data together with information about the dynamics of RNA levels after a specific intervention would give a more complete picture of the basic mechanisms of transcriptional regulation (75) than either single dataset alone, as described in Chapter 7 of this thesis. Another example shows how the proper combination of ChIP-seq data, micro-array data, and gene ontology information can decouple direct and indirect mechanisms of TF in the regulation of target genes (76). Proper data integration still represents a difficult task, mainly due to the need for extensive and complex bio-informatical approaches. For "wet lab" scientists, this can be a major obstacle (74). Although user-friendly software solutions are available (77-82), they often address only particular types of datasets and biological questions and all of them require larger promotion among the scientific audience before becoming a standard part of a scientist's toolbox. Most importantly, we have to realize that even the best data integrating software will never be a magic wand with an "analyze and write paper" button, which takes datasets on one end and outputs results hypothesis and principles on the other end, but instead these tools will always need clever input in the form of specific scientific biological questions.

OUTLINE OF THIS THESIS

In **Chapter 2**, we show that the use of shorter sequencing libraries together with increased read length improves the performance of genomic enrichment. The strength of our method was not only the best enrichment efficiency achieved at that time, but more importantly, the improved evenness of coverage through the targeted region with a more even distribution of sequencing reads on the positive and negative strands. In particular, the even distribution of reads on the positive and negative strands was the most important parameter for variant discovery and allowed us to achieve low false positive and false negative rates.

Chapter 3 reports a new strategy for the targeted re-sequencing of multiple samples. Through the re-sequencing of 770 genes in 30 N-ethyl-N-nitrosourea-mutagenized rats, we demonstrated the possibility of pooling multiple samples before genomic enrichment in a single assay. Before that, a separate enrichment of every sample was necessary before sequencing. Using our method, we were able to dramatically cut the price of targeted re-sequencing, especially for studies using a high number of samples and a candidate gene approach.

In **Chapter 4**, we present the published practical protocol for multiplexed targeted re-sequencing. Through this protocol, we demonstrate the effectiveness of our method on a wide range of applications starting from the enrichment and re-sequencing of 96 patients in a single assay for a limited number of genes to the full human exome projects.

In **Chapter 5**, we introduce the "fast-forward genetics" approach for identification of candidate variants in forward genetics studies. We successfully applied this two-step approach, which combines the traditional bulk segregant technique with the targeted genomic enrichment and next-generation sequencing technology, to two Arabidopsis mutants and identified a novel factor required for stem cell activity.

In **Chapter 6**, we demonstrate the addition of a second round of DNA fragmentation after de-cross-linking as beneficial for ChIP-seq procedures. This method helps generate sequencing libraries from the immunoprecipitated material, the quantities of which would normally not be sufficient, using the longer immunoprecipitated fragments that would normally be too long for sequencing.

In **Chapter 7**, we introduce RNA Polymerase II (Pol II) ChIP-seq as a reliable and versatile measure for gene activity. By integrating Pol II ChIP-seq, micro-array, and RNA-seq data, we were able to identify functional groups of genes with different mechanisms of expression regulation.

In **Chapter 8**, using the technique from Chapter 6, we were able to reproducibly perform ChIP-seq on β-catenin and TCF/LEF family members and identified two classes of β-catenin-bound genomic elements.

REFERENCES

1. Dahm, R. (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet, 122, 565-581.
2. Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A, 74, 560-564.
3. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74, 5463-5467.
4. Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M. and Smith, M. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature, 265, 687-695.
5. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269, 496-512.
6. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. et al. (1996) Life with 6000 genes. Science, 274, 546, 563-547.
7. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252, 1651-1656.
8. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. et al. (2001) The sequence of the human genome. Science, 291, 1304-1351.
9. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001) Initial sequencing and analysis of the human genome. Nature, 409, 860-921.
10. Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science, 296, 92-100.
11. Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science, 296, 79-92.
12. Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature, 420, 520-562.
13. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 437, 69-87.
14. Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E. et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature, 428, 493-521.
15. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 432, 695-716.
16. Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., 3rd, Zody, M.C.

et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature, 438, 803-819.

17. Smits, B.M., Mudde, J., Plasterk, R.H. and Cuppen, E. (2004) Target-selected mutagenesis of the rat. Genomics, 83, 332-334.
18. Guryev, V., Berezikov, E., Malik, R., Plasterk, R.H. and Cuppen, E. (2004) Single nucleotide polymorphisms associated with rat expressed sequences. Genome Res, 14, 1438-1443.
19. Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R.H. and Cuppen, E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. Cell, 120, 21-24.
20. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376-380.
21. Miller, W., Drautz, D.I., Ratan, A., Pusey, B., Qi, J., Lesk, A.M., Tomsho, L.P., Packard, M.D., Zhao, F., Sher, A. et al. (2008) Sequencing the nuclear genome of the extinct woolly mammoth. Nature, 456, 387-390.
22. Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H. et al. (2010) A draft sequence of the Neandertal genome. Science, 328, 710-722.
23. Miller, J.R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. Genomics, 95, 315-327.
24. Adamidi, C., Wang, Y., Gruen, D., Mastrobuoni, G., You, X., Tolle, D., Dodt, M., Mackowiak, S.D., Gogol-Doering, A., Oenal, P. et al. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res, 21, 1193-1200.
25. Kohlmann, A., Klein, H.U., Weissmann, S., Bresolin, S., Chaplin, T., Cuppens, H., Haschke-Becher, E., Garicochea, B., Grossmann, V., Hanczaruk, B. et al. (2011) The Interlaboratory RObustness of Next-generation sequencing (IRON) study: a deep sequencing investigation of TET2, CBL and KRAS mutations by an international consortium involving 10 laboratories. Leukemia.
26. Nijman, I.J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E. and Cuppen, E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods, 7, 913-915.
27. Pasche, B. and Absher, D. (2011) Whole-genome sequencing: a step closer to personalized medicine. JAMA, 305, 1596-1597.
28. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet, 11, 685-696.
29. Kloosterman, W.P., Guryev, V., van Roosmalen, M., Duran, K.J., de Bruijn, E., Bakker, S.C., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M. et al. (2011) Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum Mol Genet, 20, 1916-1924.
30. Welch, J.S., Westervelt, P., Ding, L., Larson, D.E., Klco, J.M., Kulkarni, S., Wallis, J., Chen, K., Payton, J.E., Fulton, R.S. et al. (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA, 305, 1577-1584.
31. Link, D.C., Schuettpelz, L.G., Shen, D., Wang, J., Walter, M.J., Kulkarni, S., Payton, J.E., Ivanovich, J., Goodfellow, P.J., Le Beau, M. et al. (2011) Identification of a novel TP53 cancer susceptibility mutation through whole-genome sequencing of a patient with therapy-related AML. JAMA, 305, 1568-1576.
32. Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. Nat Rev Genet, 12, 363-376.
33. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K. and Irizarry, R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet, 11, 733-739.
34. Metzker, M.L. (2010) Sequencing technologies - the next generation. Nat Rev Genet, 11, 31-46.
35. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 4, 907-909.
36. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods, 4, 903-905.
37. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet, 39, 1522-1527.
38. Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V. and Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res, 38, e116.
39. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol, 27, 182-189.

40. Kiss, M.M., Ortoleva-Donnelly, L., Beer, N.R., Warner, J., Bailey, C.G., Colston, B.W., Rothberg, J.M., Link, D.R. and Leamon, J.H. (2008) High-throughput quantitative polymerase chain reaction in picoliter droplets. Anal Chem, 80, 8975-8981.
41. Johansson, H., Isaksson, M., Sorqvist, E.F., Roos, F., Stenberg, J., Sjoblom, T., Botling, J., Micke, P., Edlund, K., Fredriksson, S. et al. (2011) Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res, 39, e8.
42. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A. et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell., 144, 27-40.
43. Streit, W.R. and Schmitz, R.A. (2004) Metagenomics--the key to the uncultured microbes. Curr Opin Microbiol, 7, 492-498.
44. Petrosino, J.F., Highlander, S., Luna, R.A., Gibbs, R.A. and Versalovic, J. (2009) Metagenomic pyrosequencing and microbial identification. Clin Chem, 55, 856-866.
45. Ozsolak, F. and Milos, P.M. (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet, 12, 87-98.
46. Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10, 57-63.
47. Faghihi, M.A. and Wahlestedt, C. (2009) Regulatory roles of natural antisense transcripts. Nat Rev Mol Cell Biol, 10, 637-643.
48. Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A. and Sharp, P.A. (2008) Divergent transcription from active promoters. Science, 322, 1849-1851.
49. Taft, R.J., Glazov, E.A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G.J., Lassmann, T., Forrest, A.R., Grimmond, S.M., Schroder, K. et al. (2009) Tiny RNAs associated with transcription start sites in animals. Nat Genet, 41, 572-578.
50. Li, M., Wang, I.X., Li, Y., Bruzel, A., Richards, A.L., Toung, J.M. and Cheung, V.G. (2011) Widespread RNA and DNA sequence differences in the human transcriptome. Science, 333, 53-58.
51. Carninci, P. (2009) Is sequencing enlightenment ending the dark age of the transcriptome? Nat Methods, 6, 711-713.
52. Rajagopalan, R., Vaucheret, H., Trejo, J. and Bartel, D.P. (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev, 20, 3407-3425.
53. Ruby, J.G., Jan, C., Player, C., Axtell, M.J., Lee, W., Nusbaum, C., Ge, H. and Bartel, D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell, 127, 1193-1207.
54. Linsen, S.E., de Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R.K., Fritz, B., Wyman, S.K., de Bruijn, E., Voest, E.E. et al. (2009) Limitations and possibilities of small RNA digital gene expression profiling. Nat Methods, 6, 474-476.
55. Solomon, M.J., Larsen, P.L. and Varshavsky, A. (1988) Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell, 53, 937-947.
56. Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. et al. (2000) Genome-wide location and function of DNA binding proteins. Science, 290, 2306-2309.
57. Hatzis, P., van der Flier, L.G., van Driel, M.A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W. et al. (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cell Biol, 28, 2732-2744.
58. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 10, 669-680.
59. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science, 316, 1497-1502.
60. Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. Cell, 129, 823-837.
61. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods, 4, 651-657.
62. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. et al. (2009) Real-time DNA sequencing from single polymerase molecules. Science, 323, 133-138.
63. Schadt, E.E., Turner, S. and Kasarskis, A. (2010) A window into third-generation sequencing. Hum Mol Genet, 19, R227-240.
64. Branton, D., Deamer, D.W., Marziali, A., Bayley, H., Benner, S.A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X. et al. (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol, 26, 1146-1153.
65. Luan, B., Peng, H., Polonsky, S., Rossnagel, S., Stolovitzky, G. and Martyna, G. (2010) Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. Phys Rev Lett, 104, 238103.
66. Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet, 12, 443-451.

67. (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061-1073.
68. Li, R., Yu, C., Li, Y., Lam, T.W., Yiu, S.M., Kristiansen, K. and Wang, J. (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, 25, 1966-1967.
69. Schadt, E.E., Linderman, M.D., Sorenson, J., Lee, L. and Nolan, G.P. (2010) Computational solutions to large-scale data management and analysis. Nat Rev Genet, 11, 647-657.
70. Kim, J.I., Ju, Y.S., Park, H., Kim, S., Lee, S., Yi, J.H., Mudge, J., Miller, N.A., Hong, D., Bell, C.J. et al. (2009) A highly annotated whole-genome sequence of a Korean individual. Nature, 460, 1011-1015.
71. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. et al. (2008) The diploid genome sequence of an Asian individual. Nature, 456, 60-65.
72. McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C. et al. (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Res, 19, 1527-1541.
73. Vermaat, J.S., Nijman, I.J., Koudijs, M.J., Gerritse, F.L., Scherer, S., Mokry, M., Roessing, W., van Diest, P.J., Cuppen, E. and Voest, E.E. (submitted) Primary tumors and their metastasis are genetically different: implications for selection of patients for targeted treatment.
74. Hawkins, R.D., Hon, G.C. and Ren, B. (2010) Next-generation genomics: an integrative approach. Nat Rev Genet, 11, 476-486.
75. Mokry, M., Hatzis, P., Schuijers, J., Lansu, N., Ruzius, F.P., Clevers, H. and Cuppen, E. (submitted) Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady state RNA levels identifies differentially regulated functional gene classes.
76. Westendorp. B, Mokry, M., Koerkamp, M.G., Holstege, F.C., Cuppen, E. and de Bruin, A. (submitted) E2F7 represses a transcriptional network of oscillating cell cycle genes to control S-phase progression.
77. Rosenbloom, K.R., Dreszer, T.R., Pheasant, M., Barber, G.P., Meyer, L.R., Pohl, A., Raney, B.J., Wang, T., Hinrichs, A.S., Zweig, A.S. et al. (2010) ENCODE whole-genome data in the UCSC Genome Browser. Nucleic Acids Res, 38, D620-625.
78. Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A. et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res, 39, D876-882.
79. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, 25-29.
80. Taylor, J., Schenck, I., Blankenberg, D. and Nekrutenko, A. (2007) Using galaxy to perform large-scale interactive data analyses. Curr Protoc Bioinformatics, Chapter 10, Unit 10 15.
81. Blankenberg, D., Taylor, J., Schenck, I., He, J., Zhang, Y., Ghent, M., Veeraraghavan, N., Albert, I., Miller, W., Makova, K.D. et al. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. Genome Res, 17, 960-964.
82. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M. and Wong, W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 26, 1293-1300.
83. Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y. et al. (2010) The sequence and de novo assembly of the giant panda genome. Nature, 463, 311-317.
84. Wetterstrand, K. (July 2011).

2

ACCURATE SNP AND MUTATION DETECTION BY TARGETED CUSTOM MICROARRAY-BASED GENOMIC ENRICHMENT OF SHORT-FRAGMENT SEQUENCING LIBRARIES

Mokry M, Feitsma H, Nijman IJ, de Bruijn E, van der Zaag PJ, Guryev V, Cuppen E: **Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries.** *Nucleic Acids Res* 2010, **38**(10):e116.

ABSTRACT

Microarray-based enrichment of selected genomic loci is a powerful method for genome complexity reduction for next-generation sequencing. Since the vast majority of exons in vertebrate genomes are smaller than 150 nt, we explored the use of short fragment libraries (85–110 bp) to achieve higher enrichment specificity by reducing carryover and adverse effects of flanking intronic sequences. High enrichment specificity (60–75%) was obtained with a relative even base coverage. Up to 98% of the target-sequence was covered more than 20× at an average coverage depth of about 200×. To verify the accuracy of SNP/mutation detection, we evaluated 384 known non-reference SNPs in the targeted regions. At ~200× average sequence coverage, we were able to survey 96.4% of 1.69 Mb of genomic sequence with only 4.2% false negative calls, mostly due to low coverage. Using the same settings, a total of 1197 novel candidate variants were detected. Verification experiments revealed only eight false positive calls, indicating an overall false positive rate of less than 1 per ~200 000 bp. Taken together, short fragment libraries provide highly efficient and flexible enrichment of exonic targets and yield relatively even base coverage, which facilitates accurate SNP and mutation detection. Raw sequencing data, alignment files and called SNPs have been submitted into GEO database http://www.ncbi.nlm.nih.gov/geo/ with accession number GSE18542.

INTRODUCTION

The need for detection of SNPs and mutations in large genomic segments is increasing rapidly, partially as a result of genome-wide association studies (GWAS) that have pinpointed many genomic loci of interest for specific diseases and disease susceptibilities (1-5). Furthermore, the vastly increased throughput of massively parallel next-generation sequencing technologies enables the interrogation of unprecedented numbers of genes in a single analysis, permitting, for instance, the investigation of the complete protein-coding transcriptome (6). Enrichment of genomic loci by microarray hybridization followed by massively parallel sequencing has become an important method for targeted re-sequencing. The approach is based on hybridization of fragmented and adapter-ligated DNA to capturing probes printed on microarray slides (7-11), present in solution (12,13) or PCR products immobilized on filters (14) that are specifically designed for the regions of interest. After hybridization, non-targeted fragments are washed away and only the captured fragments are eluted for deep sequencing on any of the next-generation sequencing platforms (7-10). Generally, DNA fragment libraries with 500-bp fragment size are recommended and used for optimal efficiency because shorter fragments were reported to increase the number of off-target reads (8). However, many re-sequencing projects are focused on exons of protein coding genes. Because the median size of a human exon is only 120 bp (with 70% of all exons shorter than 200 bp) (Figure 1), long-fragment libraries can have severe limitations. Most importantly, many of the specifically captured DNA from long-fragment libraries will consist of sequences derived from introns flanking the exons of interest, which decreases the effective sequencing yield. To address this issue, we have explored the efficiency of enrichment of DNA fragment libraries with much

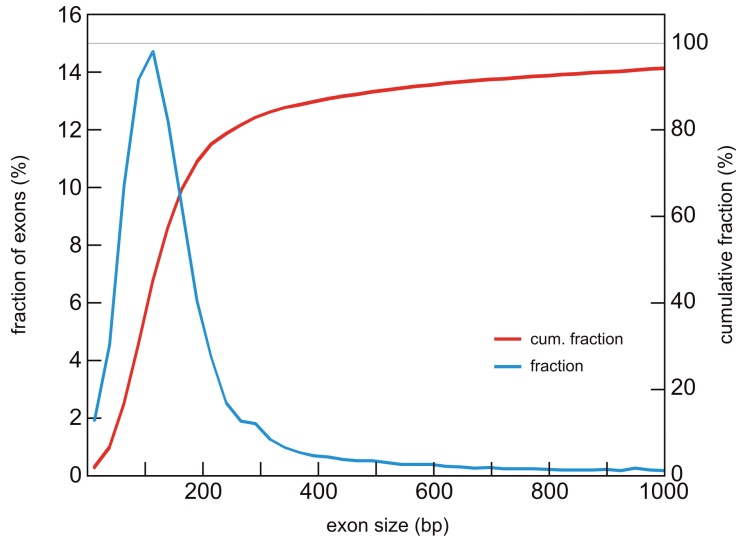


Fig 1: Size distribution of human exons. The median size of human exons is only 120 bp (with 70% of all exons shorter than 200 bp). Therefore many of the specifically captured DNA from long-fragment libraries will consist of sequences derived from introns flanking the exons of interest, which decreases the effective sequencing yield.

shorter fragments (85–110 bp). We used human genomic DNA and an exon-centric capturing design on commercially available custom microarray slides. We developed an effective probe design strategy with tiled oligonucleotides for capturing both genomic strands, resulting in targeting of 1.69 Mb of exonic sequences on a 244 K array. To validate the performance of the microarray-based enrichment strategy, we have tested the specificity of enrichment and evaluated the effectiveness of retrieval of known SNPs that were present in the targeted regions. When applying highly stringent filtering criteria, we were able to evaluate 93% of the target regions without any false-negatives. Moreover, independent verification experiments on newly discovered polymorphisms revealed an overall false-positive rate of less than 1 per 200 kb. These results indicate that the use of short fragment libraries results firstly in highly efficient enrichment with relatively even base coverage over the targeted region and that secondly the sensitivity and specificity of SNP/mutation calling is very high.

MATERIALS AND METHODS

Enrichment array design

Exonic sequences of 1621 genes were selected from hg18 build of the human genome spanning a total length of targeted sequence of 1.69 Mb. Genes were selected based on the potential presence of exonic SNPs with presumed functional effects and thus potential clinical relevance (15). Sixty nucleotide long probes were designed with an average tilling density of 10 bp for both negative and positive strands. The probe selection strategy was set using the following rules: all possible 60-mer probes starting in a 10 bases long window were collected and a single probe with the lowest penalty score (see below) was selected. This procedure was repeated for every 10 nt bin in the region of interest, which presented all coding exons of selected genes. Penalty scores were calculated as follows: 4 points if T_m is <77°C or >81°C with T_m defined as:

$$Tm = (64.9 + 41 \times (n_G + n_C - 16.4)/N) \times ^{\circ}\mathrm{C} \qquad (1)$$

where n_G is total number of guanidines, n_C is total number of cytosines and N represents oligonucleotides length, 2 points per homo-polymer longer than 5 bp, 1 point per each base over or below the limit (C or G fraction <15% and >25%, A or T fraction <25% and >35%). To exclude potentially repetitive elements from the design, all probes were compared to the reference genome using BLAST and those returning more than one hit (as defined by *E*-value cutoff <0.01) were discarded from the design. Probes were synthesized on custom 244 k Agilent arrays with randomized positions.

Library preparation

DNA was fragmented for 6 min using a Covaris S2 sonicator (6 × 16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping). After fragmentation, fragments were blunt-ended and phosphorylated at the 5'-end using End-it Kit (Epicentre) according to the manufacturer's instructions, followed by ligation of double-stranded short adapters (adapter 1: pre-annealed duplex of 5'-CTA TGG GCA GTC GGT GAT-3' and 5'-ATC ACC GAC TGC CCA TAG TTT-3' and adapter 2: pre-annealed duplex of 5'-CGC CTT GGC CGT ACA GCA G-3' and 5'-GCT GTA CGG CCA AGG CG-3'; all oligo's were acquired through Integrated DNA Technologies (Coralville, IA, USA) and pre-annealing was done by mixing complementary oligonucleotides at 500 µM concentration and running on thermocycler with the following program: 95°C for 3 min, 80°C for 3 min, 70°C for 3 min, 60°C for 3 min, 50°C for 3 min, 40°C for 3 min and 4°C hold). Ligation was performed using Quick ligation kit (New England Biolabs) with 1 µg of fragmented DNA, 750 nM adaptor 1 and adaptor 2, 150 µl of 2× Quick ligation buffer and 5 µl Quick Ligase in a total volume of 300 µl. Samples were purified on Ampure beads (Agencourt) and run on a native 6% polyacrylamide gel. Fragments ranging from 125 to 150 bp were excised; the piece of gel containing fragments was shredded and dispersed into 400 µl of Platinum PCR Supermix with 750 nM of both amplification PCR primers (provide sequence of amplification primers), 2.5 U of Pfu DNA polymerase (Stratagene) and 5 U Taq DNA polymerase (Bioline). Before ligation-mediated amplification, the PCR sample was incubated

at 72°C for 20 min in PCR mix to let the DNA diffuse from the gel and to perform nick translation on non-ligated 3'-ends. After eight cycles of amplification, the library DNA was purified on Ampure beads and the quality was checked on a gel for the proper size range and the absence of adapter dimers and heterodimers. This library served as a stock for all subsequent hybridization experiments.

Enrichment hybridization and elution

Prior to hybridization, 50 ng of stock library was amplified using 10 cycles in 1000 µl of Platinum PCR Supermix with 750 nM of both amplification PCR primers to produce a sufficient amount of library DNA necessary for enrichment (Table 1). Amplified library DNA was subsequently purified using a MinElute Reaction Cleanup Kit (Qiagen). Amplified DNA was mixed with 5 × weight excess of human Cot-1 DNA (Invitrogen) and concentrated using a speedvac to a final volume of 12.3 µl. DNA was mixed with 31.7 µl Nimblegen aCGH hybridization solution and denatured at 95°C for 5 min. After denaturing, the sample was hybridized for 65 h at 42°C on a 4-bay MAUI hybridization station using an active mixing MAUI AO chamber (MAUI). After hybridization, the array was washed using the Nimblegen Wash Buffer Kit according to the user's guide for aCGH hybridization. The temperature of Wash buffer I for Library 2 was 42°C instead of room temperature. Elution was performed using 800 µl of elution buffer (10 mM Tris pH 8.0) in an Agilent Microarray Hybridization Chamber at 95°C for 30 min. After 30 min, the chamber was quickly disassembled and elution buffer collected into a separate 1.5 ml tube. Microarray slides were dipped into re-distilled water and stored for re-use. Eluted library DNA was concentrated in a speedvac to a final volume of 50 µl and amplified with a limited number of PCR cycles (12–14 cycles) with full-length primers (amp-P1: 5'-CCA CTA CGC CTC CGC TTT CCT CTC TAT GGG CAG TCG GTG AT and amp-P2: 5'-CTG CCC CGG GTT CCT CAT TCT CTN NNN NNN NNN CTG CTG TAC GGC CAA GGC G, where N represent unique barcode sequence for each library) to introduce barcode sequences as well as adapter sequences required for SOLiD sequencing.

SOLiD sequencing

To achieve clonal amplification of library fragments on the surface of sequencing beads, emulsion PCR (emPCR) was performed according to the manufacturer's instructions (Applied Biosystems). A total of 600 pg of double stranded library DNA was added to 5.6 ml of PCR mix containing 1× PCR Gold Buffer (Applied Biosystems), 3000 U AmpliTaq Gold, 20 nM emPCR primer 1, 3 µM of emPCR primer 2, 3.5 mM of each deoxynucleotide, 25 mM MgCl2 and 1.6 billion SOLiD sequencing beads (Applied Biosystems). PCR mix was added to SOLiD emPCR Tube containing 9 ml of oil phase and emulsified using ULTRA-TURRAX Tube Drive (IKA). The PCR emulsion was dispensed into 96-well plate and cycled for 60 cycles. After amplification, the emulsion was broken with butanol, beads were enriched for template-positive beads, 3'-end extended and covalently attached onto sequencing slides. Four physically separated samples were deposited on one sequencing slide and sequenced using SOLiD system version 2 to produce 35-base long reads. Library 1 has been additionally sequenced using the SOLiD version 3 system to produce 50-base long reads in a barcoded experimental setup.

Mapping of sequencing data and SNP calling

Sequencing reads were mapped against the reference genome (hg18 assembly, NCBI build 36) using the Maq package (16), which allows mapping in SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA with following settings: number of maximum mismatches that can always be found –n 3, threshold on the sum of mismatching base qualities –e 150. Raw variant positions were called by the Maq package and filtered using custom scripts (available upon request). For stringent SNP calling, we used the following filtering settings: (i) positions with <20× and >5000× coverage were excluded, (ii) each of non-reference alleles had to be supported by at least three independent reads (as determined by different read start positions) separately on positive and negative strand with quality >10, (iii) the non-reference allele should account for at least 20% of the reads covering the polymorphic position, and (iv) the ratio between + and – strand reads should be between 1/9 and 9 (Table 2). Positions that passed these filtering settings were considered as SNPs. A SNP was qualified as homozygous when the fraction of non-reference alleles was above 95% and heterozygous when the fraction of non-reference alleles was between 20% and 95%.

Calculation of evenness score

A crucial parameter for assessing the effectiveness of any enrichment method is the evenness of coverage (12). Here, we introduce a dedicated parameter to represent the evenness of coverage score, *E*. This score intends to describe the

TABLE 1. Enrichment statistics

	Library1	Library2	Library3	Library1
Length of sequenced tags	35	35	35	50
Microarray slide	new	new	reused	new
Amount of hybridized DNA (µg)	3	3	6.5	3
Washing temperature	RT	42°C	RT	RT
Mappable tags (millions)	6.64	18.88	17.05	14.10
Uniquely mappable tags (millions)	4.97	12.70	13.98	10.78
Mappable sequence on target	56%	40%	61%	53%
Uniquely mappable sequence on target	67%	60%	75%	69%
Bases covered 1x	99.59%	99.38%	99.88%	99.97%
Bases covered 10x	92.49%	93.18%	98.08%	99.37%
Bases covered 20x	83.17%	86.77%	95.46%	98.13%
Bases covered with 10% of average coverage*	95%	90%	95%	98%
Bases covered with 25% of average coverage*	86%	77%	86%	92%
Bases covered with 50% of average coverage*	69%	60%	69%	76%
Evenness score E (%)**	70.2	62.4	70.1	74.8

The last column gives the results of an experiment in which the library resulting from a first enrichment experiment was sequenced using the Solid V3 update, which provides 50 bp read lengths.
*The percentage of bases covered with a given percentage of the average coverage is a better measurement for comparison of coverage evenness than the percentage of bases covered with a certain depth, because it is independent on overall depth of sequencing.
**Evenness score represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. This is a measurement that is relatively independent on sequencing depth.

uniformity of base coverage over targeted regions. Together with the percentage of sequenced bases on target, which determines the enrichment level, this score can be used as an objective measure to compare different enrichment experiments as the parameter *E* is quite insensitive to sequencing depth (Figure 2C). The evenness score, *E*, represents the fraction of whole-sequencing throughput that is correctly distributed. Consequently, 1-E represents the fraction of the (whole) sequencing output that still has to be redistributed from positions with coverage above average to positions with coverage below average (by better enrichment) to get the ideal even coverage over all targeted positions. The more even the coverage, the higher the evenness score: *E* will be 100% for completely uniform coverage of every base in the targeted regions and approaches 0% in case of extreme non-uniform distributions. From Figure 2B, one can appreciate that *E* is equivalent to the area under the curve. Hence, a formula for *E* can be readily arrived at by summing for all percentage positions up to the normalized coverage of 1, as in the ideal case all positions will have a coverage at least equal to the average coverage. Thus the evenness score, *E* is defined as:

$$E = \left\{ \sum_{i=1}^{C_{ave}} \frac{P_i}{C_{ave} \cdot N_{TP}} \right\} \cdot 100\% = \left\{ \frac{1}{C_{ave} \cdot N_{TP}} \cdot \sum_{i=1}^{C_{ave}} P_i \right\} \cdot 100\% \quad (2)$$

Where P_i is defined as number of targeted positions with *at least* coverage C_i, C_{ave} is defined as the average coverage through all targeted positions and N_{TP} is defined as a total number of targeted positions. This formula can be rewritten in a form which is numerically more attractive, as C_{ave} may not be an integer number:

$$E = \int_0^1 F(i) \times di \cdot 100\% \quad (3)$$

Where $F(i)$ is the fraction of positions with normalize coverage of at least $C(i)/C_{ave}$. This fraction equals P_i/N_{TP}, where P_i is percentage of position with a coverage of at least $C(i)$ and N_{TP} is the total number of targeted positions. C_{ave} is the average coverage over all targeted positions. This integral corresponds to the area under the curve

TABLE 2. SNP calling statistics

	Library1	Library2	Library3	Library1
Length of sequenced tags	35	35	35	50
Average coverage	67x	148x	204x	213x
SNP positions validated	384	384	384	384
SNP positions filtered due to low coverage (<20x)	18.8%	25.8%	3.4%	1.8%
SNPs with enough coverage filtered out for other reasons*	26.5%	21.1%	27.9%	6.0%
SNPs identified after filtering (with at least 20x coverage)	54.7%	53.1%	68.7%	92.2%
False negative discovery rate	45.3%	46.9%	31.3%	7.8%

*due to low base quality, or large strand bias

of the graph between a normalized coverage of 0 and 1 (Figure 2B).

The relative independence of the evenness score *E* on sequencing coverage is brought about by the normalization to the average coverage. Hence, *E* solely reflects the quality of targeted genome selection.

Illumina SNP genotyping

The DNA sample that was used in this study was genotyped using an Illumina HumanHap550+ Genotyping BeadChip through 23andMe services (http://www.23andme.com). A total of 384 genotyped SNPs, which are either heterozygous or homozygous non-reference in the sample, are located in the 1.69 Mb region of our interest. These positions were used as a reference set for identifying false negatives in our sequencing dataset.

RESULTS AND DISCUSSION

Specificity of enrichment

Short fragment (85–110 nt) libraries were made using focused acoustic fragmentation (Covaris) and used for enrichment on custom-designed 244 K Agilent microarrays. The specificity of the enrichment was determined by sequencing on an ABI SOLiD sequencer version 2 with a quadrant slide capacity per sample (loaded at various densities). This resulted in 219–676 millions of mappable bases out of which 40–61% mapped directly to the targeted regions (Table 1). Increasing the stringency by raising the washing temperature from room temperature to 42°C (Library 2) did not increase the enrichment efficiency (Table 1). In contrast, the evenness of coverage E did decrease appreciably from 70% to 62 % (Table 1 and Figure 2A), suggesting selective loss of specific target regions. Increasing the amount of DNA for hybridization (Library 3) had no effect at all.

The percentage of on-target bases increases to 60–75% when only taking into account bases from reads that could be placed completely uniquely on the genome. Our results contrast with previously reported results (8), where the use of 100–200 bp fragments resulted in only 29% percent of reads mapping to targeted exons. The observed difference could possibly be explained by differences in probe design strategy and/or array platform (Nimblegen versus Agilent). A different approach for enrichment using 170-nt long biotinylated RNA probes (12), resulted in 42–50% of sequenced bases mapping directly to targeted exons. An alternative solution-based approach used molecular inversion probes (MIPs) (13) for selective capturing of 55 000 human exons and resulted in >99% of all reads mapping to the targeted regions. However, since the first 20 bp of each

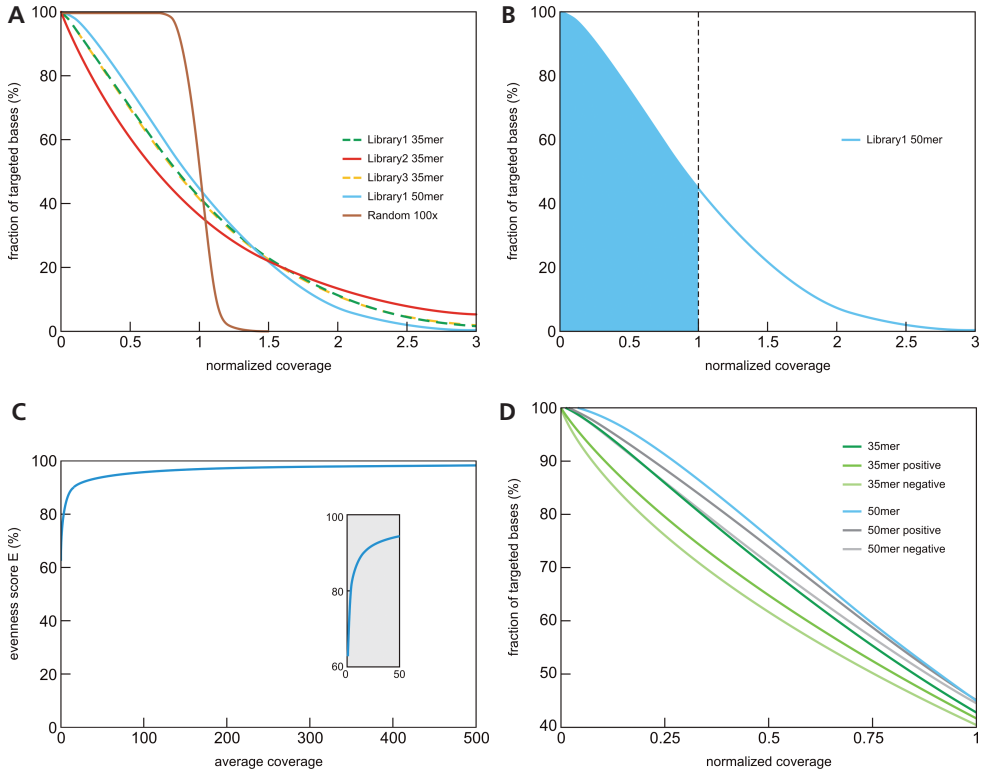


Fig 2: Comparison of sequence coverage evenness after enrichment. The fraction of target positions with at least that coverage was as compared to the average coverage. A) Comparison of various enrichment (washing temperature and input DNA) and sequencing (35 vs 50-mer) conditions. Library 1 sequenced by 50-mer reads results in the most even coverage compared to other libraries. The brown curve depicts the best possible evenness for an ideal evenly enriched sample with 100x average coverage where the unevenness is purely caused by statistical randomness in the coverage assuming a Poisson distribution of the sequencing reads. B) The evenness score, *E*, represents the fraction of whole sequencing throughput that is correctly distributed (marked area below the curve). Consequently, 1-E represents the fraction of the (whole) sequencing output which has to be redistributed from positions with coverage above average to positions with coverage below average (by better enrichment) to get the ideal even coverage over all targeted positions. The more even the coverage, the higher the evenness score. C) Correlation of evenness score E for randomized sets to the sequencing depth. In this simulation the unevenness of these datasets is purely caused by the random distribution of reads and fits a Poisson distribution of sequence coverage. When the discrete character of the data is reduced by sufficient depth of coverage, E changes only slightly with increasing average coverage and thus can be characterized as relatively independent of sequencing depth. D) Comparison between + and – strand coverage for 35 and 50-mer reads. In the case of 50-mer reads the coverage is more even with fewer positions covered by extremely low (or high) numbers of sequencing tags. This difference is more prominent when the coverage is determined separately for the positive or negative strand. Independent strand coverage is better for 50-mer than for 35-mer sequencing.

sequenced read are always coming from the MIPs only the remainder of the sequence is informative for genotyping and a substantial proportion of the overall sequencing output thus has to be discarded as non-informative. Moreover, exons longer than the maximum read length of the sequencing platform will need multiple probe designs for capturing and sequencing. As no commercial solutions are available for cost-effective synthesis of high-quality, long oligonucleotides (>100 nt), which are required for this approach, widespread implementation of this method is questionable.

Evenness of coverage

During enrichment not all DNA hybridizes to capture probes with the same efficiency. As a result, targeted sequences are covered unevenly with sequencing reads. For a limited number of regions there even is no coverage at all. Generally, the more even the coverage distribution is, the less overall sequencing depth is required for variant detection. In our experiments 60–76% of targeted regions are covered with >50% of average coverage, 77–92% of targeted regions are covered with >25% of average coverage and 90–98% of reads have >10% of average coverage (Table 1 and Figure 2A). In addition, we covered 99.38–99.97% of targeted positions with at least one read suggesting significantly better evenness of coverage compared to other studies, where 82% (10); 95% (12) or 99.21% (17) of targeted positions were covered with at least one read.

For better comparison of coverage evenness, we introduce a new parameter, the evenness score *E* (see section 'Materials and Methods' section and Figure 2B). *E* is relatively independent of sequence coverage and thus enables the comparison of different libraries with varying sequencing depth. The relative independence of the evenness score *E* to sequencing depth is shown in Figure 2C. In this figure, the correlation of *E* for a randomized read sets assuming a Poisson read distribution, is calculated as a function of the sequencing depth. Figure 2C shows that once the sequencing coverage is sufficiently high (>50×), the evenness score *E* changes only slightly with increasing average coverage and thus can be characterized as relatively independent of the sequencing depth. The evenness score represents the area under the curve in a fraction of the positions with at least that coverage vs normalized coverage at $X = 1$ (see in Figure 2B). In other words, *E* represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. The evenness score, $E = 100\%$ for perfectly uniform coverage and approaches 0% in cases of extreme non-uniform distributions. By using short-fragment libraries, we achieved even distributions with evenness scores (*E*) ranging from 62.4% (Library 2) to 74.8% (Library 1 sequenced with 50-nt read length). Sequencing of the same enriched library on SOLiD V3 with 50-nt long reads instead of 35-nt long reads with SOLiD V2 resulted in a better evenness score with *E*, rising from 70.2% to 74.8% (Table 1 and Figure 2A). The most obvious explanation for this observation would be that the longer reads resulted in improved bridging of regions with lower capture efficiency due to fragments captured by well-performing flanking probes. In addition, due to low complexity of genomic regions, 35-mer tags cannot be mapped uniquely to a substantial part of the genome and this fraction is reduced with 50-mers.

To illustrate that the described approach universally results in high evenness, we analyzed additional experiments that were performed with the same experimental protocol, but with different array designs and/or species (Supplementary Table S1 and Supplementary Figure S1). We do find a consistent high evenness score, even

between organisms (human, rat and *Arabidopsis*). Moreover, we reanalyzed publically available datasets from recently published genomic enrichment experiments (6,11,18) showing that our experiments result consistently in more even coverage, especially when considering strand-specific coverage (Supplementary Table S1 and Supplementary Figure S1). The latter is especially important, as we observed in our dataset that proper coverage on both strands is instrumental for reducing false positive heterozygote SNP calling. This is most likely due to systematic errors that are introduced by the sequencing process due to platform-specific biases in sequencing chemistry, which is in all cases context dependent and thus different for the + and – strand.

Sequencing of positive and negative DNA strand

We found that having sequencing data mapping to both positive and negative strand is an important factor in reducing false positive variant calls. Therefore, we evaluated the evenness of coverage with respect to DNA strand. A substantial part of the targeted sequence was covered by sequencing tags coming from only one strand in case of libraries sequenced as 35-mers with SOLiD V2 chemistry (Figures 2D and 3). Increasing the read length to 50-mers improved double-stranded coverage markedly, in line with the observed overall base coverage (Figures 2D and 3).

Improvements of sequencing coverage

Since the contribution of the random character of sequencing to unevenness of coverage was minor (Figure 2C), further improvements in the evenness of the sequencing coverage could be obtained by improvement of the enrichment procedure. The strategy used in our array design resulted in overlapping probes and did not provide much opportunity for redesigning probes for poorly enriched regions. Therefore, we included various probes at variable quantities in our test design. We found a very strong correlation between the number of probes and eventual base coverage (Figure 4). Spotting more copies of the same probe for underperforming regions therefore seems an effective strategy for improving *E*. Furthermore, these results also indicate that a limiting factor for DNA yield after elution could be the number of probe molecules and their saturation after 65 h of hybridization and not the depletion of targeted library molecules. This is supported by the observation that increasing the amount of library DNA during hybridization had no effect on enrichment efficiency and coverage. However, we cannot exclude that the efficiency of mixing during hybridization is limited and that local depletion of target sequences occurred, without saturating the capturing probes. Presence of capture probes at physically separated locations (which is the case in the random design used in these experiments) would in such case also improve capturing efficiency.

Yet another possibility for improvement of sequence coverage can be expected from further increasing sequencing read length, in line with our results obtained for 35- versus 50-mer reads.

Identification of polymorphic positions

To determine the accuracy of SNP detection, we compared SNPs called from sequencing data obtained from enriched short-fragment libraries and those genotyped by an Illumina HumanHap550+ Genotyping BeadChip. A total of 384 SNP genotyped positions were different from the reference allele (either heterozygous or homozygous) and were located in the targeted regions. From those 384 SNP positions, 53.1– 92.2% passed the stringent criteria of our SNP filtering pipeline, which required sufficient high-quality base coverage on both DNA strands. In concordance with evenness of coverage, Library 1 sequenced on SOLiD V3 with 50-nt

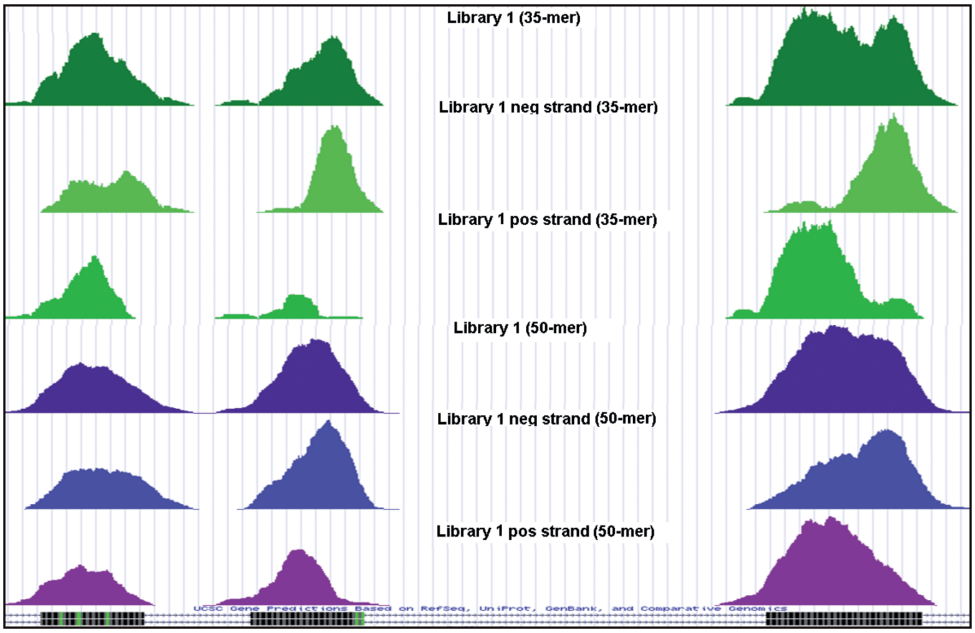


Fig 3: Exemplary representation of target coverage after enrichment. Sequencing results of library 1 are shown for 35-mer (green) and 50-mer (purple) sequencing. Total, positive and negative strand reads are shown independently. Coverage is more equal and better represented by both strands for the longer sequencing reads.

long reads gave the best results with only a 7.8% false negative discovery rate over the complete targeted region. From this total of 7.8% false-negative SNP positions, 3.6% had not been covered with (i) at least 20 reads, and (ii) at least 3 reads from each of the strands or (iii) did not have a coverage ratio from both strands within the limits set at 1/9 and 9. Consequently, this part of the targeted regions could already be marked as not surveyed, even prior to SNP calling, since this part would not pass our minimal requirements of SNP calling. Such a prediction will be important for clinical diagnostic purposes because this enables one to predict which regions have not been sequenced sufficiently deep for reliable SNP calling. Taken together, at ~200× average sequence coverage, we were able to survey 96.4% of 1.69 Mb of genomic sequence with only 4.2% of false negative calls, while 3.6% of targeted regions had to be marked as unsurveyed.

The better performance of Library 1 could be explained by better evenness of coverage (more positions have sufficient base coverage for reliable SNP calling) as well as by better strand balance where more positions have good coverage coming from tags mapping to both negative and positive strands (Figures 2D and 3). Another explanation for the better performance of longer reads could be a reduction of mappability bias. While mapping sequencing reads, non-reference alleles are more likely to be discarded due to low mapping quality, since they already have one mismatch (two mismatches in SOLiD color space) compared to reads coming from reference alleles. Additionally, capture of non-reference DNA molecules to reference capture probes may be slightly less effective. Indeed, the overall frequency of non-reference allele

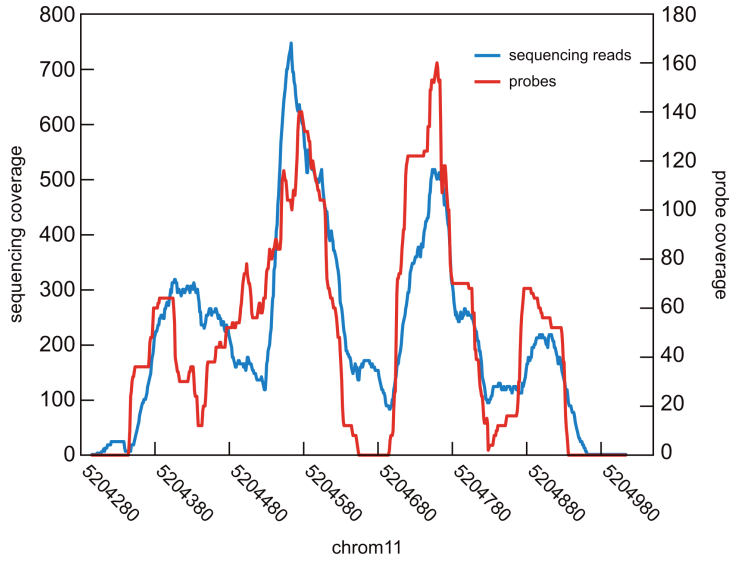


Fig 4: Correlation of probe density and sequencing coverage. Each genomic region was represented on the array with a variable number of capture probes throughout the region. The sequencing coverage per base (blue line) linearly correlates with probe density (red line).

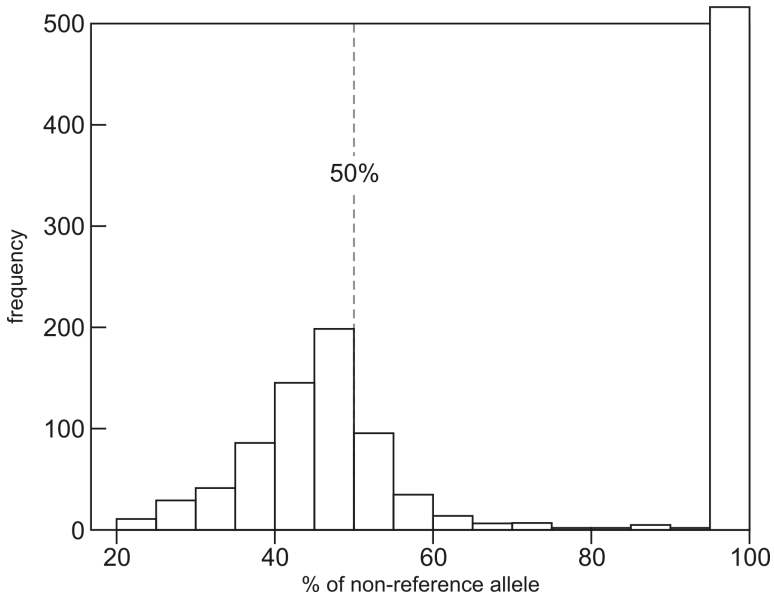


Fig 5: Distribution of non-reference allele reads. The percentage of non-reference allele reads was calculated for every heterozygous and homozygous non-reference allele position in the targeted region (n = 1,197) and is represented in bins of 5%. For heterozygous calls the distribution is skewed towards reference allele reads.

reads for heterozygous positions was shifted downward form the 50% position, although not dramatically (Figure 5). The mappability issue is even more serious for SNPs that have additional linked SNPs in close vicinity. This bias is less prone for longer reads since one mismatch contributes less to overall mapping accuracy in 50-mers compared to 35-mers.

Detection of novel variants

We analyzed the results of Library 1 (50-mer reads) for the presence of polymorphisms. A total of 1197 SNPs were identified within the targeted regions plus 30-bp flanking intronic regions using our SNP detection pipeline. This set included the 384 previously genotyped SNPs as well as 759 other polymorphisms that were already present in dbSNP129 or the Ensembl database. We considered these 1143 (95.5%) SNPs as validated and set out to validate the remaining 54 SNPs by PCR-based dideoxy resequencing. We failed to develop working assays for 10 candidate SNPs, most likely due to the repetitive nature of the genomic environment. We found that only eight of the remaining candidates, all heterozygote scores, were identified as false positives. Altogether, our results indicate a false positive rate of less than one per ~200 000 bp (0.0005%). All eight false positive SNPs tended to have a lower than average percentage of non-reference allele reads and/or low overall base coverage compared to true positives (Figure 6). Although we only used planar microarrays in our experiments, we believe that the characteristics of short-fragment libraries described here will be equally applicable to any hybridization-based approach, including in-solution methods.

Input DNA requirements

Relatively high amounts of DNA are normally used for enrichment procedures, which can

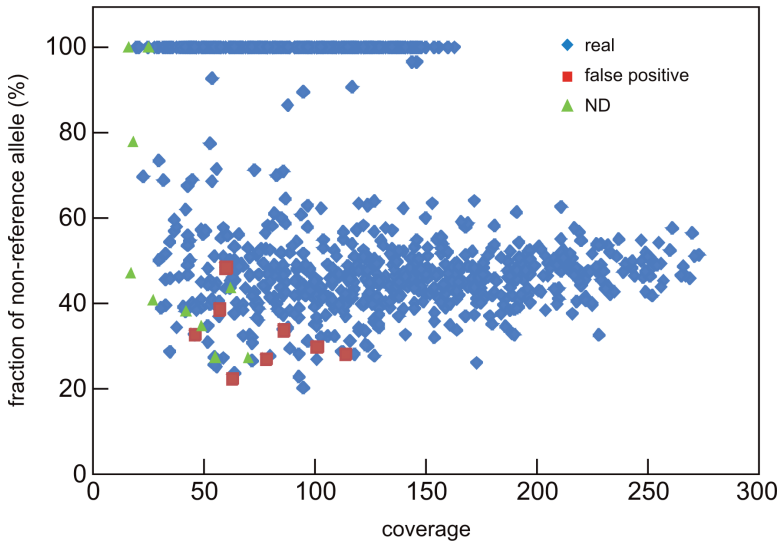


Fig 6: Sequencing coverage and percentage of non-reference allele distribution for validated and non-validated SNPs. All polymorphic and non-reference positions that were identified by the SNP detection pipeline are plotted as a function of total base coverage versus non-reference read frequency. Validated SNPs (either by their presence in dbSNP or by resequencing) are indicated in blue, non-validated SNPs are shown in red and positions for which no working validation assay could be designed in green. False-positive SNPs tend to have a lower percentage of non-reference allele reads and/or low overall coverage.

be a limiting factor for many clinical applications. In our experiments, we used only 1 µg of genomic DNA for all experiments shown. The fragment library was made and amplified with only eight PCR cycles to produce a stock library, which was sufficient for at least 20 independent enrichment procedures as described here. Before enrichment a small proportion (50 ng) of the initial library was amplified with 10 PCR cycles to produce sufficient material for hybridization. After enrichment the eluted library was amplified by an additional 13 cycles to produce amounts sufficient for accurate quantification and sequencing. By using these logistics, the stock library could be used multiple times for different enrichment experiments, taking away the need for re-isolating DNA or re-preparing sequencing libraries. In addition, most of the amplification cycles (18 out of 31 in total) were done before hybridization. Potential biases in coverage caused by PCR could, in theory, be normalized again during the hybridization step. However, this requires that capturing probes, rather than target molecules, are the limiting factor in this step.

Detailed analysis of our deep-sequencing results revealed no unexpected clonality bias due to the amplification steps.

CONCLUSIONS

Short-fragment libraries provide highly efficient enrichment characteristics for exonic targets. The relative even base and strand coverage facilitates accurate SNP and mutation retrieval and discovery. To measure the evenness of coverage, which is relatively independent of sequencing depth, we have introduced the parameter *E* [see Equations (2 and 3) and Figure 2]. This score can be applied to any genomic enrichment experiment and in combination with the percentage of reads on target it provides the possibility to compare the efficiency of different approaches.

ACKNOWLEDGEMENTS

We would like to thank Wim Verhaegh, Anja van de Stolpe and Dennis Merkle for their useful comments and help with manuscript preparation.

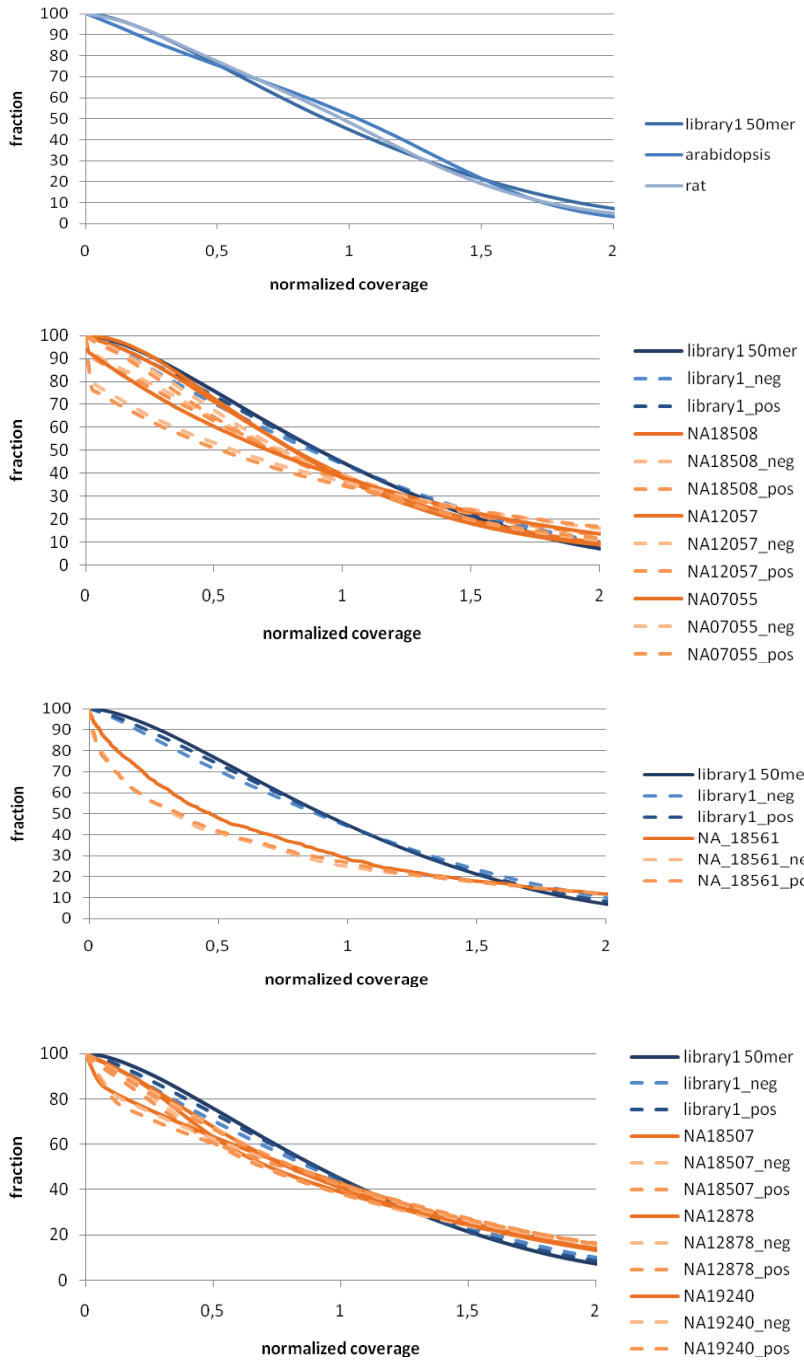
REFERENCES

1. Zheng, W., Long, J., Gao, Y.T., Li, C., Zheng, Y., Xiang, Y.B., Wen, W., Levy, S., Deming, S.L., Haines, J.L. *et al.* (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet*, **41**, 324-328.
2. Song, H., Ramus, S.J., Tyrer, J., Bolton, K.L., Gentry-Maharaj, A., Wozniak, E., Anton-Culver, H., Chang-Claude, J., Cramer, D.W., DiCioccio, R. *et al.* (2009) A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nat Genet*, **41**, 996-1000.
3. Papaemmanuil, E., Hosking, F.J., Vijayakrishnan, J., Price, A., Olver, B., Sheridan, E., Kinsey, S.E., Lightfoot, T., Roman, E., Irving, J.A. *et al.* (2009) Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nat Genet*, **41**, 1006-1010.
4. Kathiresan, S., Voight, B.F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P.M., Anand, S., Engert, J.C., Samani, N.J., Schunkert, H. *et al.* (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, **41**, 334-341.
5. Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C.S., Humphreys, M.K., Platte, R., Morrison, J., Maranian, M., Pooley, K.A., Luben, R. *et al.* (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet*, **41**, 585-590.
6. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272-276.
7. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. *et al.* (2007) Direct selection of human genomic loci by microarray hybridization. *Nat Methods*, **4**, 903-905.
8. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. *et al.* (2007) Genome-wide in situ exon capture for selective resequencing. *Nat Genet*, **39**, 1522-1527.
9. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat Methods*, **4**, 907-909.
10. Bau, S., Schracke, N., Kranzle, M., Wu, H., Stahler, P.F., Hoheisel, J.D., Beier, M. and Summerer, D. (2009) Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. *Anal Bioanal Chem*, **393**, 171-175.
11. Summerer, D., Wu, H., Haase, B., Cheng, Y., Schracke, N., Stahler, C.F., Chee, M.S., Stahler, P.F. and Beier, M. (2009) Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. *Genome Res*, **19**, 1616-1621.
12. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, **27**, 182-189.
13. Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A. and Shendure, J. (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. *Nat Methods*, **6**, 315-316.
14. Herman, D.S., Hovingh, G.K., Iartchouk, O., Rehm, H.L., Kucherlapati, R., Seidman, J.G. and Seidman, C.E. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods*, **6**, 507-510.
15. Carriaso, M. and Lennon, G. (2009).
16. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**, 1851-1858.
17. Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W. and Hannon, G.J. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat Protoc*, **4**, 960-974.
18. Okou, D.T., Locke, A.E., Steinberg, K.M., Hagen, K., Athri, P., Shetty, A.C., Patel, V. and Zwick, M.E. (2009) Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. *Ann Hum Genet*, **73**, 502-513.

Supplementary Table

		Evennes score*		
Dataset	Library	Both strands	Positive strand	Negative strand
Human	Library1_50	75.7%	74.2%	74.2%
Arabodipsis	193	74.8%	73.1%	71.1%
Rat	vip	76.6%	74.0%	74.4%
Summerer et al.	NA18561	49.8%	41.8%	41.7%
Okou et al.	NA18508	70.8%	63.3%	64.8%
	NA12057	71.3%	65.5%	66.9%
	NA17005	62.0%	52.4%	56.6%
Ng et al.	NA18507	65.8%	62.7%	63.9%
	NA12878	63.0%	59.2%	61.1%
	NA19240	65.8%	61.8%	63.0%

Performance of our enrichment method with different probe designs and different species with comparison to other publically available datasets.
* Evenness score represents the fraction of sequenced bases that do not have to be redistributed from above-average coverage to below-average coverage positions to obtain completely even coverage for all targeted positions. This is a measurement that is relatively independent on sequencing depth.



Supplementary Figure: Performance of our enrichment method with different probe designs and different species with comparison to examples of other publicly available datasets. Evenness of coverage is reproducible among different capture designs and is better compared to other available datasets NA18508(1), NA12057(1), NA07055(1), NA18561(2), NA18507(3), NA12878(3), NA19240(3).

3

MUTATION DISCOVERY BY TARGETED GENOMIC ENRICHMENT OF MULTIPLEXED BARCODED SAMPLES

*Nijman IJ, *Mokry M, *van Boxtel R, Toonen P, de Bruijn E, Cuppen E: **Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples**. *Nat Methods* 2010, **7**(11):913-915.
*equal contribution

ABSTRACT

Targeted genomic enrichment followed by next-generation DNA sequencing dramatically increased the efficiency of mutation discovery efforts. We describe a protocol for genomic enrichment of pooled barcoded samples in a single assay that substantially increases experimental flexibility and efficiency. We screened a 770 genes (1.4 Mb) in 30 rat individuals and identified known variants at >96 % sensitivity, as well as novel mutations at a false positive rate of <1 in 8 Mb.

While throughput of next-generation sequencing equipment continues to increase one may argue whether genomic enrichment strategies are sufficiently cost- and labour-efficient compared to whole genome sequencing. However, for many applications, including most diagnostic assays, genetic variation in only part of the genome (e.g. coding sequences or specific sets of genes) is of interest with the rest of the genome being irrelevant or non-interpretable.

Targeted genomic enrichment for high-throughput sequencing was first demonstrated by using planar microarrays with tiling paths of probes complementary to the target regions of interest, while more recently in solution approaches using DNA or RNA probes (in review (1)) have become increasingly popular because of experimental versatility. However, in all cases, only a single sample can be enriched per assay, limiting throughput and making the approach laborious and costly. As a result, efficient protocols for large-scale mutation discovery efforts that include many samples as well or involve many genes are still lacking. We developed a universal and efficient protocol for large-scale mutation discovery and demonstrated a proof-of-principle in a gene-driven target-selected mutagenesis approach (also known as TILLING, Targeting Induced Local Lesions In Genomes, for examples (2,3)) for the generation of genetic knockout models in the rat (4). This reverse genetics approach is commonly used in a wide-range of other model animals and

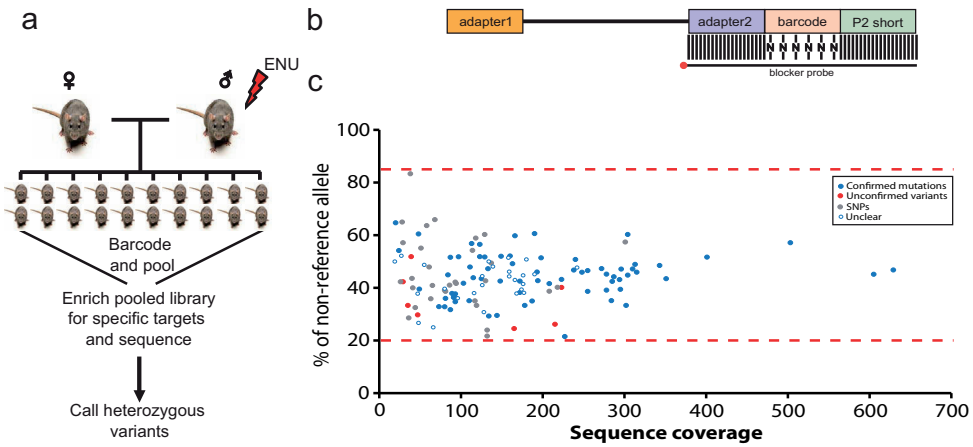


Figure 1. (**a**): Schematic representation of the next-generation reverse genetics procedure. This procedure can be divided into 4 different experimental steps (for details of each step see Supplementary Figure 1). First, mutagenized founder individuals are outcrossed to generate an F1 resource from which DNA is isolated. Second, the DNA is used to generate sequencing libraries. SOLiD barcodes are introduced for every individual sample before pooling and multiplexed enrichment for genes of interest. Next, the enriched multiplexed libraries are sequenced, followed by barcode decoding and bioinformatic data analysis for the presence of heterozygous mutations. All identified variants were verified by an independent assay using capillary sequencing. (**b**): Specific blocking oligonucleotide to increase hybridization efficiency in multiplexed genomic enrichment experiments. The oligonucleotide sequence is complementary to the adapter 2 and the P2-short adapter, which are required for SOLiD sequencing. Between these sequences degenerate bases are present that are complementary to the different barcodes. Furthermore, a dideoxy group is present at the 3′ end to prevent contribution to the post-enrichment PCR. (**c**): Frequency and coverage plot of identified variants in the mutation discovery screen. The allele frequencies of the identified variants are plotted against sequence coverage in the SOLiD data. The red dashed lines indicate the upper and lower allele frequency boundaries between which variants are considered to be heterozygous. False positive variants (red dots) tend to have a low percentage of non-reference alleles and a low coverage.

plants (2,5) and is based on the random introduction of (point-)mutations in the germline of the organism of interest by chemical mutagenesis. Next, progeny is screened for induced mutations in target genes of interest, with the objective to identify alleles that affect protein function, e.g. by introducing a premature stopcodon or by changing functionally important residues. To this end, various screening methods have been applied, including nuclease- (6) or Mu transposase-based mismatch detection (7), CSCE (8) (Conformation Sensitive Capillary Electrophoresis), HRM (8) (High Resolution Melting Analysis) or PCR-based dideoxy resequencing (9,10). However, these assays are typically limited to single genes or exons, making them unsuited for the systematic generation of genome-wide collections of knockout mutants employing the large archives that have been or are currently being build for various species (7,9).

To be able to screen a much larger number of genes in parallel, we multiplexed individually barcoded samples before targeted genomic enrichment and SOLiD-based next-generation sequencing (Fig. 1a). To be able to determine the false-positive rates, we made use of a previously generated repository of mutant F1 rats (10). A selected set of genes was previously screened in these animals by traditional dideoxy resequencing of PCR-amplified genomic segments (exons) (4) , showing a heterozygous ENU-induced point mutation frequency of 1 in 720 kb. Here, custom capture microarrays were designed to enrich for genomic regions of interest (1), including all conserved non-odorant G protein-coupled receptors plus an additional ~500 kb of sequence encoding MIM morbid genes. In total, the complete coding sequence of 770 genes was covered, adding up to 1.4 Mb of target sequence (Supplementary Table 1).

In a first experiment, we pooled equimolar amounts of individually barcoded sequencing libraries from 10 animals and enriched them for the genomic regions of interest in a single enrichment experiment. Sequencing using a single AB/SOLiD slide yielded 16.6 ± 3.8 million reads for every animal (Table 1 and Supplementary Table 2). In contrast to standard non-multiplexed enrichment, where we routinely obtain between 60 and 90 % of reads on target (11), we only acquired about 35 % of the reads on target across all samples using multiplexed enrichment. Nevertheless, an average coverage of about 200 × for the targeted regions was obtained for all samples, which is sufficient for reliable variant discovery (Table 1 and Supplementary Table 2).

We hypothesized that the relatively low enrichment specificity might be due to the long barcode-containing 3' adaptors (52 nt), which could mediate hybridization between unrelated library molecules and thereby allow for carry-over of non-targeted DNA molecules (12). Therefore, we designed specific degenerate primers to block the barcode sequences before and during hybridization-based enrichment (Fig. 1b). The amount of degeneracy of the region blocking the actual barcode can differ based on the barcode kit used. In this study the 20 barcode kit was used so degeneracy is moderate (Supplementary Table 3 and Supplementary Fig. 2a). When using the 96 barcode kit, full degeneracy is required (Supplementary Fig. 2b). New libraries were prepared for a second set of 20 animals and enrichment was performed in the presence of these blocking oligonucleotides. Now the enrichment efficiency was increased to an average of 60 % of the reads on target, resulting in 188 × average target coverage per animal from a single slide AB/SOLiD run (Table 1 and Supplementary Table 4).

To identify heterozygous mutations in the enrichment sequencing data, we developed a bioinformatic pipeline that takes into account allele frequency, independent read starts, and allele coverage on the Watson and Crick DNA strands. Using standard

parameters all 31 known mutations (from previous capillary sequencing) were picked up except for one. Manual inspection of the mutation that was missed revealed that the coverage of the mutant allele was below the threshold level, but also that this mutation was flanked by a linked SNP two basepairs downstream of the mutation. The presence of two mismatches between the probe and the mutant allele may result in decreased capturing efficiency compared to DNA fragments originating from the wild-type allele. Furthermore, the presence of two additional polymorphisms may also affect mapping efficiency of the short read (50-mer) sequencing data, which would further increase the reference bias (ie the bias where a perfect reference sequence maps better that a less matching sequence due to polymorphisms).

Besides the 30 known mutations, an additional 133 novel heterozygous variations were identified (Supplementary Tables 5 and 6). PCR amplicons were designed to confirm these mutations by capillary sequencing. For 21 variants, we failed to design a working amplicon or capillary sequencing failed due to low sequence complexity (e.g. simple repeats). The fact that these loci could be screened by AB/SOLiD sequencing can be considered a specific advantage of the current approach compared to traditional PCR-based dideoxy sequencing. A total of 105 variants were readily confirmed, while 7 mutations could not be confirmed by capillary sequencing. However, inspection of the allele frequencies in combination with sequence coverage in the SOLiD data (Fig. 1c) suggests that most of these mutations are real mutations that may be false negatives in traditional PCR-based dideoxy resequencing. In support of this, we did identify 5 novel mutations in genes that were previously screened by capillary sequencing (~230 kb in total) resulting in a false negative rate of 1 in 46 kb (Supplementary Table 6).

Nevertheless, with only 7 potential false positives (7 % of all candidate mutations), the overall false positive rate in our experiments (~58 Mb of screened target sequence) is less than 1 in 8 million base pairs surveyed. Although this rate may go up when lower coverage sequencing is used, the obtained rates are much better than routinely obtained by high-throughput capillary dideoxy resequencing. Furthermore, our results are similar as compared to results obtained in previous non-multiplexed studies. For example, a study screened 304 kb of sequence in 10 individuals with 99.5 % and 97 % accuracy for homozygous and heterozygous sites (13), respectively. In another study 97-99 % SNP calling accuracy was shown (14). It should be mentioned, however, that in these studies the average coverage depth was lower, but also that the false-negative rates were not explored.

Of the 105 confirmed variants identified here, 30 were observed in more than one F1 animal and are therefore believed to be common single nucleotide polymorphisms (SNPs) in the outbred rat strain that was used (Supplementary Table 5). The 75 remaining variants are likely mostly novel chemically induced mutations, resulting in an overall ENU mutation frequency of 1 in 775 kb, which is similar to results obtained in previous screens in this genetic background (4,10) . The identified mutations (Supplemental Tables 5 and 6) include 2 nonsense, 1 translational start site, 53 missense, 14 silent, and 5 non-coding mutations, resulting in novel candidate knockout models for the G protein-coupled receptor *GRM2*, the peroxisomal membrane protein *PXMP3* and the prolactin hormone receptor *PRLHR*.

Taken together, we have described a universal method that allows for highly efficient multiplexed mutation discovery with very low false positive and false negative rates. While we demonstrate utility of the approach with medium target size in combination with medium sample number employing planar capture microarrays, we

have successfully applied the same approach to large target sizes (SureSelect exome, 36 and 50 Mbp) in combination with small numbers of samples using commercially available in solution capture assays (E.C. unpublished results). With continuously increasing capacity of next-generation sequencing equipment, the need for versatile multiplexing targeted experiments will continue to grow and the described approach can easily be scaled further without notably increasing the lab effort or decreasing accuracy, making it amendable for a wide range of applications.

SEQUENCE AND ENRICHMENT STATISTICS

Table 1: Sequencing statistics for enrichment of multiplexed pools of barcoded rat DNA samples

		Coverage regions of interest[a]		Fold coverage[b]	
Pool size	On target (%)	> 1x (%)	> 20x (%)	Mean	Median
10 animals	35	99.4	95.1	204	183
20 animals	59	99.4	96.4	196	191

[a] target region includes coding sequences of 770 genes encompassing ~1.4 Mb
[b] including non-targeted regions (e.g. exon-flanking intronic sequences)

ACCESSION CODES

Sequencing data is available from NCBI GEO accession number (ID GSE22024).

ACKNOWLEDGMENTS

This work was supported by funds from the Cancer Genomics Centre and the award "Exploiting natural and induced genetic variation in the laboratory rat" to E.C. from the European Heads of Research Councils and European Science Foundation EURYI (European Young Investigator) Award scheme.

ONLINE METHODS

Preparation of sequencing libraries

One microgram of each individual genomic DNA sample was fragmented to ~ 100 nt double-stranded DNA fragments by sonication in 100 µl of 10 mM Tris buffer pH 8 using Covaris S2 sonicator (6× 16 mm AFA fiber Tube, duty cycle: 20 %, intensity: 5, cycles/burst: 200, frequency sweeping, 6 minutes). After fragmentation, fragments were blunt-ended and phosophorylated at 5'-prime end using End-it Kit (Epicentre) according to the manufacturer's instructions. Double stranded adapters compatible with SOLiD sequencing (Supplementary Table 3: adapter 1: pre-annealed duplex of Adapter 1_A and Adapter 1_B and adapter 2: pre-annealed duplex of Adapter 2_A and Adapter 2_B; all oligo's were acquired through Integrated DNA Technologies and pre-annealing was done by mixing complementary oligonucleotides at 500 µM concentration and running on thermocycler with the following program: 95 ˚C for 3 min, 80 ˚C for 3 min, 70 ˚C for 3 min, 60 ˚C for 3 min, 50 ˚C for 3 min, 40 ˚C for 3 min and 4 ˚C hold), were ligated to the fragmented DNA (~ 1 µg) using Quick ligation kit (New England Biolabs) with 750 mM adaptor 1 and adaptor 2, 150 µl of 2x Quick ligation buffer, 5 µl

Quick Ligase in a total volume of 300 µl. Samples were purified using Ampure beads (Agencourt) and amplified in 400 µl of Platinum PCR Supermix with 750 mM of both amplification PCR primers (P1_short and BC-primer 1-20: 5'-CTGCCCCG-GGTTCCTCATTCTCTNNNNNNNNNNCTGCT-GTACGGCCAAGGCG-3' where N represents unique barcode sequence for each library, Supplementary Table 3 and Fig. 2a), 2.5 U of Pfu DNA polymerase (Stratagene) and 5 U Taq DNA polymerase (Bioline). Before ligation-mediated amplification, the PCR sample was incubated at 72 ˚C for 5 minutes in PCR mix to perform nick translation on non-ligated 3'-ends. After 6 cycles of amplification, the library DNA was purified on Ampure beads and the quality was checked on a gel for the proper size range and the absence of adapter dimers and heterodimers. Finally, libraries were equimolarly pooled and size selected on 4 % agarose gel for the 125-175bp fraction.

Array hybridization and elution

Here, we have chosen to use custom-designed microarrays for targeted genomic enrichment, because of their high flexibility in terms of design, their robust performance in heterozygous mutation discovery experiments (11), and their cost-effectiveness. The array costs (about 500 euro/dollar) currently contribute only about 5-10 % of the costs of a multiplexed mutation discovery experiment. Because at least 20 samples can be multiplexed and arrays can be re-used after extensive stripping (preferably using different sets of barcodes to fully exclude potential carry-over) up to 100 samples can be enriched using a single array, bringing enrichments costs below 10 euro/dollar per sample. However, the described approach using microarrays for enrichment is not very amendable to parallelization and automation, making in solution approaches more attractive alternatives. Although not described here, we explored the possibility to use the same approach as described here in combination with standard in solution enrichment (Agilent SureSelect) and found that enrichment of 20 pre-barcoded samples resulted in similar enrichment efficiencies and mutation retrieval rates.

Prior to hybridization size-selected library was amplified with 10 PCR cycles (95˚C for 5 min, 10 cycles: 95 ˚C for 15 s, 54 ˚C for 15 s and 70 ˚C for 45 s) in 1000 µl of Platinum PCR Supermix with 750 mM of both amplification PCR primers (Supplementary Table 5: P1_short P2_short) and 6.25 U of Pfu DNA polymerase (Stratagene) to produce amounts of DNA necessary for enrichment (2-7 µg) and purified using MinElute Reaction Cleanup Kit (Qiagen). Amplified DNA was concentrated to 12,3 µl of volume by speedvac together with non-specific, repeat enriched DNA (10 × weight excess of Hybloc rat DNA (Applied Genetics Laboratories, Melbourne, FL)). In addition for rat pool2 10 × weight excess of both barcode blocking oligos (Supplementary Table 3: block1 and block2) was used. DNA was mixed with 31.7 µl of Nimblegen aCGH hybridization solution and denatured at 95˚**C** for 5 minutes. After denaturing the sample was hybridized to a custom Agilent microarray for 72 hours at 42˚**C** on a MAUI hybridization station. After hybridization, the array was washed using Nimblegen Wash Buffer Kit according the user's guide for aCGH hybridization. Elution was performed using 800 µl of elution buffer (10 mM Tris pH 8.0) in an Agilent Microarray Hybridization Chamber at 95˚**C** for 30 minutes. After 30 minutes the chamber was quickly dissembled and eluted sample was collected. Eluted library DNA was concentrated by speedvac to a volume of 50 µl, mixed with 400 µl of Platinum PCR Supermix with 750 mM of both full length amplification PCR primers (Supplementary Table 3: amp-P1 and amp-P2) and 2.5 U of Pfu DNA polymerase (Stratagene), and amplified with 13 PCR cycles (activation 95˚C for 5 min, 13 cycles: 95˚C for 15 s, 54 ˚C for 15 s and 70˚C for 45 s).

SOLiD Sequencing

To achieve clonal amplification of library fragments on the surface of sequencing beads, emulsion PCR (ePCR) was performed according to the manufacturer's instructions (Applied Biosystems). 1500 pg of double stranded library DNA was added to 5.6 ml of PCR mix containing 1 × PCR Gold Buffer (Applied Biosystems), 3000 U AmpliTaq Gold, 20 nM ePCR primer 1, 3 µM of ePCR primer 2, 3.5 mM of each deoxynucleotide, 25 mM $MgCl_2$ and 1.6 billion SOLiD sequencing beads (Applied Biosystems). PCR mix was added to SOLiD ePCR Tube containing 9 ml of oil phase and emulsified using ULTRA-TURRAX Tube Drive (IKA). The PCR emulsion was dispensed into 96-well plate and cycled for 60 cycles. After amplification emulsion was broken with butanol, beads were enriched for template positive beads, 3'-end extended and covalently attached onto sequencing slides. Sequencing was performed on an AB/SOLiD sequencer with V3 chemistry (50 bp fragments).

Rat mutagenesis

All animal experiments were approved by the Animal Care Committee of the Royal Dutch Academy of Sciences according to the Dutch legal

ethical guidelines. Experiments were designed to minimize the number of required animals and their suffering. ENU treatment of male *MSH6*$^{-/-}$ rats (Msh6^{1Hubr}) was done as described (10). Animals were housed under standard conditions in groups of two to three per cage per gender under controlled experimental conditions (12-h light/dark cycle, 21 ± 1°C, 60 % relative humidity, food and water *ad libitum*).

Gene-centric rat exon enrichment array design
We extracted the exonic sequence from 770 genes from Ensembl (build53; *see* Supplementary Table 1) with a footprint of 1,392,385 bp. A custom PERL script was used to design 60 bp oligos within a sliding window of 10 bp. Within each window the most optimal probe is selected based on melting temperature and absence of homopolymer stretches. To exclude potentially repetitive elements from the design, all probes were compared to the reference genome using BLAST and those returning more than one hit (as defined by a 60% match of probe sequence) were discarded from the design. 97.5% (1,357,860 bp) of the requested regions was covered by 239,110 probes matching these criteria. Probes were synthesized on custom 244k Agilent arrays with randomized positions.

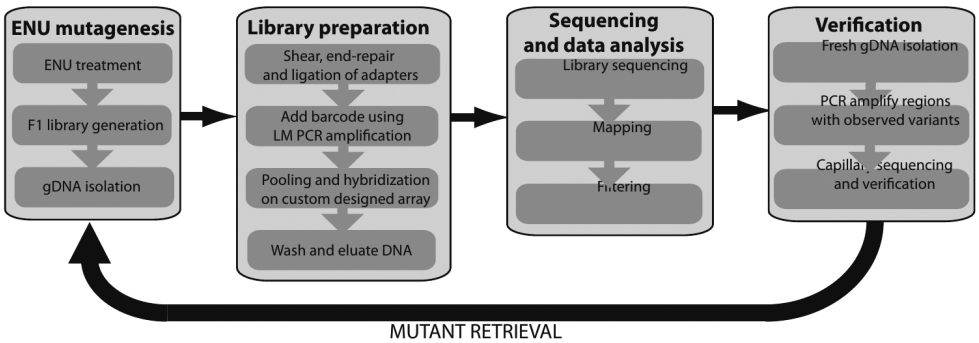
Rat sequencing data analysis and SNP calling
Sequencing reads were mapped against the reference genome (Rat Ensembl Build 56.34x) using the Maq package (15) (V0.7.1, options: -c, n3, e180, C10), which allows mapping in SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA. Raw variant positions were called by the Maq package and filtered using custom scripts. For stringent variant calling we used the following filtering settings: 1) positions with lower than 20x and higher than 5000x coverage were excluded, 2) each of non-reference alleles had to be supported by at least 3 independent reads (as determined by different read start positions) separately on positive and negative strand with quality > 10, 3) the non-reference allele should account for at least 20% of the reads covering the polymorphic position, and 4) the ratio between + and – strand reads should be between 1/9 and 9. Positions that passed these filtering settings were considered as candidate variant. Since the ENU-induced mutations are heterozygous in the assayed F1 animals, a variant was qualified as heterozygous when the fraction of non-reference alleles was between 20 % and 85 %.

Material availability
All custom scripts used in this study are available from the authors upon request or from http://www.hubrecht.eu/research/cuppen/next_gen_tilling_scripts.zip. All rat mutant models will be made available through the rat knockout consortium (www.knockoutrat.org).

REFERENCES

1. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. Nat. Methods, 7, 111-118.
2. Henikoff, S., Till, B.J. and Comai, L. (2004) TILLING. Traditional mutagenesis meets functional genomics. Plant Physiol., 135, 630-636.
3. Stemple, D.L. (2004) TILLING--a high-throughput harvest for functional genomics. Nat. Rev. Genet., 5, 145-150.
4. van Boxtel, R., Vroling, B., Toonen, P., Nijman, I.J., van Roekel, H., Verheul, M., Baakman, C., Guryev, V., Vriend, G. and Cuppen, E. (2010) Systematic generation of in vivo G protein-coupled receptor mutants in the rat. Pharmacogenomics J., doi: 10.1038/tpj.2010.44.
5. Moens, C.B., Donn, T.M., Wolf-Saxon, E.R. and Ma, T.P. (2008) Reverse genetics in zebrafish by TILLING. Brief. Funct. Genomic. Proteomic., 7, 454-459.
6. Till, B.J., Burtner, C., Comai, L. and Henikoff, S. (2004) Mismatch cleavage by single-strand specific nucleases. Nucleic Acids Res., 32, 2632-2641.
7. Mashimo, T., Yanagihara, K., Tokuda, S., Voigt, B., Takizawa, A., Nakajima, R., Kato, M., Hirabayashi, M., Kuramoto, T. and Serikawa, T. (2008) An ENU-induced mutant archive for gene targeting in rats. Nat. Genet., 40, 514-515.
8. Gady, A.L., Hermans, F.W., Van de Wal, M.H., van Loo, E.N., Visser, R.G. and Bachem, C.W. (2009) Implementation of two high throughput techniques in a novel application: detecting point mutations in large EMS mutated plant populations. Plant Methods, 5, 13.
9. Cuppen, E., Gort, E., Hazendonk, E., Mudde, J., van de Belt, J., Nijman, I.J., Guryev, V. and Plasterk, R.H. (2007) Efficient target-selected mutagenesis in Caenorhabditis elegans: toward a knockout for every gene. Genome Res., 17, 649-658.

10. van Boxtel, R., Toonen, P.W., Verheul, M., van Roekel, H.S., Nijman, I.J., Guryev, V. and Cuppen, E. (2008) Improved generation of rat gene knockouts by target-selected mutagenesis in mismatch repair-deficient animals. BMC Genomics, 9, 460.
11. Mokry, M., Feitsma, H., Nijman, I.J., Bruijn, E.d., Zaag, P.J.v.d., Guryev, V. and Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res.
12. Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W. and Hannon, G.J. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat. Protoc., 4, 960-974.
13. Okou, D.T., Locke, A.E., Steinberg, K.M., Hagen, K., Athri, P., Shetty, A.C., Patel, V. and Zwick, M.E. (2009) Combining microarray-based genomic selection (MGS) with the Illumina Genome Analyzer platform to sequence diploid target regions. Ann. Hum. Genet., 73, 502-513.
14. Hoischen, A., Gilissen, C., Arts, P., Wieskamp, N., van der Vliet, W., Vermeer, S., Steehouwer, M., de Vries, P., Meijer, R., Seiqueros, J. et al. (2010) Massively parallel sequencing of ataxia genes after array-based enrichment. Hum. Mutat., 31, 494-499.
15. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18, 1851-1858.



Supplementary Figure 1: Detailed overview of the experimental steps in the next-generation reverse genetics approach. The procedure can be divided into 4 different experimental steps. First, mutagenized individuals are outcrossed to generate an F1 library of which DNA is isolated. Second, the DNA is used to generate a sequencing library. SOLiD barcodes are introduced in order to pool F1 animals before the multiplex enrichment for genes of interest. Next, the enriched DNA is sequenced and the data is analyzed. Finally, the identified variants are verified by an independent assay, e.g. traditional capillary sequencing.

4

MULTIPLEXED ARRAY-BASED AND IN-SOLUTION GENOMIC ENRICHMENT FOR FLEXIBLE AND COST-EFFECTIVE TARGETED NEXT-GENERATION SEQUENCING

*Harakalova M, *Mokry M, Hrdlickova B, Renkens I, Duran K, van Roekel H, Lansu N, Van Roosmalen M, de Bruijn E, Nijman IJ et al:
Multiplexed array-based and in-solution genomic enrichment for flexible and cost-effective targeted next-generation sequencing
Nature Protocols 2011, ~~in press~~ 1
*equall contribution

ABSTRACT

The unprecedented increase in the throughput of DNA sequencing driven by next-generation technologies now allows efficient analysis of the complete protein-coding regions of genomes (exomes) for multiple samples in a single sequencing run. However, sample preparation and targeted enrichment of multiple samples has become a rate-limiting and costly step in high-throughput genetic analysis. Here, we present an efficient protocol for parallel library preparation, followed by targeted enrichment of pooled multiplexed barcoded samples. The procedure is compatible with microarray-based and solution-based capture approaches. The high flexibility of this method allows multiplexing of 3-5 samples for whole exome experiments, 20 samples for targeted footprints of 5Mb, and 96 samples for targeted footprints of 0.4 Mb. From library preparation to post-enrichment amplification, including hybridization time, this protocol takes 5-6 days for array-based enrichment and 3-4 days for solution-based enrichment. Our method provides a cost-effective approach for a broad range of applications, including targeted re-sequencing of large sample collections (e.g. GWAS follow-up studies), as well as whole exome or custom mini-genome sequencing projects.

INTRODUCTION

Application, protocol development, and comparison with other methods

Next-generation sequencing (NGS) technologies enable efficient high-throughput sequencing of full genomes or genomic regions of interest for individual samples (1-4). However, the interpretation of the enormous amounts of variants in complete genomes is still extremely challenging and not a routine procedure (5,6). Therefore researchers often prefer to use targeted genomic enrichment (TGE) approaches to reduce complexity by focusing on subparts of the genome (e.g., exome) (7,8). TGE requires less sequencing and computational capacity compared to whole-genome sequencing, which allows it to be cost-effective by the inclusion of more samples in a single study and often facilitates analysis and biological interpretation (9). Thanks to our relatively good understanding of protein-coding sequences, exon-centric TGE approaches have already succeeded in detecting causal variants for several diseases (10-16).

Due to the rapidly increasing throughput and accuracy of NGS technologies, it is now possible to sequence many individual exomes in a single sequencing run. Ongoing improvements to next-generation sequencers are bringing routine sequencing of complete genomes within reach. This has initiated a debate over whether targeted re-sequencing is actually cost-effective, as costs for enrichment procedures have become significant compared to the sequencing costs themselves. The multiplexing of samples could in principle meet the increased throughput of sequencers, but the available commercial solutions only support enrichment of single samples, followed by the post-enrichment introduction of barcodes/indices. Recently, we and others have shown that up to 20 barcoded samples can be multiplexed in a single enrichment reaction with no adverse effects on coverage distribution and variant calling (9,17,18). However, for sequencing of candidate genes in larger sample cohorts for research and diagnostics, the multiplexing of higher numbers of samples is desired (19).

NGS platforms support indexing of samples with up to 96 barcodes and protocols for high-throughput 96-well plate library preparation have already been established (20). Here, we describe a protocol for multiplexed targeted genomic enrichment for both microarray-based and solution-based enrichment platforms based on our previous publications that describe methods and applications for efficient mutation detection using genomic enrichment (9,21) (Fig. 1). The protocol is compatible with the SOLiD sequencing platform and allows simultaneous enrichment of 3-5 samples for a human whole-exome capture (38 Mb or 50 Mb, respectively), 20 samples for a medium size design (5Mb), or 96 samples for a 0.4Mb target in a single assay. Our approach enables the routine application of TGE in research and diagnostics for significantly lower cost and effort compared to traditional non-multiplexed approaches or whole genome sequencing. The protocol is adaptable to other enrichment methods and sequencing platforms and offers a broad range of applications in human, animal, or plant high-throughput variant detection and discovery screens. We provide a detailed description of the experimental steps. We also give recommendations for the calculation of approximate numbers of samples and sizes of target footprints to ensure sufficient coverage for accurate mutation discovery in individual samples when using a single SOLiD sequencing slide.

In short, in our protocol, genomic DNA samples are purified using standard isolation protocols, sheared to short fragments, and end-repaired for subsequent ligation to truncated adaptors. Next, individual

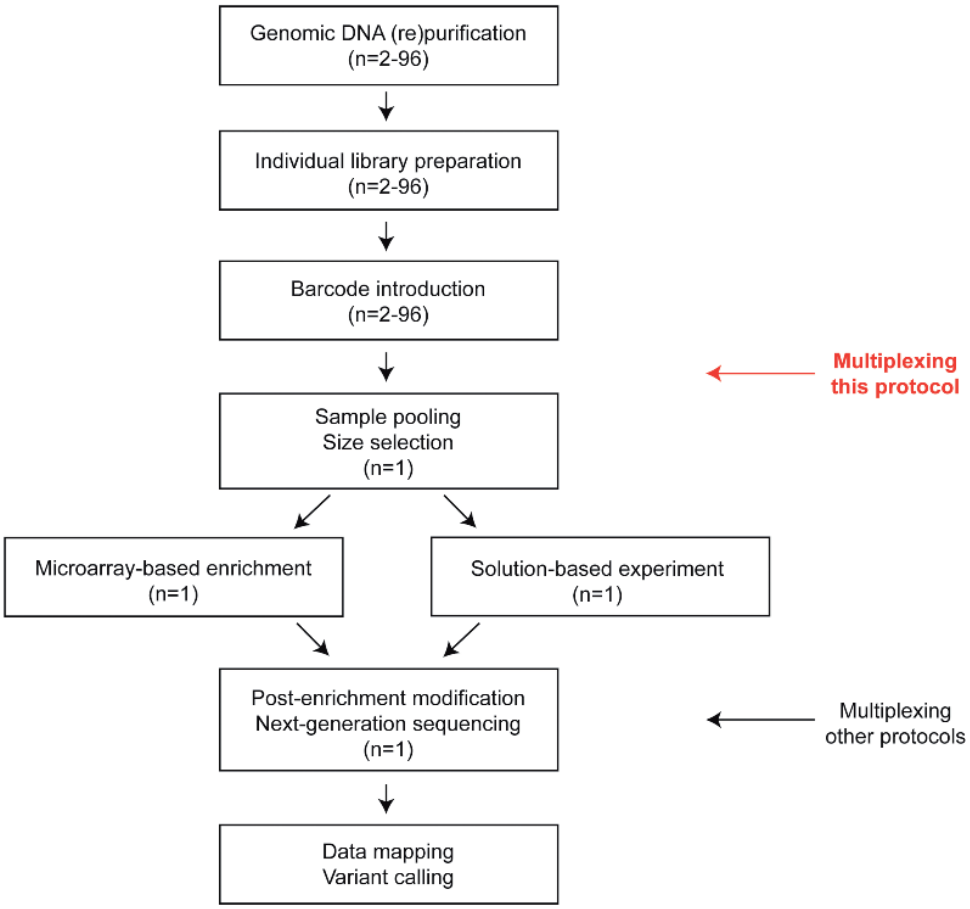


Figure 1: Workflow of the protocol for highly multiplexed enrichment. The procedure for multiplexed library preparation is compatible with microarray-based and solution-based enrichment approaches and applicable to any next-generation sequencing platform. While previous protocols supported only multiplexed sequencing runs, this protocol allows multiplexing before enrichment.

barcodes are introduced using a few PCR cycles and specific tailed oligonucleotides. Alternatively, barcoded adaptors can be ligated, removing the need for post-adaptor ligation PCR. The abovementioned steps can easily be automated on a liquid handling robot. After barcoding, samples are pooled in equimolar ratios and selected for the correct fragment sizes. For the enrichment step, the protocol provides options for both solution-based Agilent SureSelect as well as custom microarray-based capturing (21). This protocol describes the use of barcode blocking oligonucleotides designed for all possible barcode decamers, which significantly increase the multiplexed enrichment efficiency (9). Our protocol allows cost-effective, efficient, and flexible targeted genomic enrichment for next-generation sequencing-based experiments.

LIMITATIONS

Sequencing of genomic regions of interest during variant discovery focuses on pre-selected candidate genes and loci in contrast with the unbiased and hypothesis-free full genome sequencing approach. Another limitation of this method is that not all coding regions can be captured due to low complexity, the presence of pseudogenes, allelic competition, or occurrence of multiple variants in close-by sequences. In particular, library molecules with indel sizes >5 bp may have decreased hybridization efficiency and may be lost during enrichment due to competition with other molecules without indels. Moreover, current enrichment techniques cause accumulation of the coverage depth over the center of the probe, while a more even distribution would be preferred. It should be mentioned, though, that these limitations are common to all available enrichment strategies that are currently available.

EXPERIMENTAL DESIGN

Planning the experiment: The degree of multiplexing for a given experiment can be calculated by multiplying the number of reads that typically map to the genome by the read length and the enrichment efficiency that is usually obtained for the design size. Microarray-based enrichments with design sizes smaller than 1 Mb typically result in enrichment efficiencies of 30-50% (M.H., M.M., I.J.N., E.C., unpublished results), while larger design sizes (1-6 Mb) result in higher enrichment efficiencies (>60%) (9)·(21). When using solution-based human whole-exome capture (37 Mb or 50 Mb), the enrichment efficiency typically increases to 65-75% (22). The resulting number (amount of bases mapping to the target) should be divided by the size of the target design and also divided by the desired average coverage per sample. Average target coverage for heterozygous variant detection needs to be higher in enrichment experiments than in whole genome sequencing, due to uneven coverage resulting from differential capture efficiency. The resulting number of samples should be used as the upper limit of multiplexing for a single enrichment assay in combination with a single sequencing run as shown in Suppl. Fig. 1. A worked example of this formula can be calculated as follows. A typical SOLiD 4 run yields at least 350 M mapped reads of 50nt in length, giving a total amount of mapped bases of 17.5 Gbp. When the target is the human whole exome (footprint of the enrichment design is 50 Mb), typically 80% of all mapped nucleotides can be assigned to the enrichment region (on target), which totals 14 Gbp. When a single sample would be applied, this would result in an average base coverage of 280x. When a mean coverage of 50x is desired, up to 5 exomes can be multiplexed per enrichment and single sequencing slide.

Selection of enrichment method: There are various commercial solutions for TGE (7). The two most commonly used techniques are the microarray-based and solution-based methods, which are supported mainly by the Roche/NimbleGen EZCap and Agilent SureSelect or SurePrint products. In solution TGE methods work on a principle of hybridization of DNA library fragments to single stranded RNA (22) or DNA molecules (23) in a liquid phase and microarray-based methods enable hybridization to DNA probes fixed on a glass surface (8,24). In addition to these products, Halo Genomics supports enrichment of targeted fragments into circular DNA molecules and is suited for large sample cohorts for enrichment of smaller design sizes (25). Although solution-based methods are more user-friendly,

more scalable, automatable, and generate less inter-experiment variation compared to microarray-based methods, we and others have seen better performance using microarray-based enrichment as measured by evenness of coverage (18). However, microarrays have a maximum target design due to limitations in the number of probes that can be printed. Nevertheless, arrays can be ordered by the piece as custom products, making them well suited for flexible research projects, whereas in-solution designs typically require a high upfront investment for a minimum of approximately 10 assays. Here, we describe two multiplexed enrichment techniques compatible with SOLiD sequencing: microarray-based enrichment using custom-designed Agilent SurePrint arrays and solution-based Agilent SureSelect enrichment (21). When we began these experiments, we chose Agilent arrays because of their higher design flexibility; however, it should be noted that these protocols can in principle also be used or adapted for other brands of arrays or in solution capture methods.

Custom probe design for microarray based enrichments: The probes are designed on a repeat-masked target sequence using a sliding window approach [typically 2-10 base pairs (bp)], where the most optimal probe in each window is selected based on melting temperature, GC content, and homopolymer stretches, as described before (21). The resulting collection of probes is blasted to the reference genome to identify possible alternative targets. When more than one additional location with >60% identity for the complete probe length is detected, the probe is discarded. Typically, approximately 95% of coding sequence can be covered with a tiling of probes. However, this percentage can be significantly lower (70-90%) in non-coding sequences due to lower GC content or simple sequence regions. If the number of features on the array allows, we repeat the complete design multiple times with different sliding window settings to fill up the array completely, thereby allowing for better exposure of the targets to multiple slightly different probes. Probes are spotted on standard CGH arrays (Agilent) where we used both the 244 k and 1M platforms. In addition, various companies offer a portfolio with ready-to-use designs (human whole exome, exome on chromosome X, kinome, etc.). In case these products do not meet the researcher's criteria, custom design can be made using web-based tools offered by the manufacurer or by using custom scripts. We have developed Perl-based scripts for making custom microarray-based designs (9)-(21), which are available upon request from I.J.N.

Genomic DNA quality and concentration measurement: The amount and purity of the genomic DNA is a crucial factor for successful library preparation, sequencing, and subsequent data analysis and confirmation. NGS labs often have to deal with DNA samples from collaborators that use a variety of isolation processes, resulting in highly variable quality. Often, DNA samples are collected from old archives and only limited information is available on the extraction procedure that was used. Therefore, we implement an extra re-purification step at the beginning of our protocol for samples of 'unknown origin.' We prefer genomic DNA purification columns for this step because, in our hands, cleaner RNA-free and protein-free genomic DNA is obtained compared to phenol-chloroform purification, but also because the reproducibility is higher and the procedures can more easily be scaled or automated (e.g., in 96-well plate format).

The purity of DNA samples is an important factor because contaminants can inhibit enzymatic reactions and the presence of large amounts of RNA can affect DNA shearing. Furthermore, the integrity of the DNA is important. Therefore, analyzing the genomic DNA on an agarose gel, bioanalyzer,

and sometimes measurement of the same sample using DNA, RNA, and protein assays is highly recommended. It is important to measure the concentration of genomic DNA with a direct double-stranded DNA-specific binding method, preferably with the Qubit® Quantitation Platform from Invitrogen - dsDNA BR (Broad Range) Assay Kit or the Picogreen method (Invitrogen). Nanodrop or other indirect spectrophotometric methods should be avoided at any step during the library preparation or enrichment procedure since these methods are non-specific, sensitive to contamination, and in some cases can overestimate the real concentration of nucleic acid by up to 10-fold.

Highly degraded input DNA from old samples or from formalin-fixed paraffin embedded tissue can cause problems during library preparation. While libraries can be readily made from only 200 ng of high quality genomic DNA, in the presence of DNA degradation, fragments with sizes lower than 100 bp can be lost during the clean-up step after shearing and this may result in loss of up to 70% of total genomic DNA. This will decrease the complexity of the library and more PCR cycles will be required during the procedure, which may result in increased clonality after enrichment. However, low amounts of DNA are acceptable in highly multiplexed experiments, where sufficient input DNA for the enrichment step can be obtained by pooling many low concentration samples.

Adaptors and oligonucleotides used during library preparation: We use truncated versions (~20 bp long) of SOLiD sequencing adaptors for library preparation, Adaptor 1 and Adaptor 2, which allow PCR-based incorporation of barcodes after ligation. In the case of Adaptor 1, no oligonucleotide blocking is necessary during the enrichment procedure because it is only 20 bp long. Adaptors are restored to full-length sequences during post-enrichment PCR with tailed primers.

After enrichment, single stranded library fragments contain a truncated version of SOLiD Adaptor P1 and a full-length barcoded version of SOLiD Multiplexed Adaptor P2. PCR after enrichment results in double stranded DNA library fragments with the full-length adaptors required for SOLiD sequencing. We do not recommend using enriched libraries without detectable bands after more than 13 PCR cycles. It is possible that including too many PCR cycles in the library preparation process may affect the complexity of a sample. This means that some alleles may be lost (false negatives) and others overamplified (false positive homozygous calls). Unfortunately, such biases can only be detected after finishing the experiments and analyzing the data. However, in the detailed procedures, we do give indications for the number of cycles at which we never identified any significant PCR or clonality bias.

Increasing the enrichment efficiency: Enrichment efficiency is influenced by two major factors that can cause carryover of non-targeted DNA molecules (8). The first important factor is the carryover of flanking non-targeted intronic regions in exon-centric enrichment strategies. This carryover can be prevented by using sequencing libraries with shorter insert sizes (80-120 bp), which are long enough for efficient hybridization but are too short for carryover of larger flanking sequences (8,21) . Shorter insert sizes also result in better balance between captured Watson and Crick strands, which is essential for reliable SNP calling (21).

The second factor affecting enrichment efficiency is the hybridization of non-targeted DNA molecules to flanking repeat sequences or adaptors (8). To compete for repeat binding, a 5× weight excess of Cot-1 DNA can be used. Repeat-mediated cross-hybridization is also reduced in libraries with short insert sizes (80-120 bp) because by definition every library molecule can harbour less repeat-sequence in addition to

the targeted sequence. To prevent adaptor mediated cross-hybridization, truncated forms of the adaptors (18-20 bp) can be used. Adaptors are elongated to their full length by post enrichment PCR amplification using full-length adaptors as primers. However, this is not compatible with the use of barcoded libraries that have longer universal adaptor sequences.

Alternatively, oligonucleotide barcode blockers can be added in 10× weight excess to the hybridization reaction to compete for non-specific hybridization to barcode sequences (9). For barcoded libraries, adaptor-barcode constructs cannot be truncated and must be blocked with a mixture of oligonucleotides complementary to each individual adaptor-barcode construct. For short barcodes of up to 10 bp long flanked on both sides by a universal adaptor (ABi/SOLiD barcoding strategy), a single pair of oligonucleotides with degenerate bases at the barcode position can be used. Normally, we obtain an enrichment efficiency of 60-70% for singleplex enrichments, but in multiplexed setups, the efficiency is decreased to 30-35%. After implementation of a single pair of oligonucleotides (barcode blockers), we were able to increase the enrichment efficiency back to 60% (9).

Two enrichment rounds have been shown to increase enrichment efficiency, especially for smaller designs (26). However, double-round enrichment involves more amplification rounds and may increase the number of clonal reads, introduce more sequencing errors, and/or increase the unevenness of sequencing coverage (unpublished results W.P.K., I.J.N.)..

Automation of the protocol: The procedure as described here can be scaled or automated on liquid handling robots in multi-well format. 96-well format DNeasy kits are available for purification of genomic DNA. Shearing can be scaled with the Covaris™ E220 System for 96-well format, which shears one sample at a time, or with the Covaris LE220 System that shears 8 samples simultaneously. We have used both platforms successfully in combination with the protocol described here (M.H., W.P.K., E.C., unpublished data). Subsequent library preparation steps can be automated and executed on suitable liquid handling platforms with multiple samples being processed simultaneously (up to 96 in parallel). For all purification steps throughout the protocol, we recommend purification by paramagnetic beads rather than the column purification used in standard protocols (SOLiD v4 library preparation manual). The bead-based method is more user-friendly, has a lower chance for cross contamination between libraries, is easily adapted to 96-well plate automation (following the manufacturer's protocols using a 96-well magnet), and offers efficient removal of adaptor dimers and heterodimers. All enzymatic reactions, such as end-repair and phosphorylation, ligation, and PCR can be set up and performed in a 96-well format as well. The only limiting step is the volume that can be fitted in multi-well plates. However, this can be solved by using deep 96-well plates or dividing the reactions over multiple plates. Therefore, each step up to the pooling of barcoded samples is automatable and the laborious steps of size selection and enrichment are performed for the barcoded pool alone. Further details for automation are provided in BOX1.

MATERIALS

CRITICAL: Unless specifically indicated in this protocol (REAGENTS and EQUIPMENT section), the following user's guides list all equipment and reagents necessary for multiplexed enrichment and sequencing with the SOLiD platform.

» Applied Biosystems SOLiD 4 System – Templated Bead Preparation Guide, April 2010
» Applied Biosystems SOLiD 4 System – Instrument Operation Guide, April 2010
» SureSelect Target Enrichment System for SOLiD Fragment and Paired-End Sequencing Protocol, Version 1.2, September 2010

REAGENTS

» RNA-free and protein-free genomic DNA sample, 2 µg (or output of procedure described in BOX2)
» Trizma® hydrochloride (Sigma-Aldrich, cat. no. T6666)
! CAUTION Skin, eye and respiratory irritant.
» UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen, cat. no. 10977015)
» DNeasy Blood & Tissue Kit (Qiagen, cat. no. 69506) containing Proteinase K, Buffer AL, DNeasy Mini spin columns, Buffer AW1, Buffer AW2, Buffer AE
» RNase A (Qiagen, cat. no. 19101)
» Ethanol 96% (Merck, cat. no. 1009672500)
! CAUTION Highly flammable liquid and vapour.
» Quant-iT™ dsDNA BR Assay Kit (Invitrogen, cat. no. Q32853)
» Quant-iT™ dsDNA HS Assay Kit (Invitrogen, cat. no. Q32854)
» Agencourt AMPure XP 450 ml Kit (Beckman Coulter Genomics, cat. no. A63882)
» EB buffer (supplied with Qiagen MinElute PCR Purification Kit, cat. no. 28004)
» Agilent High Sensitivity DNA Kit (Agilent Technologies, cat. no. 5067-4626)
» End-It™ DNA End-Repair Kit (Epicentre Biotechnologies, cat. no. ER81050) containing End-Repair Enzyme Mix, End-Repair 10X Buffer, dNTP Solution (2.5 mM each), ATP (10 mM)
» Quick Ligation™ Kit (New England BioLabs, cat. no. M2200L) containing 2X Quick Ligation Reaction Buffer and Quick T4 DNA Ligase
» Adaptor Oligonucleotide 1A, 1B, 2A, 2B, HPLC purified, diluted to 1 mM working solution; sequences available in Supplementary Table 1 (Intergrated DNA Technologies)
» Adaptor 1 and Adaptor 2. Sequences available in Supplementary Table 1. See REAGENT SETUP for adaptor preparation (Intergrated DNA Technologies)
» Primer 1, Primer 2, Primer 3, Primer 4, desalted and lyophilized, diluted to 50 µM working solution. Sequences available in Supplementary Table 1 (Intergrated DNA Technologies)
» Platinum® PCR SuperMix (Invitrogen, cat. no. 11306-016)
» Pfu DNA Polymerase (Promega, cat. no. M7741)
» FlashGel DNA Cassette 2.2%,16+1well, double tier (Lonza, cat. no. 57032)
» FlashGel DNA Marker 50 bp-1.5 kb (500 µl; Lonza, cat. no. 57033)
» FlashGel Loading Dye (Lonza, cat no. 50462)
» Agarose MP (Roche, cat. no. 11388991001)
» Ethidium bromide (Sigma-Aldrich, cat. no. 46067)
! CAUTION Mutagen and potential carcinogen.
» Tris Base (Roche, cat. no. 10708976001)
» Boric Acid (Merck, cat. no. 1.00165.1000)
» EDTA (Sigma-Aldrich, cat. no. 6381-92-6)
» Orange G (Sigma-Aldrich, cat. no. 861286-25G)
» Glycerol (Sigma-Aldrich, cat. no. G5516)
» QIAquick Gel Extraction Kit (Qiagen, cat. no. 28706)
! CAUTION Buffer QG contains materials that cause damage to the skin; may be harmful if swallowed or inhaled.
» Barcode Block 1 and Barcode Block 2, desalted and lyophilized, diluted to 10 µg/µl working solution; sequences available in Supplementary Table 1
» NimbleGen Hybridization Kit (NimbleGen, cat. no. 05583683001)
» NimbleGen Wash Buffer Kit (NimbleGen, cat. no. 05584507001)
» Repetitive sequence fraction of DNA (depending on species):
» Human Cot-1 DNA (Invitrogen, cat. no. 15279-011)

» Mouse Cot-1 DNA (Invitrogen, cat. no. 18440-016)
» Rat Hybloc (Applied Genetics Laboratories, cat. no. RHB)
» Other species from Applied Genetics Laboratories

EQUIPMENT
» Waterbath (56°C): LAUDA Ecoline Star edition E106T (LAUDA, cat. no. LCM 0091)
» Low bind 1.5-ml centrifuge tubes (Applied Biosystems, cat. no. am12450)
» Microfuge 18 with F241.5P Rotor, 24 × 1.5-2.0 ml (VWR/Beckman Coulter, cat. no. BK367160)
» VWR Microcentrifuge Mini 115V (VWR, cat. no. 37000-700)
» Savant SpeedVac ® DNA 110 Concentrator (Savant, cat. no. DNA11-240)
» VWR Signature™ Digital Vortex Mixer 230V (VWR, cat. no. 12620-854)
» Stuart® rocker & roller mixer SRT6 (Aldrich, cat. no. Z671711)
» Qubit™ Quantification Platform (Invitrogen, cat. no. Q32860)
» Qubit™ Assay Tubes (Invitrogen, cat. no. Q32856)
» CovarisTM S2 System (Covaris, S-series)
» DynaMag-2 Magnet (Invitrogen, cat. no. 123-21D)
» Covaris Water Conditioning System for the S-series (Covaris, cat. no. 500195)
» THQ micro HOLDER (Covaris, cat. no. 500114)
» 6 × 16 mm Round bottom glass tube, AFA fiber, and (pre-slit) snap-cap system (100 µl; Covaris, cat. no. 520045)
» Thermomixer® R (42°C; Eppendorf, cat. no. 022670158)
» Exchangeable thermoblock for 24 × 1.5-ml (Eppendorf, cat. no. 022670522)
» Bioanalyzer Agilent 2100 (Agilent)
» 96-Well GeneAmp® PCR System 9700 (Life Technologies, cat. no. N8050200)
» FlashGel Dock System (Lonza, cat. no. 57025)
» Microwave
» Safe Imager™ 2.0 Blue-Light Transilluminator (Invitrogen by Lifetech, cat. no. G6600EU)
» UV transilluminator ProXima C16+ Phi (Isogen Life Science, cat. no. IM-520-0750)
! CAUTION UV radiation can cause damage to unprotected eyes and skin.
» MAUI Mixer AO, Hybridization Chamber Mixers, (Biomicro Systems, cat. no. 02-A008-10) in package with plastic bay clamp stickers
CRITICAL Other mixers may partially overlap with the probe print on the microarray.
» Filter Tips, 100 µl /200 µl (Greiner Bio-one, cat. no. 772288)
» CRITICAL Other filter tips may not fit with the port on mixer while loading.
» NimbleGen Array Processing Accessories (NimbleGen, cat. no. 05223539001) containing Slide Rack, Wash Tank and Slide Container
» NimbleGen Hybridization System 4 (220V; NimbleGen, cat. no. 05223687001)
» Compressed air
» 5317 Desiccator Cabinet (Nalgene, cat. no. 5317-0180)
» Hybridization Gasket Slide Kit (100) – 1 microarray per slide format (Agilent Technologies, cat. no. G2534-60005)
» Hybridization Chamber Kit – SureHyb enabled, Stainless (Agilent Technologies, cat. no. G2534A)
» Surgical blade

REAGENT SETUP
10 mM Tris Buffer (pH 8-9) Prepare a 10 mM Tris Buffer working solution by 50× dilution of 500 mM Trizma-hydrochloride buffer in nuclease-free water. Maximal recommended storage time and storage temperature: 3 months at room temperature (18-25°C).

70% Ethanol Prepare a stock of 70% Ethanol working solution (vol/vol) by diluting 35 ml 96% Ethanol (vol/vol) in final volume of 50 ml nuclease-free water. Maximal recommended storage time and storage temperature: 1 week at room temperature.

2% agarose gel (wt/vol) Add 2 g agarose to 100 ml of 1× TBE electrophoresis buffer in an Erlenmeyer flask. Mix well by shaking. Heat in a microwave oven until the agarose is completely melted. Occasionally shake the Erlenmeyer flask during heating to allow homogenous melting of agarose powder. Add

5 µl ethidium bromide (10 mg/ml) to 100 ml of gel for visualization of DNA after electrophoresis. After cooling to 50-60°C, pour gel into a casting tray containing a gel comb and allow it to solidify at room temperature. Use a gel comb with left and right border wells for DNA ladder and tape the middle wells to create one big single well with sufficient volume for the library pool. Leave at least one empty well between the ladder and the sample pool. Use 100 ml gels for smaller volumes and 400 ml gels for pools of up to 96 samples. Always make a fresh gel before loading samples. Use immediately, storage is not recommended.

1x TBE Buffer Prepare 10x TBE buffer stock by dissolving 108 g Tris Base, 55 g Boric Acid, and 7.4 g EDTA in a final volume of 1 L of nuclease-free water. Prior to use, dilute the 10× TBE buffer stock in nuclease-free water to obtain 1× TBE buffer working solution. Maximal recommended storage time and storage temperature: 1 month at room temperature.

10x OrangeG Loading Buffer Dissolve 100 mg of Orange G in 25 ml of nuclease-free water; add 25 ml of 100% glycerol and vortex for 1 min. Maximal recommended storage time and storage temperature: 6 months at 4ºC.

Adaptor 1 and 2 annealing: Mix complementary Adaptor Oligonucleotides (Adaptor Oligonucleotide 1A and 1B; Adaptor Oligonucleotide 2A and 2B) to get final concentration of 500 µM each and run on the thermocycler with the following program: 95°C for 3 min, 80°C for 3 min, 70°C for 3 min, 60°C for 3 min, 50°C for 3 min, 40°C for 3 min, and 4°C hold. Dilute 10-fold to obtain a 50 µM working solution. Use 10 mM Tris, pH 8 for any dilutions. Store in 50 µl aliquots. Maximal recommended storage time and storage temperature: 1 year at -20ºC. Avoid repeating freeze-thaw cycles.

EQUIPMENT SETUP

Shearing settings: Shear the genomic DNA to 100-150 bp fragments with mean size 125 bp on CovarisTM S2 System using the following settings: number of cycles: 3, bath temperature: 4°C, bath temperature limit: 8°C, mode: frequency sweeping, duty cycle: 20%, intensity: 5, cycles/burst: 200, time per cycle: 1 min and 45 s.

PROCEDURE

Measure concentration of genomic DNA prior to library preparation

TIMING 1 h per sample, 0.25 h hands-on

1. Measure concentration of a 2 µl DNA sample on the Qubit™ Quantification Platform using the dsDNA BR Assay Kit, following the manufacturer's instructions.
   CRITICAL STEP Make sure to obtain at least 2 µg RNA-free and protein-free pure DNA. If necessary, perform the re-purification on the remaining genomic DNA sample to obtain the required amount of pure DNA (see BOX2).
   CRITICAL STEP It is recommended to use a sample of water instead of genomic DNA as a negative control alongside the whole library preparation process. This sample only serves as a control for possible contamination of reagents used in library preparation process (steps 2-33) and is not included in the preparing of library pool, size selection and enrichment and sequencing procedure.
   ?TROUBLESHOOTING
2. If necessary, concentrate the genomic DNA in a speedvac at 30-40°C to 50-100 ng/µl.

Shear genomic DNA to 100-150 bp fragments with a median size of 125 bp

TIMING 0.25 h per sample, 0.15 h hands-on

3. Dilute 2 µg purified genomic DNA sample in nuclease-free water to 100 µl in a 1.5-ml centrifuge tube.
4. Transfer 100 µl diluted genomic DNA in a 6 × 16 mm AFA fiber tube. Spin down briefly.
5. Shear genomic DNA using the CovarisTM S2 System. See EQUIPMENT SETUP for shearing settings.
   If desired, reserve 1 µl DNA sample to check the size distribution of the sheared fragments on the Agilent Bioanalyzer 2100 using the Agilent High Sensitivity DNA Kit before the purification step to allow for comparison with the unsheared sample and the sample after purification with the Agencourt AMPure XP system

CRITICAL STEP Check the water level in the tank. Shearing with the water level lower than required can negatively affect the resulting fragment size.

6. Transfer the DNA sample to a new 1.5-ml centrifuge tube.

Purify the sample with the Agencourt AMPure XP system

TIMING 0.5 h per sample, 0.25 h hands-on

7. Add 2 volumes of the Agencourt AMPure XP Reagent to the sample. Mix by vortexing and incubate for 10 min at room temperature on a rock and roller or shaker with little agitation (600-650 rpm).
8. Spin down briefly. Place tube into magnetic rack and wait 2-3 min until supernatant becomes clear. Discard the supernatant. Spin down briefly once more and discard the remaining supernatant.
9. Add 500 µl freshly prepared 70% ethanol to the bead suspension. Spin down briefly. Place tube into magnetic rack. Wash beads in ethanol by turning the tube 180° horizontally in the rack and wait until beads move back to the magnet. Repeat horizontal turning of tube twice.
   CRITICAL STEP For optimal results on Agencourt AMPure XP system, prepare fresh 70% ethanol solution weekly.
10. Discard the supernatant. Spin down briefly and put back into magnetic rack to discard the remaining supernatant.
11. Add another 500 µl of freshly prepared 70% ethanol to the bead suspension. Spin down briefly. Place each tube into magnetic rack. Wash beads in ethanol by turning the tube 180° horizontally in the rack and wait until beads move back to the magnet. Repeat horizontal turning of tube twice.
12. Discard the supernatant. Spin down briefly and put again back into magnetic rack to discard the remaining supernatant.
    CRITICAL STEP Make sure all ethanol has been removed as residual ethanol may negatively affect elution efficiency and/or subsequent reactions.
13. Air dry pellet in a heat block at 42°C for 5-7 min.
    CRITICAL STEP Take care not to over-dry the beads (beads appear cracked), as this may decrease elution efficiency. Make sure all ethanol has evaporated.
14. Elute DNA from beads by resuspending the pellet in 40 µl EB buffer and vortexing. Incubate the beads for 2-3 min at room temperature. Spin down briefly.
15. Place tube into magnetic rack and collect supernatant in a new 1.5-ml centrifuge tube. Discard the old tube containing beads.
    If necessary, place the new tube with the eluate back into the magnetic rack and collect the supernatant in a new clean 1.5-ml centrifuge tube.
    At this point, 2 µl DNA sample can be reserved for concentration measurements using the Qubit™ Quantification Platform and the dsDNA HS Assay Kit, following the manufacturer's instructions. 1 µl DNA sample can be reserved to check the size distribution of the library fragments on the Agilent Bioanalyzer 2100 using the Agilent High Sensitivity DNA Kit.
    PAUSE POINT If necessary, sample can be stored at -20°C up to one month.

End-repair and phosphorylate the 5' ends of the DNA fragments

TIMING 1.75 h per sample, 0.5 h hands-on

16. Prepare the following master mix in a tube of suitable volume. Mix carefully by pipetting up and down or flicking the tube, and spin down.

Component	Amount per sample (µl)	Final
End-Repair Buffer (10×)	7.5 µl	1x
dNTP Solution (2.5 mM each)	7.5 µl	250 µM
ATP (10 mM)	7.5 µl	1 mM
End-Repair Enzyme Mix	1.0 µl	1.0 µl
Nuclease free water	11.5 µl	

    CRITICAL STEP All reagents are stored at -20°C. Thaw reagents on ice in advance. Prepare the master mix on ice.
17. Add 35 µl master mix to fragmented DNA sample in a total reaction volume of 75 µl. Mix and spin down. Incubate for 1 h at room temperature.

18. Follow the sample purification described in Steps **7-15**, including the concentration and size distribution checks.
    PAUSE POINT Store samples at -20°C for up to 1 month.

Ligate Adaptor 1 and 2 to end-repaired and 5'end-phosphorylated library fragments

TIMING 1.25 h per sample, 0.5 h hands-on

19. Prepare the following master mix in a tube of suitable volume. Mix carefully by pipetting up and down or flicking the tube, and spin down.

Component	Amount per sample (µl)	Final
Adaptor 1 (50 µM)	4 µl	2 µM
Adaptor 2 (50 µM)	4 µl	2 µM
Quick Ligation Buffer (2×)	50 µl	1 x
Quick Ligase	2 µl	2 µl

    CRITICAL STEP All reagents are stored at -20°C. Thaw reagents on ice in advance. Do not heat during thawing, as heating can cause denaturation of double stranded adaptors. Prepare master mix on ice.

20. Add 60 µl master mix to each fragmented DNA sample in a total reaction volume of 100 µl. Mix and spin down. Incubate for 30 min at room temperature.
21. Follow the sample purification described in Steps **7-15**, including the concentration and size distribution checks.
    PAUSE POINT Store sample at -20°C for up to 1 month.

Nick-translate the non-phosphorylated and non-ligated 3'-ends and add barcodes to each library by PCR

TIMING 1.5 h per sample, 0.5 h hands-on

22. Prepare the following master mix in a tube of suitable volume. Mix carefully by pipetting up and down or flicking the tube, and spin down.

Component	Amount per sample (µl)	Final
Primer 1 (50 µM)	3 µl	0.33 µM
Primer 2 (50 µM)	3 µl	0.33 µM
Platinum® PCR SuperMix	400 µl	
Pfu DNA Polymerase	1 µl	1 µl

    CRITICAL STEP All reagents are stored at -20°C. Thaw reagents on ice in advance. Prepare the master mix on ice.

23. Add 407 µl master mix to 40 µl of sample. Mix and spin down. Divide 55 µl of resulting mixture into each tube in a PCR strip (8 tubes in total). Keep on ice.
24. Amplify using the following PCR conditions:

Cycle number	Nick translate	Denature	Anneal	Extend	On hold
1	72°C, 20 min				
2		95°C, 5 min			
3-7 (5 cycles)		95°C, 15 s	54°C, 15 s	70°C, 1 min	
8				70°C, 4 min	
9					4°C

25. Add 1 µl Lonza Loading Dye to a 4 µl aliquot of each PCR reaction and transfer into one well of FlashGel DNA Cassette 2.2%. Add 2 µl Lonza Marker to the right and left border wells. It is not necessary to leave one well between the marker and library.

26. Run the FlashGel at 285 V for 3 min.
27. Check the FlashGel under Safe Imager. If a smear around 175-225 bp is visible, continue directly with Step 30. Otherwise, continue with Step 28.
28. Put the sample back into the cycler and run the following program:

Cycle number	Denature	Anneal	Extend	On hold
1-3 (3 cycles)	95°C, 15 s	54°C, 15 s	70°C, 1 min	
6			70°C, 4 min	
7				4°C

29. Repeat Steps 25-27. If you see a smear around 175-225 bp, continue directly with Step 30.
    ?TROUBLESHOOTING
30. Pool all eight PCR aliquots together in a new 1.5-ml centrifuge tube.
31. Follow the sample purification described in Steps **7-15**.
    PAUSE POINT Store samples at -20°C for up to 1 month.

Measure the concentration of amplified libraries

TIMING 0.75 h per sample, 0.25 h hands-on

32. Use 2 µl of each sample for concentration measurement using the Qubit™ Quantification Platform and the dsDNA HS Assay Kit, following the manufacturer's instructions. As accurate quantification is very important at this step to allow for subsequent equimolar pooling, each concentration measurement could optionally be done in duplicate..
33. Reserve 1 µl of each DNA sample to check the library size on the Bioanalyzer 2100 using the Agilent High Sensitivity DNA Kit.
    CRITICAL STEP Make sure that the library size is similar for all libraries in a pool.
    ?TROUBLESHOOTING

Pool barcoded samples

TIMING 0.5 h per sample, 0.25 h hands-on

34. Calculate the amount of DNA per sample needed for pooling. This step can be performed using option A or option B depending on the type of enrichment used.
    CRITICAL STEP The total amount of DNA required is 2 µg for microarray-based enrichment and 500 ng for solution-based enrichment. Since a size selection step involving DNA loss is required before the enrichment step itself, increase the amount of DNA prior to pooling.
    A Microarray-based enrichment
        i Divide 3× the total amount of DNA required per enrichment (2 µg) by the number of samples (e.g., if 2 µg are required prior to enrichment for a pool of 96 samples, use the equation (3 × 2000 ng)/96 = 62.5 ng per library).
    B Solution-based enrichment
        i Divide 3× the total amount of DNA required per enrichment (500 ng) by the number of samples (e.g., if 500 ng is required before enrichment for a pool of 5 samples, use the equation (3 × 500 ng)/5 = 300 ng per library).
    ?TROUBLESHOOTING
35. Pool all samples together into a new 1.5-ml centrifuge tube by pipetting the amount of DNA per library, according to results of the calculations in Step 34.
    CRITICAL STEP Carefully check for lack of volume in pipette tip. Double-check the expected total volume.
    PAUSE POINT Store samples at -20°C for up to 1 month or continue with the next step.

Size select the pool of barcoded libraries

TIMING 1 h per pool, 0.5 h hands on

36. Prepare a 2% agarose gel using 1× TBE buffer (see REAGENT SETUP). Agarose gels may be prepared earlier the same day to save time.
    CRITICAL STEP If the volume of the pool is too large, speedvac at 30-40°C to obtain a more suitable volume.

37. Add 1/10 volume of 10× OrangeG Loading Buffer to the sample pool and mix well by pipetting. Load 20 µl of FlashGel DNA Marker 50-1.5 kb to right and left border wells. Load the entire volume of pooled samples with loading dye into the large middle well, leaving at least one empty well between the ladder and the sample.
    CRITICAL STEP It is important to leave a well between the ladder and the sample to prevent cross-contamination.
38. Run the gel at 90-100 V for 30-40 min (or until the OrangeG is approximately 5 cm away from the starting point). View the gel on Safe Imager or UV transilluminator.
39. Excise a gel slice at the desired size for enrichment. This step can be performed using option A or option B depending on the type of enrichment used.
    A Microarray-based enrichment
        i Using a clean scalpel, excise a gel slice between 150-225 bp and transfer to a 15- or 50-ml tube, depending on the size of the gel slice.
    B Solution-based enrichment
        i Using a clean scalpel, excise a gel slice between 175-250 bp and transfer to a 15-or 50-ml tube, depending on the size of the gel slice.
    CRITICAL STEP Directly proceed to gel purification. Store the gel slice at 4°C overnight in QG buffer only if necessary (QIAquick Gel Extraction Kit).

Purify the gel slice after size selection

TIMING 0.75 h per sample, 0.5 h hands-on

40. Use the QIAquick Gel Extraction Kit to purify the DNA from the agarose slice following the manufacturer's instructions (QIAquick® Spin Handbook, March 2008) with the following critical exception: Incubate at room temperature until the gel slice has completely dissolved instead of incubating at 50°C for 10 min. Heating can impair column purification, because short DNA library molecules will become partially single stranded. Cut the gel slice into smaller pieces. To help dissolve the gel, mix on a rock and roller or shaker with little agitation (600-650 rpm) at room temperature.
    CRITICAL STEP Up to 400 mg agarose can be processed per spin column with a maximum volume of 700 µl per spin cycle. Since the gel slice from pooled sample size-selection is large, spin down the entire gel slice solution using several spin cycles on several columns, considering the agarose gel weight and spin column volume limitations.
    PAUSE POINT Store samples at -20°C for up to 1 month.

Measure the concentration of size selected library pool

TIMING 0.25 h per sample, 0.2 h hands on

41. Take 2 µl of each DNA sample to determine the concentration with the Qubit™ Quantification Platform using dsDNA HS Assay Kit following the manufacturer's instructions.
    If desired, take 1 µl of each DNA sample to check the size distribution of the size-selected fragments on bioanalyzer using the Agilent High Sensitivity DNA Kit.
    CRITICAL STEP In not enough DNA was obtained after the size selection step, perform an extra PCR amplification step as described in BOX3.

Enrich the multiplexed library pool for regions of interest

42. Perform the enrichment of the multiplexed barcoded library pool for the regions of interest. This step can be performed using option A or option B depending on the type of enrichment used.
    CRITICAL STEP Perform the steps in rapid succession. There should be no pause points during the enrichment procedure.
    A Microarray-based enrichment
    TIMING 3 days per sample, 3 h hands-on
        i Set the NimbleGen Hybridization System to 42°C. With the cover closed, allow at least 3 h for the temperature to stabilize. Be aware that the temperature of the NimbleGen Hybridization System may fluctuate.
        ii Mix the DNA library with 5× weight excess of repetitive sequence fraction of DNA and 0.5 µl Barcode Block 1 and 0.5 µl Barcode Block 2.
        iii CRITICAL STEP Use repetitive sequence fraction of DNA suited only for the species of interest [see Reagents section; e.g., for human samples use Human Cot-1 DNA from

Invitrogen (concentration 1 µg/µl)]. For 2 µg DNA library mix, use 10 µl of Human Cot-1 DNA.

iv Speedvac at 30-40°C to pellet the library pool mixed with repetitive sequence fraction of DNA.

v Resuspend the pellet in 12.3 µl MQ. Vortex and spin down.

vi Set a standard heat block (Thermomixer) to 95°C.

vii Prepare the following hybridization master mix in a new 1.5-ml centrifuge tube using the NimbleGen Hybridization Kit. Vortex to mix and spin down.

Component	Amount per array (42.5 µl total)
2× Hybridization Buffer	29.5 µl
Hybridization Component A	11.8 µl
Nuclease free water	1.2 µl

viii Add 31.7 µl hybridization master mix to 12.3 µl DNA library. Vortex well (15 s) and spin down.

ix Place the 1.5-ml centrifuge tube with the remaining 10.8 µl hybridization master mix in the NimbleGen Hybridization System. Do not discard the hybridization master mix at this step. CRITICAL STEP Remaining hybridization master mix warmed to 42°C can be useful later while loading the slide (Step 42 A (xvi)).

x Incubate the tube with DNA library pool mixed with repetitive sequence DNA and hybridization master mix at 95°C in a heat block for 5 min with closed lid.

xi Move the tube immediately from the heat block to the NimbleGen Hybridization System pre-heated to 42°C for at least 5 min or until ready for loading.

xii Take the enrichment slide out of the storage box from the dessicator cabinet. Wear gloves and touch only the edges to prevent scratching away the probe print. For best results, blow compressed air across the slide to remove any dust or debris.

xiii Open a chamber in the NimbleGen Hybridization System and place the slide in with the remaining numbered barcode sticker oriented at the right bottom side.

xiv Open the AO mixer and remove the thin adhesive gasket from the surface with a forceps or a pipette tip. Place the mixer precisely on the Agilent slide starting at 0.5-1 mm from the left side of the slide (already placed in the chamber), with the adhesive part facing the slide.
CRITICAL STEP Loading of samples should be performed within 30 min of opening the vacuum packaged AO mixer.

xv Take the slide sticking to the AO mixer out of the chamber and tightly glue the edges by applying moderate pressure with a piece of plastic. Place the slide with the AO mixer back into the chamber.
CRITICAL STEP Make sure the edges are glued well enough to prevent evaporation and formation of bubbles during loading and hybridization.

xvi Take a pipette with a Filter Tip, 100/200 µl from Greiner Bio-One set to aspirate 50 µl. Aspirate the sample and inspect the pipette tip for air bubbles. Dispense and reload the pipette if bubbles exist.

xvii Load the sample on the enrichment slide. While loading, keep the pipette tip perpendicular to the slide to avoid possible leakage at the fill port. Apply gentle pressure of the tip into the port to ensure a tight seal while loading the sample. Wait until the fluid reaches the right end of the slide.
CRITICAL STEP The volume of AO mixers is approximately 44±4 µl (every mixer is different), so be prepared to quickly add some additional hybridization solution of the remaining master mix from Step 42 A (viii) to ensure the whole surface of the slide is covered as dry parts and larger air bubbles will impair mixing and hybridization efficiency. Be careful not to insert air bubbles during this process, although a single small air bubble should not impair the quality of enrichment because the mixer is moving the fluids during the hybridization.

xviii Gently dry any overflow from the fill port with tissue and close both holes on the mixer with two plastic bay clamp stickers. Place over the holes without pressing, and then press both with fingers simultaneously to close well.

xix Close the lid and turn on the Mixing Panel on the Hybridization System and press the mix button to start mixing. Ensure the mix mode is set to B and the system recognizes each slide, as indicated by the green light. Hybridize the samples for 64-72 h.

xx After 64-72 h, prepare NimbleGen washing buffers in four wash tanks with suitable volume, according to the following scheme:

Wash tank	Wash Buffer	Nuclease-free water	Buffer Stock (10×)	Total washing volume (1×)	Washing temperature
1st	Wash Buffer 1	180 ml	20 ml	200 ml	42°C
2nd	Wash Buffer 1	270 ml	30 ml	300 ml	room temperature
3rd	Wash Buffer 2	270 ml	30 ml	300 ml	room temperature
4th	Wash Buffer 3	270 ml	30 ml	300 ml	room temperature

xxi Preheat 200 µl 1× Wash Buffer 1 to 42°C for usage in stap xxii.

xxii Double-glove both hands. Double-gloving facilitates easy removal of the outer gloves between removal of mixer and washes.

xxiii Take the enrichment slide from the chamber of the NimbleGen Hybridization System and place it directly into the wash tank with Wash Buffer 1 heated to 42°C. Gently but firmly remove the AO mixer from the enrichment slide inside the washing solution with fingers.

xxiv Move the slide immediately from the 1st tank into the slide rack in the 2nd tank (Wash Buffer 1 at room temperature) without touching the slide holder, remove outer gloves, and then actively wash the slide for 2 min (washing solution should become foamy).

xxv Move the slide rack with enrichment slide from the 2nd tank into the 3rd tank (Wash Buffer 2 at room temperature) and actively wash the slide for 1 min.

xxvi Replace gloves with a new pair. Move the slide rack with enrichment slide from the 3rd tank into the 4th tank (Wash Buffer 3 at room temperature) and actively wash the slide for 15 s. Slowly remove the array from the washing buffer and place for a few minutes on a dry space with the barcode facing down (probes facing up).

xxvii Place a gasket slide in the Agilent Hybridization Chamber device (rubber on the upper side) and pipette 800 µl nuclease-free water onto the gasket slide. Place the enrichment slide carefully on the gasket slide (barcode facing up) and lock the Agilent Hybridization Chamber device.

xxviii Incubate the Agilent Hybridization Chamber device with the slide for 30 min at 95°C in hot air incubator.

xxix Quickly dismantle the Agilent Hybridization Chamber device to release the enrichment slide and gasket slide and place them on clean aluminum foil with gasket slide placed at the bottom. Put the edge of a surgical blade between the gasket slide and enrichment slide and carefully dislodge them by turning the blade. Quickly pipette the fluid into a new 1.5-ml centrifuge tube with a pre-prepared pipette (collect approximately 400-600 µl).

xxx CRITICAL STEP Be careful not to spill out the elution fluid, since this contains the eluted enriched library fragments.

xxxi CAUTION! To avoid burns, be careful not to touch the hot metal bracket with bare hands.

xxxii Speedvac the eluted enriched library pool to 30-40 µl at 30-40°C

xxxiii PAUSE POINT Store samples at -20°C for up to 1 month.

B Solution-based enrichment

TIMING 2 days per sample, 3 h hands-on

i Add 0.5 µl Barcode Block 1 and 0.5 µl Barcode Block 2 to the size-selected library pool.

ii Speedvac at 30-40°C to precipitate the library pool and Barcode Block mix.

iii Resuspend the pellet in 3.4 µl nuclease free water. The input DNA can now be used directly for enrichment.

iv Follow steps 3-12 of the manufacturer's protocol for enrichment: pages 38-46 of SureSelect Target Enrichment System for SOLiD Fragment and Paired-End Sequencing Protocol, Version 1.2, September 2010.
PAUSE POINT Store samples at -20°C for up to 1 month.

Amplify enriched library fragments by PCR

TIMING 1.5 h per sample, 0.5 h hands-on

43. Prepare the following master mix in a tube of suitable volume. Mix carefully by pipetting up and down or flicking the tube, and then spin down.

Component	Amount per sample (µl)	Final
Primer 3 (50 µM)	1.5 µl	0.33 µM
Primer 4 (50 µM)	1.5 µl	0.33 µM
Platinum® PCR SuperMix	200 µl	
PfuTurbo® DNA Polymerase	0.5 µl	0.5 µl

44. Add 203.5 µl master mix to each sample. Mix and spin down. Divide approximately 50 µl aliquots into each of four tubes in a PCR strip. Prepare on ice.
45. Amplify using the following PCR conditions:

Cycle number	Denature	Anneal	Extend	On hold
1	95°C, 5 min			
2-11 (10 cycles)	95°C, 15 s	54°C, 15 s	70°C, 1 min	
12			70°C, 4 min	
13				4°C

CRITICAL STEP The number of cycles for amplifying the library pool depends on the size of the enrichment design. For small designs (up to 0.5 Mb) we recommend using up to 13 cycles, while for whole exome samples (50 Mb), 10 cycles should be sufficient.

46. Add 1 µl Lonza Loading Dye to 4 µl aliquot of each PCR reaction and transfer to a well of a FlashGel DNA Cassette 2.2%. Add 2 µl Lonza Marker in the right and left border wells. It is not necessary to leave one well between the marker and library.
47. Run the FlashGel at 285 V for 3 min.
48. Check the FlashGel under Safe Imager. If a smear corresponding to the size of selected library band is visible, continue directly with Step 49.
?TROUBLESHOOTING
49. Pool all PCR samples together into a new 1.5-ml centrifuge tube.
50. Follow the sample purification described in Steps **7-15**.
PAUSE POINT Store samples at -20°C for up to 1 month or continue with the next step.

Measure the concentration of library pool prior to SOLiD run preparation

TIMING 0.75 h per sample, 0.25 h hands-on

51. Use 2 µl of each DNA sample to measure the concentration using the Qubit™ Quantification Platform and dsDNA HS Assay Kit, following the manufacturer's instructions.
52. Use 1 µl of each DNA sample to check the size distribution of the enriched fragments on a bioanalyzer using the Agilent High Sensitivity DNA Kit. The enriched library pool is now ready for SOLiD sequencing.
?TROUBLESHOOTING

BOX1: AUTOMATION OF THE PROCEDURE

Additional list of equipment required for 96-well format library preparation

» DNeasy 96 Blood & Tissue Kit (Qiagen, cat. no. 69581)
» CovarisTM E220 System (Covaris, E-series)
» Multichannel pipette or liquid handling robot
» Deep 96-well plate
» Agencourt SPRI plate Super Magnet Plate (Backman Coulter Genomics, cat. no. A32782)
» Steps in this protocol that can be scaled or automated in 96-well format:
» **Step 3-6:** Shearing (CovarisTM E220 or LE220 System)
» **Step 7-15, 18, 21, 31, 50, BOX3 (9):** Purification (deep 96-well plate, Agencourt SPRIPlate Super Magnet Plate, follow the manufacturers' protocol)
» **Step 16-17:** End repair and phosphorylation (deep 96-well plate)
» **Step 19-20:** Ligation (deep 96-well plate)
» **Step 21-30, 43-49, BOX3 (1-8):** PCR reactions (deep 96-well plate)

BOX 2: RE-PURIFICATION OF ISOLATED GENOMIC DNA OF UNKNOWN QUALITY OR ISOLATION PROCESS

TIMING 0.75 h per sample, 0.5 h hands-on

1. Dissolve genomic DNA in 10 mM Tris Buffer working solution to a final volume of 200 µl in a 1.5-ml centrifuge tube. Spin down briefly.
2. Add 20 µl Proteinase K stock and 200 µl Buffer AL (without added ethanol) to the genomic DNA sample and immediately mix thoroughly by vortexing. Spin down briefly. Incubate at 56°C for 5 min.
3. Add 4 µl RNase A (100 mg/ml) to genomic DNA sample, mix by vortexing, spin down briefly, and incubate at 56°C for an additional 5 min.
4. Add 200 µl Ethanol (96-100%) to genomic DNA sample and mix thoroughly by vortexing. Spin down briefly
5. Pipet mixture into DNeasy Mini spin column placed in a 2 ml collection tube. Centrifuge at ≥6,000 × *g* (8,000 rpm) for 1 min. Discard flow-through and collection tube.
6. Place DNeasy Mini spin column in a new 2 ml collection tube, add 500 µl Buffer AW1 and centrifuge at ≥6,000 × *g* (8,000 rpm) for 1 min. Discard flow-through and collection tube.
7. Place DNeasy Mini spin column in a new 2 ml collection tube, add 500 µl Buffer AW2, and centrifuge at ≥20,000 × *g* (14,000 rpm) for 3 min to dry the DNeasy membrane. Discard flow-through and collection tube.
8. Place DNeasy Mini spin column in a new 1.5-ml centrifuge tube and pipet 200 µl Buffer AE onto the DNeasy membrane. Incubate at room temperature for 1 min, and then centrifuge for 1 min at ≥6,000 × *g* (8,000 rpm) to elute. For maximum DNA yield, elute the membrane once more into a new 1.5-ml centrifuge tube with 100 µl Buffer AE. After pooling of both eluates, the final volume is 300 µl.
   PAUSE POINT Store sample at -20°C for up to 1 month.
   CRITICAL STEP Measure DNA concentration after thawing stored samples prior to library preparation.

BOX 3: OPTIONAL PCR AMPLIFICATION TO INCREASE INPUT MATERIAL PRIOR TO ENRICHMENT

TIMING 1.5 h per sample, 0.5 h hands-on

CRITICAL Perform an additional PCR amplification only if the amount of size-selected library pool is lower than required for enrichment (2 mg for micro-array, 500 ng for Sureselect in-solution enrichment).

1. Concentrate re-purified genomic DNA in a speedvac at 30-40°C to a final volume of ≤50 µl.
2. Prepare the following master mix in a tube with suitable volume. Mix carefully by pipetting up and down or flicking the tube, and spin down.

Component	Amount per sample (µl)	final
Primer 1 (50 µM)	1.5 µl	0.33 µM
Primer 3 (50 µM)	1.5 µl	0.33 µM
Platinum® PCR SuperMix	200 µl	
PfuTurbo® DNA Polymerase	0.5 µl	0.5 µl

   CRITICAL STEP All reagents are stored at -20°C. Thaw reagents on ice in advance. Prepare the master mix on ice.
3. Add 203,5 µl master mix to each sample. Mix and spin down. Divide 50 µl of each mix into each of 4 tubes in a PCR strip. Keep on ice.
4. Amplify using the following PCR conditions:

Cycle number	Denature	Anneal	Extend	On hold
1	95°C, 5 min			
2-6 (5 cycles)	95°C, 15 s	54°C, 15 s	70°C, 1 min	
7			70°C, 4 min	
8				4°C

5. Add 1 µl Lonza Loading Dye to a 4 µl aliquot of each PCR reaction and transfer into one well of a FlashGel DNA Cassette 2.2%. Add 2 µl Lonza Marker to the right and left border wells. It is not necessary to leave one well between the marker and library.
6. Run the FlashGel at 285 V for 3 min.
7. Check the FlashGel under the Safe Imager. If a smear corresponding to the size of the selected library band is visible, continue directly with Step 8.
   ?TROUBLESHOOTING
8. Pool all PCR aliquots together into a new 1.5-ml tube.
9. Follow the sample purification described in Steps 7-15 of the main procedure protocol.
   PAUSE POINT Store samples at -20°C for up to 1 month.

Measure the concentration of amplified size selected library pool

10. TIMING 0.25 h per sample, 0,2 h hands on
11. Use 2 µl of each DNA sample to measure the concentration using the Qubit™ Quantification Platform and the dsDNA HS Assay Kit, following the manufacturer's instructions.
12. If desired, use 1 µl of each DNA sample to check the size distribution of the library fragments on a bioanalyzer using the Agilent High Sensitivity DNA Kit.

Timing

The timing of each step is calculated for a single sample; however, when processing multiple samples, the time per sample will decrease since enrichment can be performed on a pool of libraries. It is convenient to simultaneously prepare batches of 10-20 libraries and pool them when they are all ready.

Microarray-based enrichment (5-6 d)
Preparation and pooling of libraries (Step 1-35) 1-2 d
Size-selection of libraries, preparation of enrichment procedure (Step 35 - 42 A (xvii)) 1 d
Hybridization (Step 42 A (xviii)) 2 d
Washing and elution, PCR amplification, and elongation of adaptors (Step 42 A (xix)-52) 1 d
Solution-based enrichment (3-4 d)
Preparation and pooling of libraries (Step 1-35) 1-2 d
Size-selection of libraries, preparation of enrichment procedure, and hybridization (Step 35-42 B) 1 d
Washing and elution, PCR amplification, and elongation of adaptors (Step 42B-52) 1 d

TROUBLESHOOTING

See Table 1 for Troubleshooting advice

Table 1: Troubleshooting table.

Step	Problem	Possible reason	Solution
Step 1	Low amount of RNA-free and protein-free genomic DNA.	Genomic DNA is of poor quality.	1) Re-purify original DNA samples (BOX2). 2) Consider collecting new DNA samples 3) In case solution 1 or 2 have been excluded as possibilities, proceed with library preparation anyway.
Step 29	No visible band on Flash Gel after 7 PCR cycles.	More PCR cycles necessary or possible loss during Step 3-27.	1) Perform more PCR cycles anyway However; be aware of increased clonality and decreased complexity of library. 2) Repeat Step 3-27 including all the optional DNA concentration measurements to indentify the loss of DNA (possibly poor ligation efficiency).
Step 33	Low amount of DNA after purification of the PCR product though optimal amount of PCR cycles used.	Possible loss during purification steps.	1) Make sure all ethanol evaporated. 2) Make sure the ethanol solution is freshly prepared.
Step 34	Low amount of DNA prior to pooling according to the calculation.	Number of PCR cycles too low.	1) If fewer than 8 PCR cycles were performed, add additional 2-3 PCR cycles.

ANTICIPATED RESULTS

Here, we show the typical results that can be obtained by different multiplex experiment setups. The experimental setups are chosen such that the targeting foot print in combination with the number of samples matches the capacity of a single SOLiD v4 sequencing run (Figs. 2-5). Relevant parameters for judging the success of an experiment include: 1) Accuracy of equimolar sample pooling. This can be determined by calculating the distribution of total reads or from reads mapped to the reference genome per barcode (Figs. 2a, 3a, 4a, and 5a). Typically, most samples are within a 2-fold range from the median, although individual outliers may occur and are most likely due to suboptimal source DNA quality. 2) Enrichment efficiency: This is calculated per individual sample by dividing the number of reads that overlapped with the design

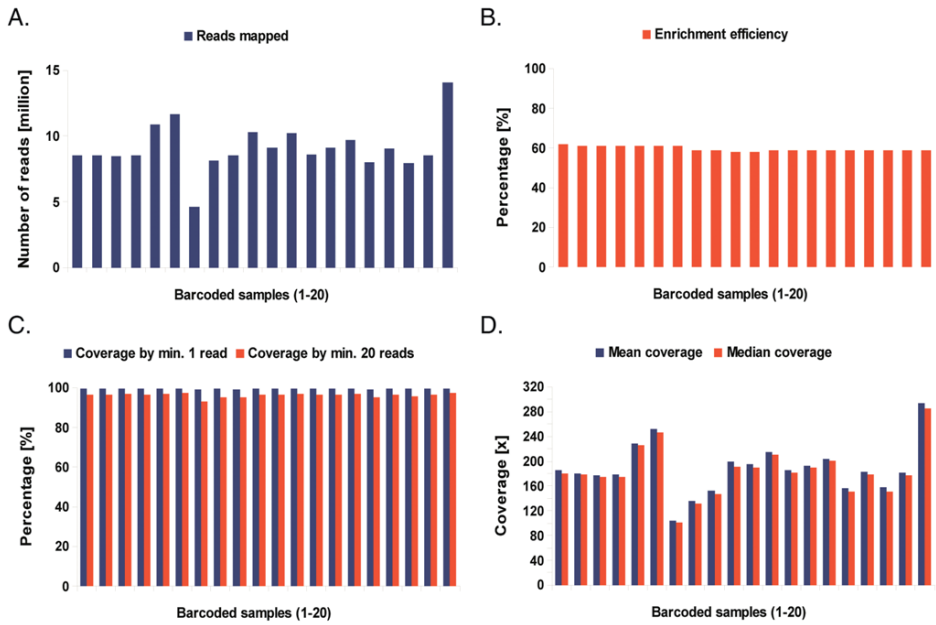


Figure 2: Mapping and enrichment statistics for a multiplexed microarray-based enrichment experiment with 20 rat samples and design size of 1.4 Mb. The figure shows an overview of the distribution of mapped reads assigned to barcodes (**A**), percentage of enrichment efficiency (**B**), design coverage by ≥1 read and ≥20 reads (**C**), and mean and median of sequencing coverage (**D**).

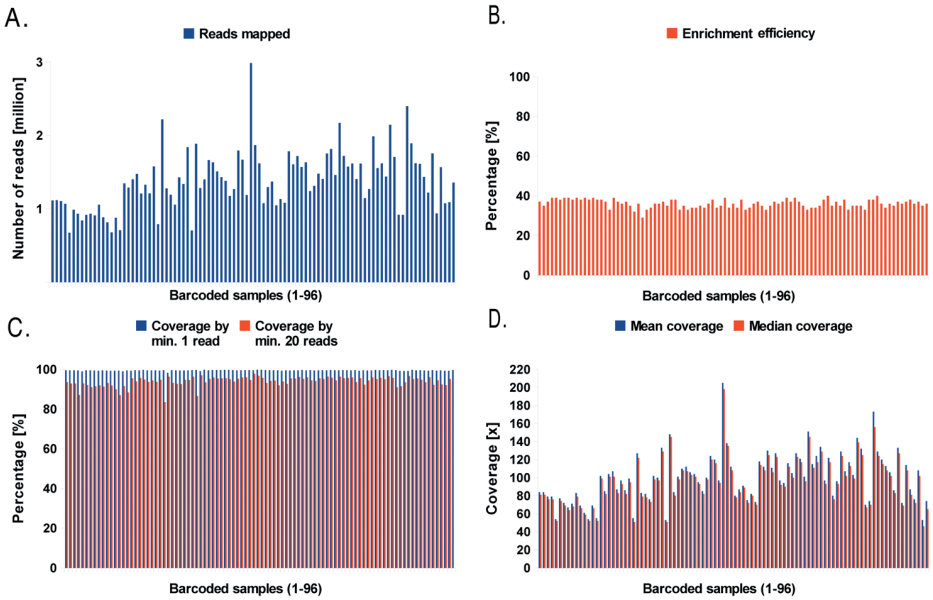


Figure 3: Mapping and enrichment statistics for a multiplexed microarray-based enrichment experiment with 96 human samples and design size of 0.4 Mb. The figure shows an overview of the distribution of mapped reads assigned to barcodes (**A**), percentage of enrichment efficiency (**B**), design coverage by ≥1 read and ≥20 reads (**C**), and mean and median of sequencing coverage (**D**).

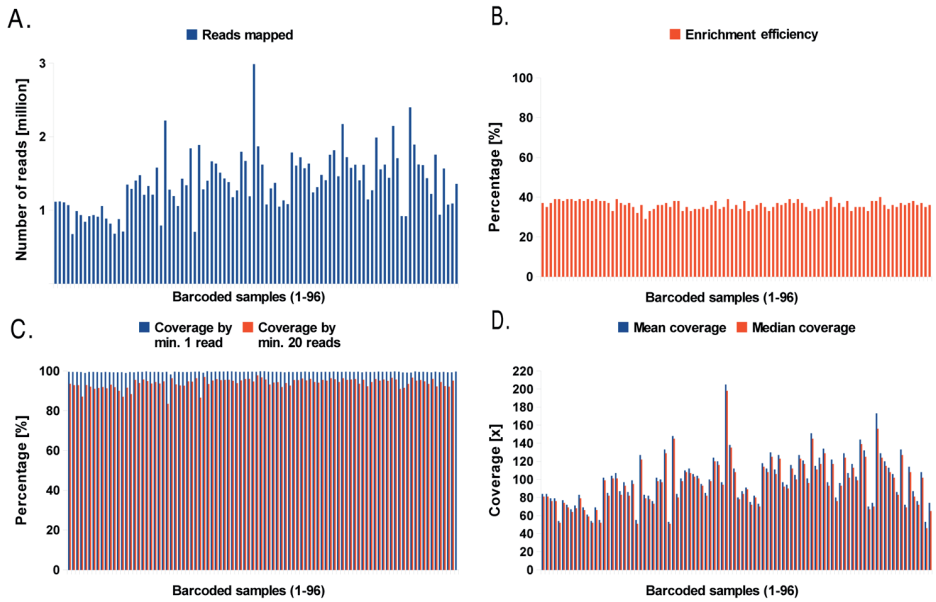


Figure 4: Mapping and enrichment statistics for a multiplexed solution-based enrichment experiment with 23 human samples and design size of 3 Mb (human exome on chromosome X). The figure shows an overview of the distribution of mapped reads assigned to barcodes (**A**), percentage of enrichment efficiency (**B**), design coverage by ≥1 read and ≥20 reads (**C**), and mean and median of sequencing coverage (**D**).

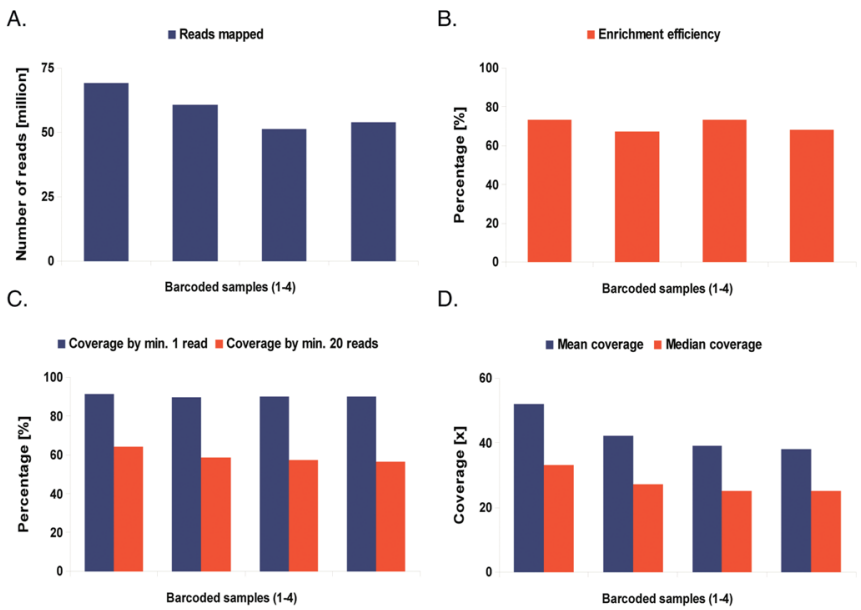


Figure 5: Mapping and enrichment statistics for a multiplexed solution-based enrichment experiment with four human samples and design size of 50 Mb (human whole exome). The figure shows an overview of the distribution of mapped reads assigned to barcodes (**A**), percentage of enrichment efficiency (**B**), design coverage by ≥1 read and ≥20 reads (**C**), and mean and median of sequencing coverage (**D**).

footprint with the total number of mapped reads (Figs. 2b, 3b, 4b, and 5b). For designs smaller than 5 Mb, the enrichment efficiency can be expected to be between 40 and 60%, while for design sizes up to 50Mb this increases to 65-75%. We have previously shown that the use of barcode blocking oligonucleotides during multiplexed enrichment significantly increases the enrichment efficiency compared to non-blocked experiments (9) coverage per target base position. The portion of the whole design covered by at least 1 read or at least 20 reads provides a good measure for completeness of the screen and ability to reliably identify heterozygous variants. Although these statistics do of course depend on the depth of sequencing, we typically aim for 50- to 100- fold average bp coverage per samples. While under those conditions similar values can be expect for the 1x coverage statistics (90-95% for in solution and >99% for array-based), values for >20x coverage can be more variable (Figs. 2c, 3c, 4c, and 5c). There are two reasons for this: first, equimolar titration of individual samples within a 2- to 3-fold range is possible but challenging. Therefore, certain samples could be covered 25×, while others are covered 125×, automatically resulting in less bases that are covered >20x in the first case. Second, the enrichment procedures introduce additional biases on top of sequence biases due to sequence-context specific effects on capture efficiency. Higher coverage sequencing could address this issue and may be required to efficiently identify heterozygous variants throughout the target region. 4) Average target coverage per sample: This is an aggregate of the statistics as indicated above at item 1) and 3). For a good experiment these statistics should mirror those of 1) and typically show a 2- to 3-fold differences between the sample with the lowest and the highest average coverage (Figs. 2d, 3d, 4d, and 5d).

One can also track the performance of individual barcodes in different experiments to detect potential sequence-specific effects, but we have never observed a systematic bias related to specific barcodes. We assume that most of the variation between the sequencing coverage is caused by differences in shearing efficiency and resultant insert size distribution differences and emulsion PCR efficiency, as well as measurement and pipetting errors. Part of this could be addressed by measuring the concentration of the specific size range fraction required for size selection rather than the total sample concentration before pooling. This could be done using for example an Agilent Bioanalyzer or 96-channel Caliper GX instrument. Low quality of genomic DNA from paraffin-embedded tissue, "old" samples or degraded DNA may also have a strong negative effect on the distribution of mapped reads, enrichment efficiency, target coverage or mean and median coverage. These mapping and enrichment statistics may become extremely uneven between indexed libraries within a pool, with up to 10-fold differences between the lower and higher value. These effects are most prominent when samples with different origin and quality are mixed in a single multiplexed experiment. Indeed, performance of a library in a pool (percentage of reads) was found to correlate with the origin and quality of the starting material. Therefore, we do recommend to only pool samples obtained from a common source.

Confirmation of the detected variants is dependent on SNP detection thresholds used. In the experiments shown here, we can typically reconfirm up to 90% of all novel variants and >90% of all known variants (dbSNP) (9) and unpublished data (M.H., I.J.N., E.C.). As there could theoretically be a competitive advantage of reference alleles to be captured above non-reference alleles, one could check the potential effect of allele frequency in the pool on the observed non-reference allele frequency per individual sample. Based on the 96-plex

enrichment experiment, we find that heterozygous calls that occur at low frequency in the total pool, do not exhibit a significant decrease in average non-reference allele % (NRA%) (R^2=0.066) or decrease in coverage (R^2<0.001) compared to heterozygous calls that occur at high frequency (Fig. 6). Similarly, homozygous calls that occur at a low frequency in a pool do not exhibit a significant decrease in average NRA% (R^2<0.001), although average coverage of low frequency homozygous calls shows a mild non-significant (R^2=0.353) decreasing trend (Figure 6). However, this effect is too small to affect the ability to reliably call the SNV, indicating that rare variants in the multiplexed pool can be detected equally reliable as more frequent variants.

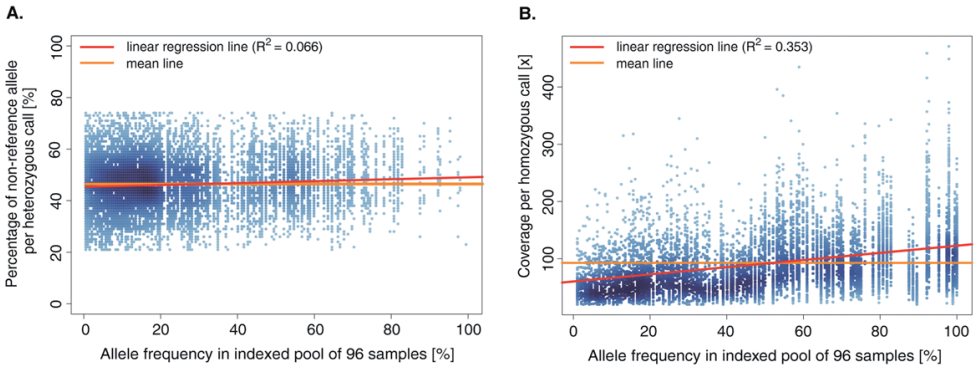
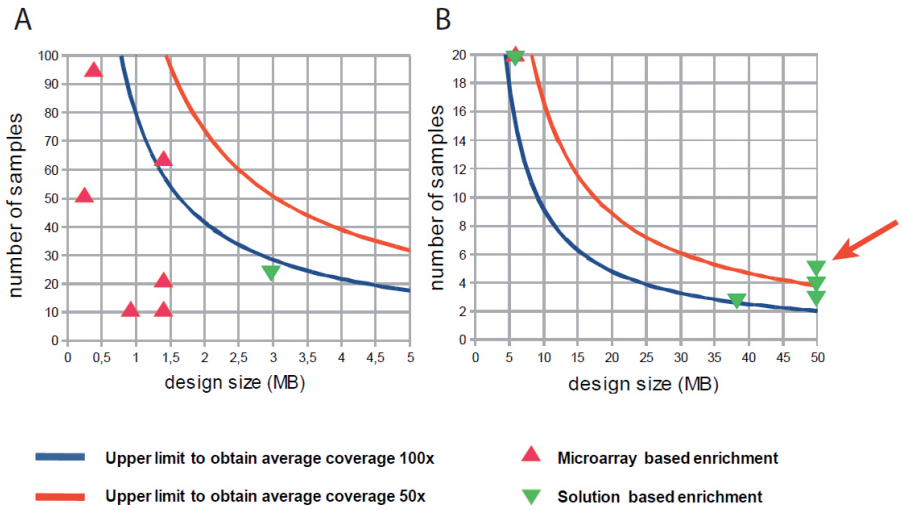


Figure 6: Effect of multiplexing level of up to 96 barcoded samples on allelic competition. The figure shows the correlation between the frequency of an allele in a pool of 96 indexed samples (192 alleles) and non-reference allele percentage (NRA%) of a heterozygous call **(A)** and frequency of an allele in a pool of 96 indexed samples (192 alleles) and coverage of a homozygous call **(B)**. A heterozygous call is defined as a call with 20-75% NRA and homozygous call is defined as a call with 75-100% NRA. Shift of the peak of heterozygous calls from expected 50% NRA was already observed also in non-multiplexed enrichment experiments as shown by Mokry et al (21).

ACKNOWLEDGEMENTS

We would like to thank Inge Wortel and Eric Slob for testing the protocol. Barbara Hrdlickova was supported by The Rector's grant MUNI/E0136/2009 provided by Masaryk University, Czech Republic.

SUPPLEMENTARY INFORMATION:



Supplementary Figure 1: Graphical example of a manual for the level of multiplexing based on the size of design. According to our observations the enrichment efficiency differs between the design sizes smaller than 5Mb (A) and designs with sizes 5-50Mb (B). Triangles indicate our already performed multiplexing experiments and based on these results we calculated the maximal possible levels of multiplexing if the average of required coverage is 50x or 100x, which are the most common used levels of sequencing coverage. Arrow indicates an experiment where we exceeded advised level of multiplexing and needed to create one additional sequencing run to obtain enough coverage.

REFERENCES

1. Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Guo, Y. et al. (2008) The diploid genome sequence of an Asian individual. Nature, 456, 60-65.
2. Li, Y., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Jiang, T., Jiang, H., Albrechtsen, A., Andersen, G., Cao, H., Korneliussen, T. et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet, 42, 969-972.
3. Metzker, M.L. (2010) Sequencing technologies - the next generation. Nat Rev Genet, 11, 31-46.
4. Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M. et al. (2010) A de novo paradigm for mental retardation. Nat Genet, 42, 1109-1112.
5. Igartua, C., Turner, E.H., Ng, S.B., Hodges, E., Hannon, G.J., Bhattacharjee, A., Rieder, M.J., Nickerson, D.A. and Shendure, J. (2010) Targeted enrichment of specific regions in the human genome by array hybridization. Curr Protoc Hum Genet, Chapter 18, Unit 18 13.
6. Durbin, R.M., Abecasis, G.R., Altshuler, D.L., Auton, A., Brooks, L.D., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. Nature, 467, 1061-1073.
7. Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J. and Turner, D.J. (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods, 7, 111-118.
8. Hodges, E., Rooks, M., Xuan, Z., Bhattacharjee, A., Benjamin Gordon, D., Brizuela, L., Richard McCombie, W. and Hannon, G.J. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc, 4, 960-974.
9. Nijman, I.J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E. and Cuppen, E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods, 7, 913-915.
10. Harbour, J.W., Onken, M.D., Roberson, E.D., Duan, S., Cao, L., Worley, L.A., Council, M.L., Matatall, K.A., Helms, C. and Bowcock, A.M. (2010) Frequent mutation of BAP1 in metastasizing uveal melanomas. Science, 330, 1410-1413.

11. Varela, I., Tarpey, P., Raine, K., Huang, D., Ong, C.K., Stephens, P., Davies, H., Jones, D., Lin, M.L., Teague, J. et al. (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. Nature, 469, 539-542.
12. Krawitz, P.M., Schweiger, M.R., Rodelsperger, C., Marcelis, C., Kolsch, U., Meisel, C., Stephani, F., Kinoshita, T., Murakami, Y., Bauer, S. et al. (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet, 42, 827-829.
13. Gilissen, C., Arts, H.H., Hoischen, A., Spruijt, L., Mans, D.A., Arts, P., van Lier, B., Steehouwer, M., van Reeuwijk, J., Kant, S.G. et al. (2010) Exome sequencing identifies WDR35 variants involved in Sensenbrenner syndrome. Am J Hum Genet, 87, 418-423.
14. Musunuru, K., Pirruccello, J.P., Do, R., Peloso, G.M., Guiducci, C., Sougnez, C., Garimella, K.V., Fisher, S., Abreu, J., Barry, A.J. et al. (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med, 363, 2220-2227.
15. Bilguvar, K., Ozturk, A.K., Louvi, A., Kwan, K.Y., Choi, M., Tatli, B., Yalnizoglu, D., Tuysuz, B., Caglayan, A.O., Gokben, S. et al. (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature, 467, 207-210.
16. Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C. et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet, 42, 790-793.
17. Cummings, N., King, R., Rickers, A., Kaspi, A., Lunke, S., Haviv, I. and Jowett, J.B. (2010) Combining target enrichment with barcode multiplexing for high throughput SNP discovery. BMC Genomics, 11, 641.
18. Teer, J.K., Bonnycastle, L.L., Chines, P.S., Hansen, N.F., Aoyama, N., Swift, A.J., Abaan, H.O., Albert, T.J., Margulies, E.H., Green, E.D. et al. (2010) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res, 20, 1420-1431.
19. Harakalova, M., Nijman, I.J., Medic, J., Mokry, M., Renkens, I., Blankensteijn, J.D., Kloosterman, W., Baas, A.F. and Cuppen, E. (2011) Genomic DNA Pooling Strategy for Next-Generation Sequencing-Based Rare Variant Discovery in Abdominal Aortic Aneurysm Regions of Interest-Challenges and Limitations. J Cardiovasc Transl Res.
20. Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L. et al. (2011) A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol, 12, R1.
21. Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V. and Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res, 38, e116.
22. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol, 27, 182-189.
23. Bainbridge, M.N., Wang, M., Burgess, D.L., Kovar, C., Rodesch, M.J., D'Ascenzo, M., Kitzman, J., Wu, Y.Q., Newsham, I., Richmond, T.A. et al. (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol, 11, R62.
24. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 4, 907-909.
25. Johansson, H., Isaksson, M., Sorqvist, E.F., Roos, F., Stenberg, J., Sjoblom, T., Botling, J., Micke, P., Edlund, K., Fredriksson, S. et al. (2011) Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res, 39, e8.
26. Lee, H., O'Connor, B.D., Merriman, B., Funari, V.A., Homer, N., Chen, Z., Cohn, D.H. and Nelson, S.F. (2009) Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. BMC Genomics, 10, 646.

5

IDENTIFICATION OF FACTORS REQUIRED FOR MERISTEM FUNCTION IN ARABIDOPSIS USING A NOVEL NEXT GENERATION SEQUENCING FAST FORWARD GENETICS APPROACH

*Mokry M, *Nijman IJ, van Dijken A, Benjamins R, Heidstra R, Scheres B, Cuppen E: **Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach**. *BMC Genomics* 2011, **12**:256.
*equall contribution

ABSTRACT

Background: Phenotype-driven forward genetic experiments are powerful approaches for linking phenotypes to genomic elements but they still involve a laborious positional cloning process. Although sequencing of complete genomes now becomes available, discriminating causal mutations from the enormous amounts of background variation remains a major challenge.

Method: To improve this, we developed a universal two step approach, named 'fast forward genetics', which combines traditional bulk segregant techniques with targeted genomic enrichment and next-generation sequencing technology

Results: As a proof of principle we successfully applied this approach to two Arabidopsis mutants and identified a novel factor required for stem cell activity.

Conclusion: We demonstrated that the 'fast forward genetics' procedure efficiently identifies a small number of testable candidate mutations. As the approach is independent of genome size, it can be applied to any model system of interest. Furthermore, we show that experiments can be multiplexed and easily scaled for the identification of multiple individual mutants in a single sequencing run.

BACKGROUND

Since the groundbreaking work of William Bateson and Thomas H. Morgan at the beginning of last century describing the key concepts of genetic linkage and linearly ordered genes on chromosomes, phenotype-driven forward genetics has developed into one of the most powerful approaches for assigning genomic elements to biological function. Many novel biological concepts have been discovered using this approach, including human disease genes like Duchenne Muscular Dystrophy (1), Huntington's (2) and Cystic Fibrosis (3) but also completely novel classes of biomolecules like microRNAs were identified this way (4,5). While many studies depended on naturally occurring mutations, the use of radiation and chemical mutagens dramatically boosted the generation of phenotypic mutants and eventually led to the first saturation screens allowing for the systematic identification of genes involved in specific processes (6).

Despite these successes, the process of positional cloning, i.e. the identification of the gene and variant(s) that causes the phenotype has remained a challenging and laborious process. Although it took until the 1980's before the gene underlying one of the most studied mutants, the Drosophila white-eye phenotype that was studied by Morgan in the 1910's, was identified (7), versatile techniques have since been developed to facilitate the cloning process. While tag-based mutagenesis approaches using e.g. transposons (8,9) or viruses (10) facilitated the cloning process, chemical-based methods using e.g. EMS or ENU (11-13) have remained most popular because of their highly random distribution and high efficiency. With the advent of next-generation sequencing technologies (14), it now has become possible to sequence complete genomes of individual mutants, thereby providing comprehensive inventories of genetic variation. However, it has also become clear that this information alone is not sufficient for unequivocally linking a phenotype to a specific genotype or single mutation. In genetically heterogeneous populations, large amounts of background variation is present, including hundreds of apparently deleterious mutations like premature stop codons in protein-coding genes (15,16). Similar problems arise in model organisms; since chemical mutagenesis typically results in many thousands of induced mutations per genome. Hence, discriminating background variation from causal mutations remains a major challenge. Indeed, a recent study in *C. elegans* demonstrated that even with the power of next-generation sequencing technologies and relatively small genome size, additional linkage information is required for mutation cloning (17). Although such data can be obtained by classical genetic mapping experiments or, alternatively, background variation could be eliminated by extensive backcrossing, these are relatively laborious processes. To address this, we developed a two-step protocol, named 'fast forward genetics', that combines a traditional bulk-segregant analysis approach (18) with state-of-the-art next-generation sequencing techniques (19). The fast forward genetics approach is highly efficient compared to traditional approaches and while we provide proof-of-principle using *Arabidopsis thaliana* it can in principle be applied to any sequenced species of interest and is largely independent of the genome size. Furthermore, we show that multiple samples can be processed simultaneously using multiplexed barcoded/indexed samples in a single sequencing run.

RESULTS

Fast forward genetics principle

First, bulk-segregant mutant and wild-type pools are generated by outcrossing the mutant ecotype to a polymorphic mapping ecotype, followed by selfing of the F1 progeny to establish the F2 generation in which the recessive phenotype segregates in a Mendelian fashion (Figure 1). From these, pools of mutant as well as wild type segregants were generated. While in the first pool, the causal mutation will be homozygous, the wild-type pool contains both heterozygous mutant as well as homozygous wild type individuals. Polymorphic alleles introduced by the mapping ecotype background, however, function as genetic markers and will be randomly distributed in both pools and appear at a frequency of ~50%, unless the marker is linked to the causal mutation. In such cases, the non-reference mapping allele will decrease gradually towards 0 % in the mutant pool upon getting closer to the causal variant, while increasing to 67% in the wild-type pool (Figure 1). Traditionally, hundreds of genetic markers (e.g. RFLPs, AFLPs, SNPs (20-22)) that are evenly distributed over the genome are tested in simple or multiplexed assays to roughly map the mutation to a chromosome or chromosomal segment. This step is normally followed by a fine-mapping step in non-pooled F2 individuals and requires additional markers in the region of interest. Finally, candidate genes are sequenced to identify the causal mutation.

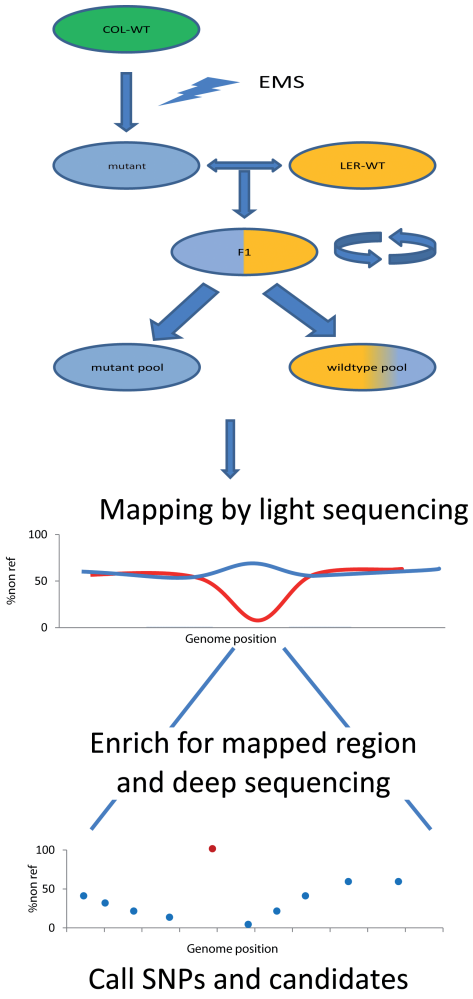


Figure 1: Schematic representation of the two step fast forward genetics approach. The mutant ecotype is crossed to a mapping ecotype background. The resulting F1 progeny are subsequently selfed to generate a F2 population in which the mutant phenotype segregates according to Mendelian rules (25% mutant, 75 % wild type). Equal numbers of 50 to 200 mutant and wild type individuals are pooled (bulk segregant pools) and used in the procedure. First, 'light sequencing (1 to 10 x genome coverage) is used to map the mutation to a genomic region, which is revealed by a strong decrease of mapping ecotype alleles in the mutant pool and a mild increase in the wild type pool. Next, a capture array is designed to capture DNA from the linked region in the mutant pool, followed by deep sequencing. This results in the identification of low frequency SNP alleles from the mapping ecotype, which allow for fine-mapping as well as abundant (homozygotic) novel non-reference alleles (red dot). These variants are prime candidates for the mutation causing the phenotype since it is the only common variant in all the individuals in the mutant pool.

In the fast forward genetics approach, all these steps are combined in a two-step experimental approach (Figure 1). When large sets of known SNPs are available (e.g. when complete genome sequences are available), the mutant pool can be 'lightly' sequenced using short read next-generation sequencing to obtain linkage information (Figure 1). Although high nucleotide coverage (>20x) is typically required for reliable genotyping by next-generation sequencing, the availability of high-density SNP panels for most commonly used model organism systems, makes it possible to use low sequence coverage (1 - 2 x) to generate medium resolution linkage information using bins of markers. Per bin of for example 100 known SNP positions, all reads covering a known polymorphic position are collected and the ratio between raw reads representing reference alleles and mapping alleles is calculated. Interestingly, this step makes the approach largely independent of the genome size. The size of the bin can be adapted to the depth of sequencing, the number of markers available and the size of the genome. While, as a consequence, the mapping resolution might be variable, it should be noted that mapping resolution depends on the number of meiotic crossovers and is thus largely determined by the size of the bulk-segregant pools.

Alternatively, when no high-density SNP data is available, the wild type pool can be sequenced at low coverage in addition to the mutant pool. The combined data can first be used to reveal variable positions between the ecotypes used. Next, these positions are typed for both pools separately and again ratios between reference and mapping alleles are calculated (Figure 1). In the study described here we applied the de-novo identification and typing of variants in the population for mapping.

After the initial mapping step, which typically results in linkage to a 5 to 10 centi-Morgan region, a capture array is designed (custom 1M or 244k Agilent SurePrint) to enrich for DNA fragments of the linked genomic region (typically 1 to 10 Mbp). In the second step the mutant bulk-segregant pool is enriched and deep-sequenced to about 5 to 10 x per allele in the pool (~1,000 x in total). This allows for simultaneous fine-mapping (step-wise decrease of mapping allele frequency due to individual cross-over events) and candidate mutation identification (fully homozygous non-reference alleles) (Figure 1).

Bulk-segregant analysis by 'light sequencing'

For a proof-of-principle, we selected two novel recessive Arabidopsis mutants. The *picup1* and *twirt1* (*twr-1*) mutants display altered root meristem function resulting in short roots, with *twr-1* mutants also affecting the shoot meristem. The mutants were generated by chemical EMS-mutagenesis of transgenic marker lines. By crossing to the Ler-1 ecotype a mapping population was generated from which the mutant as well as the wild type / heterozygous pools 200 plants each, were generated. Subsequently, DNA of the corresponding pools was converted into standard fragment libraries and sequenced in multiplex setup using barcodes that were introduced during library preparation. Next-generation sequencing was performed using AB/SOLiD technology and because of the relatively small size of the Arabidopsis genome (~120 Mbp) even the use of only about 10% of the current capacity of a single slide run results in on average 10 x genome coverage (Table 1). Sequence reads were mapped to the TAIR 8 reference genome and variable positions common to the mutant and wild-type pool are used for mapping. The mutant/mapping allele (Col/Ler) ratio was calculated for bins of 25 mapping SNPs. Simulations using part of the sequencing data showed that the linkage results are similar for even lower depth sequencing down to ~1 x coverage

Table 1: Sequencing statistics for Arabidopsis mutants using fast forward genetics

Mutant	pool	raw reads	mapped reads	on target reads (%)	Avg coverage (genome / target)
picup1	mutant	33,153,386	24,617,730		10.3x
	wild type	19,696,216	14,776,790		6.2x
	mutant enriched	51,643,460	40,185,359	35,541,544 (88%)	1,777x
twr-1	mutant	29,766,024	23,150,679		9.6x
	wild type	20,687,086	16,308,193		6.8x
	mutant enriched	46,068,250	37,584,650	32,882,209 (87%)	1,644x

(Additional file 1, Figure S1A). Below this coverage, SNP discovery becomes too unreliable and a known SNP set should be used. The number of SNPs per bin can also be varied (Additional file 1, Figure S1B). A higher number results in a smoother curve, but this is only possible when sufficient SNPs are available or discovered. When this number is limited, the effect of single SNPs becomes more pronounced and can cause large fluctuations in the frequency for a bin.

A single linkage region was identified on chromosome 1 around 21.6Mb for *picup-1* and on chromosome 5 around 22.5 Mb (Figure 2A) for *twr-1* mutant. Analysis of the wild-type pool supported these locations as the percentage of non-reference alleles increases at the same locations. In addition, the mutants were analyzed in a traditional mapping approach using a panel of PCR-based markers (not shown). The results of that experiment were fully consistent with the fast forward genetics analysis as the *picup-1* mutation on chromosome 1 between 21.6 and 21.9 Mb spanning 285 kb and the *twirt* mutation was mapped between 22.4 and 22.6 Mb on chromosome 5 spanning a 212 kb region (data not shown).

Genomic enrichment and 'deep-sequencing'

Recently, various techniques have been developed for the enrichment of specific genomic DNA fragments for down-stream next-generation sequencing (23-26). Although in principle many different platforms could be used for this step, we have chosen to use custom-designed Agilent microarrays to enrich the bulk segregant libraries, as this platform is highly flexible in terms of design and ordering. Furthermore, in combination with short-insert fragment libraries and AB/SOLiD sequencing, these arrays provide high target-enrichment rates with relatively even coverage, which was found to be instrumental for reliable heterozygous mutation discovery with high sensitivity and specificity (27,28).

A single microarray with a dense tiling set of probes targeting 1Mb of the (repeat-masked) region identified in the mapping was designed and used for the enrichment of the mutant pool. The enriched sample was sequenced using AB/SOLiD, in barcoded/multiplex way using ~ 10-15 % of the capacity of a single slide run. The enrichment was highly efficient with 87-88 % of the reads mapping to the target region of interest, resulting in 1,777 x coverage for the *picup1* mutant and 1,644 x coverage for the *twr-1* mutant (~ 4.1-4.4 x per allele in the pool) (Table 1).

To identify potential causal variants, we calculated the percentage of non-reference alleles for every informative position in the regions of interest. This percentage is expected to decrease for polymorphic mapping ecotype alleles when closing in to the causal variant and potential causal variants are expected to show up as 100%

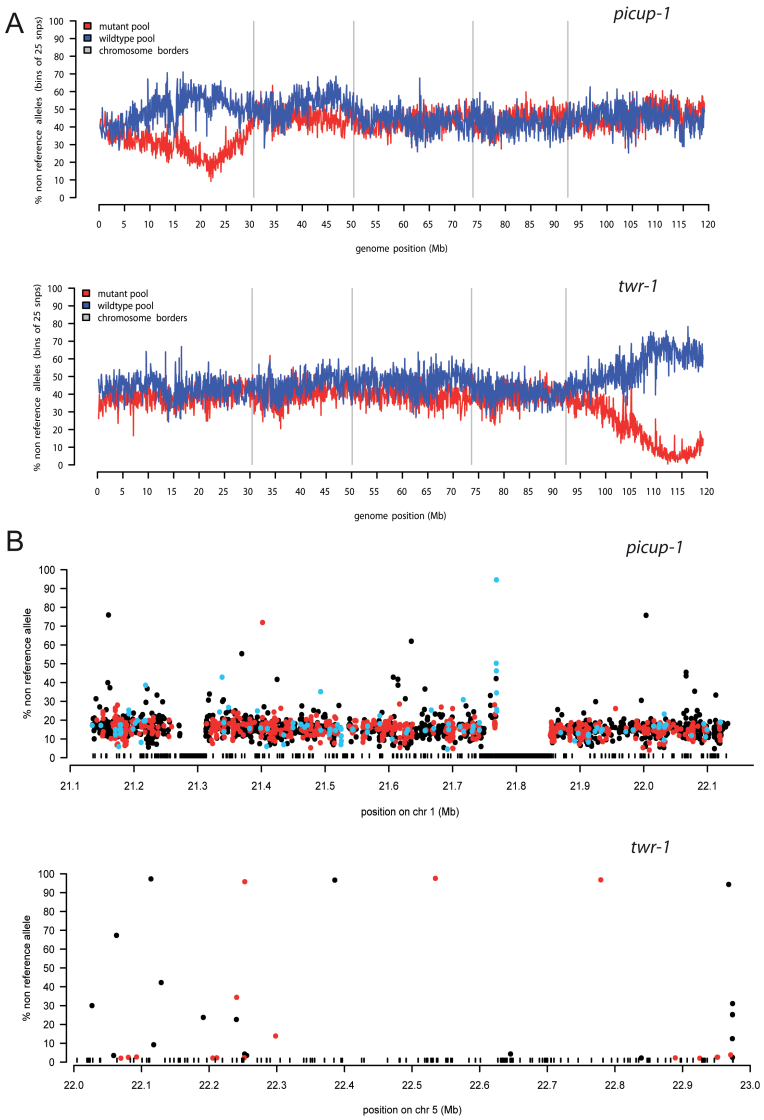


Figure 2: Results of the fast forward genetics approach: A) First step of the fast forward genetics approach. Single linkage to genomic region was identified for both mutants. The region with low frequency non-reference alleles in the mutant population indicates a common genomic fragment to the mutant pool. Verification in the wild type pool (blue graph) supports this region by showing an overrepresentation of non-reference alleles to balance the Hardy-Weinberg equilibrium. B) Second step of the fast forward genetics approach. Non-reference alleles detected in the genomic enrichment sequencing data are plotted by their location and frequency for mutants (black: non-coding, red: coding, blue: UTR). The black boxes indicate regions where no capture probes could be designed due to repeats and /or other non-unique sequences.

non-reference. To account for noise (error rates in sequencing as well as mapability issues) in the raw sequencing data as well as potential missorting of individuals in the mutant pool, relatively loose criteria (>70% non-reference alleles) were applied for candidate mutations.

Using these settings, for the *picup1* mutant, four variants were found in targeted region, two in non-coding regions, one silent and one candidate identified in the 3'-UTR of At1g58602. Although these variants could be causal mutations, the mapped region also contains a sub-region with highly repetitive sequences that was omitted from the capture design (Figure 2B). Furthermore, in this mutant 95.15% of targeted sequence was covered at more than 20x, leaving 13,391 coding bases non-inspected. A large proportion of these were in the two obvious gaps (Figure 2B) which were excluded from the array design criteria due to known repeat content or sequences present multiple times in the genome (duplications). The first gap around 21.3 Mb contains a large cluster of non-coding RNA genes and does not contain CDS regions, but the second gap around 21.8 Mb contains large transposable elements and coding sequences with high similarity matches elsewhere in the genome (some of them to mitochondrial and chloroplast). Because of their annotation as repeats (TAIR8 built) these sequences were automatically omitted from the design of the enrichment array since they would either decrease capture efficiency or introduce difficulties in mapping the sequences back to the genome.

For *twr-1* only a very limited number of candidate mutations were identified (Figure 2B, Table 2). As expected, the majority (5 out of 6) variants are G/C to A/T mutations, which is in line with the EMS-induced mutation spectrum in Arabidopsis (29). We predicted the effect of the candidates (Table2), and we found a candidate causing a premature stop codon in the gene At5g55580 (Figure 3A) making this a very likely cause for the phenotype. When we consider the additional PCR-based mapping information, it is the only candidate in the region.

On the other hand, it cannot completely be excluded that the causal mutations were missed by sequencing and are still presented in targeted regions. However, sequencing the target region for mutant showed that 99.86% of all coding bases were covered over 20x, which is sufficient for our strict SNP calling pipeline, and only 471 coding bases were not efficiently surveyed.

Mutant phenotypes and complementation test

The *picup1* mutant displays mis-localization of the normally polar localized PIN2 auxin efflux facilitator, which results in a shorter root (Figure 3E). To confirm that the causal mutation in the *picup1* mutant is residing within the identified repetitive region we ordered JatY clones (http://www.jicgenomelab.co.uk) spanning this region and used these to transform two different *picup1* alleles that came from the same mutant screen and checked for complementation in the F2 generation. Four of the JatY clones were able to complement both mutant alleles (Additional file 1, Figure S2). While these results confirm the mapping data, identification and confirmation of the causal variant will require additional experiments, which are complicated by the repetitive nature of this region and potential incompleteness of this region in the reference genome.

The *twr-1* mutant shows a reduced root growth and delayed shoot meristem activation resulting in belated outgrowth of leaves. The shoot defect was traced back to the embryo development with mature mutant embryos displaying a largely reduced shoot apical meristem compared to the dome shaped meristem in the WT (Figure 3B,C). Upon transition to flowering the plant produces fewer stems and enhanced floral

Table 2: Detected candidate mutants and their predicted effect on the genes (green: silent, yellow: synonymous, red: stop).

mutant	Chr	Position (TAIR8)	ref allele	Detected allele	gene	Predicted effect
picup-1	chr1	21,160,008	C	T	AT1G56490	pseudogene
	chr1	21,401,749	G	A	AT1G57770	S558S
	chr1	21,768,825	G	A	AT1G58602	3'UTR
	chr1	22,003,621	T	C		
twr-1	chr5	22,113,982	T	A		
	chr5	22,252,705	C	T	AT5G54730	A345T
	chr5	22,385,804	C	T		
	chr5	22,534,542	C	T	AT5G55580	Q467X
	chr5	22,779,060	C	T	AT5G56240	M405I
	chr5	22,968,138	C	T		

termination resulting in reduced seed set (Figure 3D).

To confirm that the causal mutation in the *twr-1* mutant maps to the *At5g55580* gene encoding a mitochondrial transcription termination factor (mTERF) family protein, we obtained a second allele harboring a T-DNA insertion in the fourth intron. This allele displayed phenotypes very similar to *twr-1* and subsequent complementation analysis revealed they are allelic (Figure 3D). Accordingly we renamed this mutant allele *twr-2*.

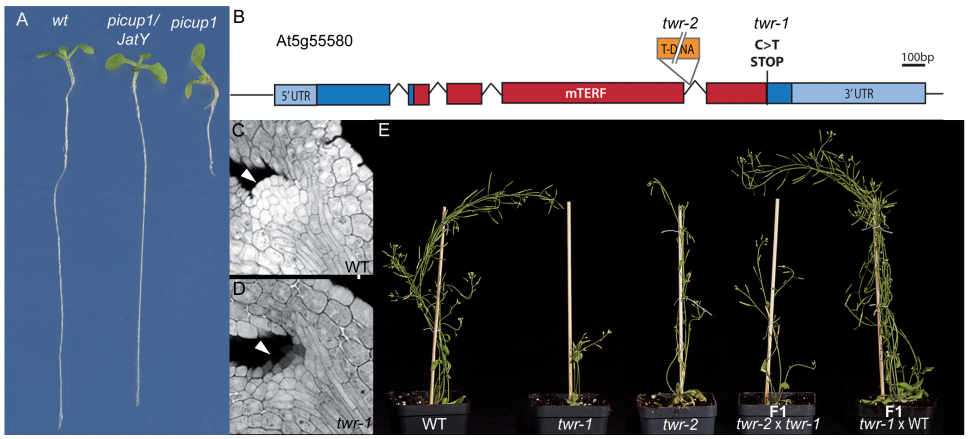


Figure 3: Identification and characterization of the mutants. A) Complementation of the *picup1* mutant by one of the JatY clones. B) Schematic representation of the TWR gene. Boxes indicate coding sequence. The conserved mTERF domain (pfam PF02536) in red, UTRs in light blue. *twr-1*: point mutation C>T causing premature stop-codon at the end of the mTERF domain. *twr-2*: T-DNA insertion in intron 4. C,D) Aniline blue staining of mature embryos showing a dome shaped group of cells in wt representing the shoot apical meristem (C), which appears absent in the *twr-1* mutant embryo (D). E) Complementation analysis of the cross between *twr-1* and *twr-2* shows identical above ground phenotypes to the single mutants. A backcross of *twr-1* to wt phenotypically resembles the wt.

DISCUSSION

We show that with the current sequencing technologies and DNA capture approaches unknown mutations can be mapped robustly to a single genomic region. Furthermore, a testable number of causal candidate mutations can routinely be obtained in a very fast and scalable two-step approach. In addition, multiple mutants can be analyzed simultaneously using a multiplexed setup with individually barcoded samples in single sequencing run. While the current capacity of next-generation sequencing equipment would already allow for the analysis of 10 to 20 Arabidopsis mutants in a single run, it is clear that the generation of mutants, bulk-segregant pools and mainly down-stream validation experiments will soon become the limiting step in forward genetics. While validation can be very laborious, the increased efficiency of the mapping and cloning steps opens up opportunities to pursue the identification of multiple independent alleles for the same locus.

The key feature of the fast forward genetics approach is the use of an old genetic trick, the use of cumulative information from bulk segregant pools. Theoretically, a pool of 100 individuals contains the information required for mapping to a resolution of about 1 cM. Translation into physical size depends on the species and chromosomal region. In general, however, a 1 cM region will typically not be larger than 1-5 Mb. While increasing pool sizes above 100 might increase mapping resolution, this would also require more sequencing depth. More importantly, however, sequencing error rates need to be significantly lower than the single allele frequency to be able to detect single cross-over events. When 500 individuals are pooled, a sequencing accuracy of 99,9 % (as currently claimed by most platforms) would mean that single allele signals equal background signal and would not be distinguishable.

Based on the known EMS-induced mutation frequency of about 1 in 100 kb (29), the regions resulting from the mapping step are expected to contain ~10 candidate mutations. We find only slightly lower numbers (four to six), which is probably due to a less efficient EMS treatment in these experiments and/or the fact that we did not efficiently cover repetitive regions. Nevertheless, fine-mapping nucleotide resolution deep-sequencing as well as bioinformatic analysis resulted in a very limited number of candidate mutations that are testable in functional assays like rescue by large-insert genomic clones or analysis of independent insertion lines from public resources (e.g. http://signal.salk.edu).

While one could in principle perform the low-resolution mapping step using traditional methods, followed by enrichment and sequencing of the non-pooled mutant ecotype genome for the linkage region, with decreasing costs of consumables, the possibility of multiplexing and increasing throughput of next generation sequencing platforms, the costs of low resolution mapping step using deep sequencing became comparable or superior to traditional methods. In any case, only a single next-generation sequencing fragment library has to be prepared and sequenced for the whole two-step procedure.

Although not strictly required for the fast forward genetics approach, the wild-type pool can be included in experiments can have advantages. First, they function as controls for the identification of real linkage locations in the mutant pools as an increase of mapping ecotype allele frequencies are expected at the same chromosomal location in the wild-type pools. Secondly, combining mutant and wild-type sequencing data allows for de novo calling of SNPs, which might be useful when no or incomplete SNP inventories are available for the ecotype

combination used. SNP discovery criteria can be set fairly stringent, as only several thousand markers genome-wide are required for the mapping step. Thirdly, sequencing data from the wild-type pool allows for the filtering of apparent candidate mutations in the second step that are actually due to genetic differences between the ecotype used for genome sequencing and the ecotype used for the mutagenesis and/or mapping.

Although every mutant requires the design of a custom microarray for enrichment, which could be both costly and take a long turnaround time, we found that microarrays can be stripped and reused several times. Therefore, depending on the size of the genome of interest only a limited set of partially overlapping enrichment arrays could be designed and generated for immediate of-the-shelf usage. For Arabidopsis a set of 10 to 15 microarrays, each targeting about 10 Mb of sequence would probably be sufficient. For larger genomes or regions, it could be considered to first design gene-centric capture arrays, although this strategy potentially ignores part of the underlying biology as certain causal mutations may be missed (e.g. in regulatory regions). Another limitation of enrichment technologies comes with challenges for probe design for capturing in repetitive regions, which as shown in case of the *picup1* mutant can potentially harbour causative mutation.

Although a successful study has been described identifying a mutation in a bulk segregant population by sequencing and mapping without using prior SNP information in a single experiment, a fairly high and complete genome coverage was necessary (30,31), making the technique challenging, especially for organisms with large genomes.

Finally, the fast forward genetics approach can be applied to any commonly used model system for which complete genome sequences are available. Sequencing requirements do not depend much on the size of a genome as mapping results in the first step largely depend on pool sizes. Candidate loci less than 10 to 20 Mb can typically be expected for any species of interest, which is compatible with the methods described here.

CONCLUSIONS

Taken together, we developed straightforward two - step approach - 'fast forward genetics' combining bulk segregant techniques with targeted genomic enrichment and next-generation sequencing technology. Method results in small number of candidate mutations which can be validated by traditional techniques.

MATERIALS AND METHODS

Plant materials, growth conditions, mutagenesis and microscopy

The *twr-1* allele was generated by EMS mutagenesis as described in (32). Similarly, the *picup1* EMS mutant was generated by screening for auxin efflux facilitator polarity defects. Seedlings and embryos were sterilized, plated and grown as described in (33,34). The Wisconsin insertion mutant WiscDsLox474E07/*twr-2* (N857510) was obtained from Nottingham Arabidopsis stock center (NASC). Primers for genotyping *twr-2*: p745-LB (Ds-Lox Wisc); AACGTCCGCAATGTGTTATTAAGTTGTC, MTERF-N857510-LP#12; CAAAACCTGGAAAAGATTGAGG, MTERF-N857510-RP#12; GGTCTTGGCATTCCTAATTCC. Aniline blue staining of mature embryos was performed as described in (35).

PCR-based mapping and generation of bulk segregant pools

Homozygous *picup1* or *twr-1* plants in Columbia-0 (Col-0) background were crossed to Landsberg

erecta (Ler-1) ecotype to create the mapping populations. For traditional mapping, *twr-1* and *picup1* mutants were selected in the F2 generation and DNA was isolated by using a CTAB-based method (36). Primers for mapping were designed using information from the CEREON collection (http://www.Arabidopsis.org/) and Primer 3 software (http://frodo.wi.mit.edu/).

A pool of 200 seedlings, each with a clear mutant phenotype (homozygous for causative mutation) and a pool of 200 seedlings showing wild-type phenotype (heterozygous for the causative mutation and wild type) were prepared and genomic DNA was isolated with the DNeasy Plant Mini Kit from QIAGEN according to manufacturer's protocol.

Preparation of SOLiD libraries

Genomic DNA (1µg) form the BSA pools was fragmented to ~ 100 nt double-stranded DNA fragments by sonication (Covaris S2, 6 x 16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping, 6 minutes). After fragmentation, fragments were blunt-ended and phosophorylated at 5'-prime end using End-it Kit (Epicentre) according to the manufacturer's instructions. Ligation of double stranded adapters (adapter 1: pre-annealed duplex of 5'-CTA TGG GCA GTC GGT GAT-3' and 5'-ATC ACC GAC TGC CCA TAG TTT-3' and adapter 2: pre-annealed duplex of 5'-CGC CTT GGC CGT ACA GCA G-3' and 5'-GCT GTA CGG CCA AGG CG-3'; all oligonucleotides were acquired through Integrated DNA Technologies (Coralville, IA) and pre-annealing was done by mixing complementary oligonucleotides at 500 µM concentration and running on thermocycler with the following program: 95˚C for 3 min, 80˚C for 3 min, 70˚C for 3 min, 60˚C for 3 min, 50˚C for 3 min, 40˚C for 3 min and 4˚C hold), compatible with SOLiD sequencing was performed using Quick ligation kit (New England Biolabs) with 750 mM adaptor 1 and adaptor 2, 150 µl of 2x Quick ligation buffer, 5 µl Quick Ligase in a total volume of 300 µl. Samples were purified using Ampure beads (Agencourt) and amplified in 400 µl of Platinum PCR Supermix with 750 mM of both amplification PCR primers, 2.5 U of Pfu DNA polymerase (Stratagene) and 5 U Taq DNA polymerase (Bioline) (nick translation 72˚C for 5 minutes, activation 95˚C for 5 min, 8 cycles: 95˚C for 15s, 54˚C for 15 s and 70˚C for 45s). After 8 cycles of amplification, library DNA was purified on Ampure beads and size selected on 4% agarose gel for 125-150bp fraction.

Array hybridization and elution

Prior to hybridization 100ng of stock library was amplified (95ºC for 5 min, 10 cycles: 95ºC for 15s, 54ºC for 15 s and 70ºC for 45s) in 1000 µl of Platinum PCR Supermix with 750 mM of both amplification PCR primers and 6.25 U of Pfu DNA polymerase (Stratagene) to produce amount of DNA necessary for enrichment (3µg) and purified using MinElute Reaction Cleanup Kit (Qiagen). Amplified DNA was concentrated by speedvac together with 20x weight excess of salmon sperm DNA and resuspended in 12.3 µl of water. DNA was mixed with 31.7 µl of Nimblegen aCGH hybridization solution and denatured at 95ºC for 5 minutes. After denaturing the sample was hybridized for 72 hours at 42ºC on MAUI hybridization station. After hybridization, the array was washed using Nimblegen Wash Buffer Kit according the user's guide for aCGH hybridization. Elution was performed using 800 µl of elution buffer (10 mM Tris pH 8.0) in an Agilent Microarray Hybridization Chamber at 95ºC for 30 minutes. After 30 minutes the chamber was quickly dissembled and elution buffer was collected. Eluted library DNA was concentrated by speedvac to a volume of 50 µl mixed with 400 µl of Platinum PCR Supermix with 750 mM of both full length amplification PCR primers and 2.5 U of Pfu DNA polymerase (Stratagene) and amplified (activation 95ºC for 5 min, 13 cycles: 95ºC for 15s, 54ºC for 15 s and 70ºC for 45s).

SOLiD Sequencing

We performed deep sequencing of enriched barcoded samples on an AB/SOLiD sequencer (Applied Biosystems) with V3 chemistry according to the manufacturer instructions to produce 50 bp long sequencing reads.

Design of enrichment arrays

Capture probes were designed on the repeat-masked sequence of the Arabidopsis reference genome (TAIR8) with a custom PERL script which selects the best 60-mer probe in a local sliding window based on criteria such as melting temperature, mono-nucleotide stretches, GC content. This design window was moved along the sequence with a 2 bp interval, resulting in a dense tiling design of probes. For the reverse strand, a completely independent design was made by offsetting the start point of the first design window. The resulting probe sequences are blasted against the reference genome, and probes with more than 2 hits, with more than 60% match are discarded from the design. The designs were uploaded through Agilent's design website E-Array (https://

earray.chem.agilent.com/earray) and produced on a 1M CGH array. The coordinates for the enriched region are: chr1:21,133,794_22,133,794 (mutant *picup1*, design window slide 2bp, 347,654 forward, 347,670 reverse probes) and chr5:22,000,600_23,000,600 (mutant *twr-1,* design window slide 2 bp, 426,762 forward, 426,731 reverse probes).

Analysis of SOLiD sequence data

Mapping sequences to the full reference genome

SOLiD sequence tags were mapped to the full genome with the MAQ package (37) (V0.7.1, options: -c, n3, e180, C10) on the Arabidopsis thaliana reference genome (build TAIR8).

SNP calling

A custom PERL script was developed to parse the MAQ pileup output and extract SNP genotypes with strict criteria: a minimal coverage of 20, more than 3 unique start sites of reads per allele, variant calls should have a quality higher than 10 and all variant alleles should be called from both strands. A maximum number of identical reads calling the same allele is set to the twice the number of individuals in the pools to suppress clonality effects. Data from both the homozygous and wild type pools is combined in the mapping phase, (creating a virtual F1), mapped and variable positions are determined. These positions are genotyped in the separate pools by either the SNP calling script or a script that checks whether a position is fully reference. Non-reference allele frequencies are calculated in windows of 25 SNPs and plotted.

Raw sequencing data are deposited on GEO archive with accession number: GSE24511. All custom scripts used in this study are available from the authors upon request.

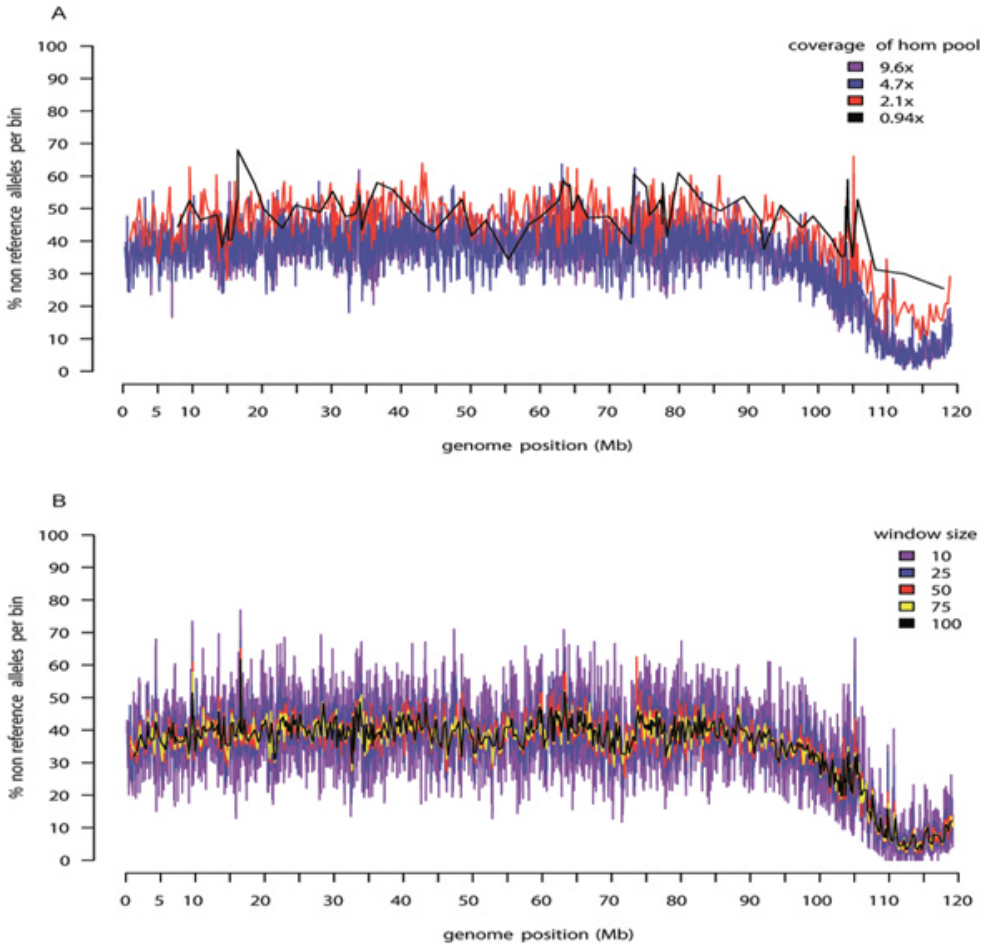
ACKNOWLEDGEMENTS

This work was financially supported by the Netherlands Bioinformatics Center through the BioRange granting scheme. RB was supported by a VENI fellowship from NWO and RH was supported by a NWO Horizon grant (050-71-054).

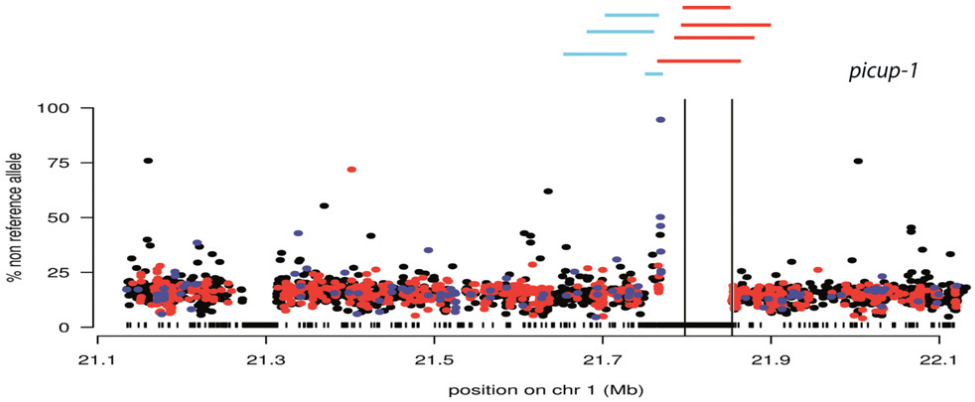
REFERENCES

1. Murray, J.M., Davies, K.E., Harper, P.S., Meredith, L., Mueller, C.R. and Williamson, R. (1982) Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. Nature, 300, 69-71.
2. (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell, 72, 971-983.
3. Rommens, J.M., Iannuzzi, M.C., Kerem, B., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N. et al. (1989) Identification of the cystic fibrosis gene: chromosome walking and jumping. Science, 245, 1059-1065.
4. Wightman, B., Ha, I. and Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell, 75, 855-862.
5. Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell, 75, 843-854.
6. Nusslein-Volhard, C. and Wieschaus, E. (1980) Mutations affecting segment number and polarity in Drosophila. Nature, 287, 795-801.
7. Goldberg, M.L., Paro, R. and Gehring, W.J. (1982) Molecular cloning of the white locus region of Drosophila melanogaster using a large transposable element. EMBO J, 1, 93-98.
8. Duverger, Y., Belougne, J., Scaglione, S., Brandli, D., Beclin, C. and Ewbank, J.J. (2007) A semi-automated high-throughput approach to the generation of transposon insertion mutants in the nematode Caenorhabditis elegans. Nucleic Acids Res, 35, e11.
9. Sivasubbu, S., Balciunas, D., Amsterdam, A. and Ekker, S.C. (2007) Insertional mutagenesis strategies in zebrafish. Genome Biol, 8 Suppl 1, S9.
10. Jao, L.E., Maddison, L., Chen, W. and Burgess, S.M. (2008) Using retroviruses as a mutagenesis tool to explore the zebrafish genome. Brief Funct Genomic Proteomic.
11. Cuppen, E., Gort, E., Hazendonk, E., Mudde, J., van de Belt, J., Nijman, I.J., Guryev, V. and Plasterk, R.H. (2007) Efficient target-selected mutagenesis in Caenorhabditis elegans: toward a knockout for every gene. Genome Res, 17, 649-658.
12. Kim, Y., Schumaker, K.S. and Zhu, J.K. (2006) EMS mutagenesis of Arabidopsis. Methods Mol Biol, 323, 101-103.

13. Maple, J. and Moller, S.G. (2007) Mutagenesis in Arabidopsis. Methods Mol Biol, 362, 197-206.
14. Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. Nat Biotechnol, 26, 1135-1145.
15. Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L. and Venter, J.C. (2008) Genetic variation in an individual human exome. PLoS Genet, 4, e1000160.
16. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461, 272-276.
17. Sarin, S., Prabhu, S., O'Meara, M.M., Pe'er, I. and Hobert, O. (2008) Caenorhabditis elegans mutant allele identification by whole-genome sequencing. Nat Methods, 5, 865-867.
18. Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A, 88, 9828-9832.
19. Darby, A.C. and Hall, N. (2008) Fast forward genetics. Nat Biotechnol, 26, 1248-1249.
20. Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M. et al. (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res, 23, 4407-4414.
21. Botstein, D., White, R.L., Skolnick, M. and Davis, R.W. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet, 32, 314-331.
22. Williams, J.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res, 18, 6531-6535.
23. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods, 4, 903-905.
24. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 4, 907-909.
25. Schracke, N., Kornmeyer, T., Kranzle, M., Stahler, P.F., Summerer, D. and Beier, M. (2009) Specific sequence selection and next generation resequencing of 68 E. coli genes using HybSelect. N Biotechnol, 26, 229-233.
26. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E.M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C. et al. (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol, 27, 182-189.
27. Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V. and Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res, 38, e116.
28. Nijman, I.J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E. and Cuppen, E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods, 7, 913-915.
29. Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E. et al. (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Res, 13, 524-530.
30. Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.E., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods, 6, 550-551.
31. Doitsidou, M., Poole, R.J., Sarin, S., Bigelow, H. and Hobert, O. (2010) C. elegans mutant identification with a one-step whole-genome-sequencing and SNP mapping strategy. PLoS ONE, 5, e15435.
32. ten Hove, C.A., Willemsen, V., de Vries, W.J., van Dijken, A., Scheres, B. and Heidstra, R. (2010) SCHIZORIZA encodes a nuclear factor regulating asymmetry of stem cell divisions in the Arabidopsis root. Curr Biol, 20, 452-457.
33. Sabatini, S., Heidstra, R., Wildwater, M. and Scheres, B. (2003) SCARECROW is involved in positioning the stem cell niche in the Arabidopsis root meristem. Genes Dev, 17, 354-358.
34. Scheres, B., Di Laurenzio, L., Willemsen, V., Hauser, M.T., Janmaat, K., Weisbeek, P. and Benfey, P.N. (1995) Mutations affecting the radial organisation of the Arabidopsis root display specific defects throughout the embryonic axis. Development, 121, 53-62.
35. Bougourd, S., Marrison, J. and Haseloff, J. (2000) Technical advance: an aniline blue staining procedure for confocal microscopy and 3D imaging of normal and perturbed cellular phenotypes in mature Arabidopsis embryos. Plant J, 24, 543-550.
36. Doyle, J.J. and Doyle, J.L. (1990) Isolation of plant DNA from fresh tissue. Focus, 12, 13-15.
37. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 18, 1851-1858.



Supplementary Figure 1: A) Simulation of the effect of reducing sequencing coverage depth on mapping results. B) Simulation of the effect of the number of SNPs per bin on mapping results.



Supplementary Figure 2: Complementation of *picup-1* mutant with YatY clones. Causal variant of *picup*-1 mutant is most likely located in large repetitive region as shown by vertical lines which represent overlapping part of the JatY clones positive in complementation experiments (blue horizontal lines depict 4 clones that did no complement; red horizontal lines depict 4 clones positive in complementation experiments). Non-reference alleles detected in the genomic enrichment sequencing data are plotted by their location and frequency for mutants (black: non-coding, red: coding, blue: UTR). The black boxes indicate repetitive regions.

6

EFFICIENT DOUBLE FRAGMENTATION CHIP-SEQ PROVIDES NUCLEOTIDE RESOLUTION PROTEIN-DNA BINDING PROFILES

Mokry M, Hatzis P, de Bruijn E, Koster J, Versteeg R, Schuijers J, van de Wetering M, Guryev V, Clevers H, Cuppen E: **Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles**. *PLoS One* 2010, **5**(11):e15092.

ABSTRACT

Immunoprecipitated crosslinked protein-DNA fragments typically range in size from several hundred to several thousand base pairs, with a significant part of chromatin being much longer than the optimal length for next-generation sequencing (NGS) procedures. Because these larger fragments may be non-random and represent relevant biology that may otherwise be missed, but also because they represent a significant fraction of the immunoprecipitated material, we designed a double-fragmentation ChIP-seq procedure. After conventional crosslinking and immunoprecipitation, chromatin is de-crosslinked and sheared a second time to concentrate fragments in the optimal size range for NGS. Besides the benefits of increased chromatin yields, the procedure also eliminates a laborious size-selection step. We show that the double-fragmentation ChIP-seq approach allows for the generation of biologically relevant genome-wide protein-DNA binding profiles from sub-nanogram amounts of TCF7L2/TCF4, TBP and H3K4me3 immunoprecipitated material. Although optimized for the AB/SOLiD platform, the same approach may be applied to other platforms.

INTRODUCTION

ChIP-seq has become the method of choice for studying functional DNA-protein interactions on a genome-wide scale. The method is based on the co-immunoprecipitation of DNA binding proteins with formaldehyde cross-linked DNA, followed by deep-sequencing of the immunoprecipitated chromatin fragments. This allows for the genome-wide identification of binding sites with high accuracy (1-5). Typical immunoprecipitated DNA fragments range in size from several hundred to several thousand base pairs. As a result, a significant part of the chromatin is not in the optimal size range for direct application to next-generation sequencing (Fig 1A). In current ChIP-seq approaches immunoprecipitated DNA fragments within the optimal sequencing range (100-200 base pairs for AB/SOLiD or 300 - 500 for Solexa/Illumina) are typically size-selected by gel-excision and converted into sequencing libraries followed by next-generation sequencing. However, this approach discards large amounts of specifically immunoprecipitated material in the larger size range, thereby increasing the demands on the amount of starting material. Furthermore, it could be possible that the observed size distribution is not random and reflects specific biology (6).

To address these limitations we applied a strategy with a first gentle shearing step before immunoprecipitation and a second more intensive shearing of purified de-crosslinked DNA after immunoprecipitation to additionally fragment all material into small fragments suitable for next-generation sequencing. We have optimized our protocols for sequencing on the SOLiD/AB platform, with an optimal fragment size of 100 - 200 bp, but this size range can be adapted at will. Furthermore, we show that the size range after the second fragmentation step is so narrow that it is possible to skip a laborious size selection in the library preparation procedure.

To demonstrate general utility, we performed ChIP-seq according to this protocol for well-characterized factors such as TBP, H3K4me3, and TCF7L2/TCF4, one of the members of the Tcf/Lef family of Wnt pathway effectors (7-9). Consensus TCF4 binding sites have been biochemically determined (9) and genome-wide binding profiles for TCF4 in colon cancer cells have been determined previously by ChIP-on-chip experiments (10). The results obtained here are in strong concordance with these previous results.

RESULTS AND DISCUSSION

Double fragmentation ChIP method

When shearing cross-linked DNA, a significant part of cross-linked chromatin remains too long for direct processing for next-generation sequencing, irrespectively of the fragmentation procedure (Fig 1A). As a consequence, a major part of the immunoprecipitated chromatin would be discarded after size selection and increases the required amount of starting material. Therefore, we introduced a double-fragmentation method for processing ChIP–seq samples with a second intensive shearing of decrosslinked immunoprecipitated chromatin into fragment lengths that are optimal for next-generation sequencing platforms (Fig 1B). This approach provides the possibility to use less starting material compared to conventional methods, as virtually all DNA is concentrated in the desired optimal size range for downstream processing. Although the size range of the second shearing step can be adjusted to any size range between 100 and 500 bp, we have focused on optimization for the

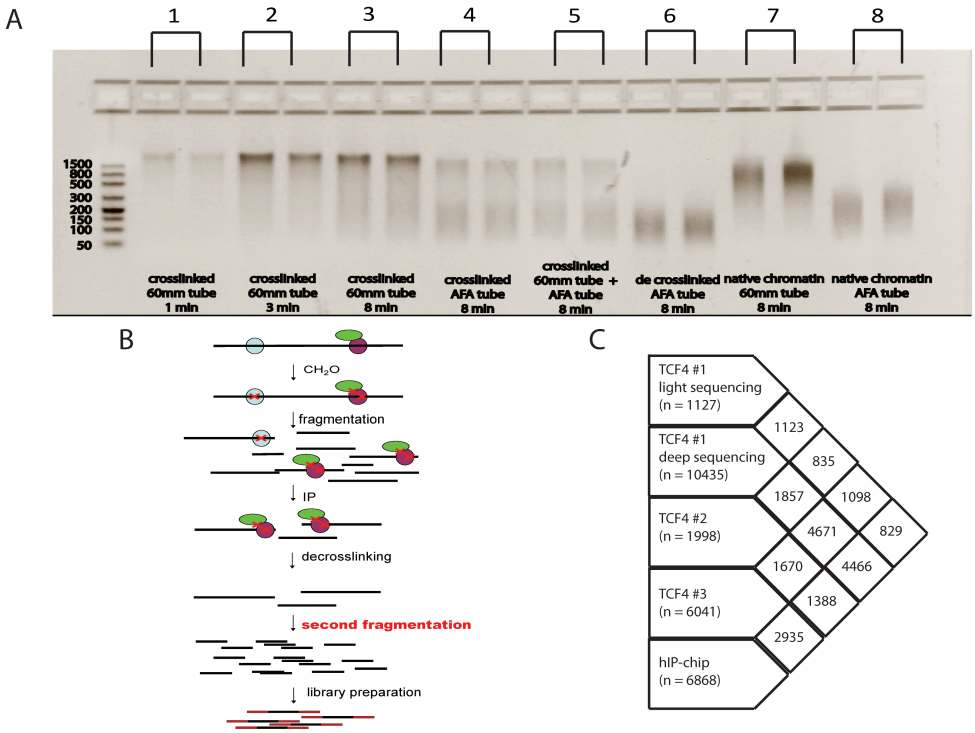


Fig 1: Double fragmentation ChIP-seq approach. A) Comparison of different shearing methods on crosslinked, de-crosslinked and native chromatin. Samples 1-3 represent crosslinked chromatin sheared at the same power intensity with increasing shearing times in 60mm tubes, sample 4 is crosslinked chromatin sheared using AFA tubes (Covaris), sample 5 is crosslinked chromatin sheared using 60 mm tubes and subsequently sheared in AFA tubes, sample 6 is crosslinked chromatin sheared in 60mm tubes, de-crosslinked and subsequently sheared in AFA tubes, samples 7 and 8 are samples of native chromatin sheared using 60mm tubes and AFA tubes, respectively. Extensive shearing of crosslinked chromatin (e.g. sample 5) still leaves a significant proportion of chromatin fragments outside the optimal range for next-generation sequencing. However, this fraction can be sheared to smaller fragments after de-crosslinking (sample 6), but not without de-crosslinking (sample 5). B) Schematic overview of the double fragmentation ChIP-seq procedure. After normal immunoprecipitation, DNA is de-crosslinked, purified and additionally sheared to concentrate all fragments in the size range that is optimal for short tag sequencers like AB/SOLiD (100 - 300 nt) or Illumina/Solexa (400 - 600 nt). C) Overlap between TCF4 ChIP-chip and ChIP-seq data. Peak sets from libraries prepared with the double shearing approach show a larger overlap with the ChIP-chip peak data.

AB/SOLiD platform. To demonstrate this, we split one of the TCF4-immunoprecipitated samples into 2 equal parts and prepared 2 independent libraries that were size-selected for fragments that are optimal for SOLiD emulsion PCR and sequencing, one with and one without second fragmentation. The library produced with the double-fragmentation method resulted in much more unique reads (1.93M unique reads from 3.4M mapped reads) as compared to the sample processed without secondary fragmentation (0.89M unique reads from 3.4M mapped reads), for this purpose when two or more reads have the same starting position on the same strand they were counted as single unique read. Although in both cases it was possible to prepare a sequencing library, the library prepared by the double shearing method is much richer in unique fragments

compared to the library prepared without second shearing, indicating that amounts of chromatin (700pg) were limiting factor in the later case. In the double-fragmentation experiments presented here, we successfully used amounts as low as 0.7 ng of immunoprecipitated DNA for sample preparation, which is much less than the 10 ng that is recommended as the lowest amount in standard ChIP-seq protocols (11). These results indicate that the double-fragmentation ChIP-seq protocol may be well suited for challenging experiments, for example with lower affinity antibodies or in cases where only limited amounts of source material is available. Furthermore, the larger chromatin fragments may be non-random and due to specific biology (6), e.g. packed in large DNA-protein complexes. In support of this, a second rigorous shearing of crosslinked DNA could not fragment all chromatin in small fragments (Fig 1A, sample 5), whereas the same treatment on de-crosslinked DNA results in subfractionation of almost all chromatin in the desired size range (Fig 1A, sample 6).

Comparison of double fragmentation ChIP-seq with ChIP-chip data

Three independent TCF4 ChIP samples obtained from the human colon cancer cell line Ls174T were prepared at different time points as described previously (10). These samples were converted into sequencing libraries using the double fragmentation approach and analyzed by AB/SOLiD sequencing (Table I) (raw sequencing files, alignment files and called peaks have been submitted into GEO database with accession number: GSE18481). In addition, sample 1 was sequenced twice at variable depth. In the case of sample 3, DNaseI was used for second fragmentation. We identified 948 to 10,435 binding regions with the number of identified peaks strongly depending on sequencing depth and false discovery rate settings (Table I).

The results from all libraries showed a strong overlap with each other and with a previously published set of peaks that were obtained by ChIP-on-chip (10) (Fig 1C). Since virtually all of the large peaks identified from libraries prepared by the double fragmentation method do overlap nicely with the ChIP-chip dataset, we can conclude that no major artifacts are introduced by the double fragmentation procedure. Even from the first test sequencing run of Sample 1, where only 1.2 millions of uniquely mapped sequencing reads were generated, 1,127 binding regions were called, out of which 829 (73.5%) mapped to a previously published set of 6,868 high-confidence peaks as determined by ChIP-on-chip (10). However, data from the other experiments illustrate that determination of the complete genome-wide set of TCF4 binding sites is complex and that the number of peaks strongly depends on sequencing depth and enrichment efficiency of the ChIP. Our results also suggest that many weaker TCF4 binding sites exist, which are likely missed by ChIP-on-chip or lower-depth ChIP-seq (12). However, it remains to be demonstrated if these 'weaker' peaks are of biologic relevance.

Versatility and simplification of the procedure

To demonstrate general utility of the described method we processed chromatin immunoprecipitation samples of TATA-binding protein (TBP) and H3 lysine 4 trimethylation histone mark (H3K4me3) in the same way (Table 1). Since virtually all chromatin fragments after the second fragmentation were in the range that is suitable for SOLiD/AB sequencer, the size selection step during library preparation was omitted for these samples. Peaks called from TBP (n = 8,734) and H3K4me3 (n = 15,671) ChIP libraries were predominantly found within 5kb from transcription start sites of protein coding genes (65,0 % in case of TBP and 79,7% in case of H3K4me3) with only a

Table I: TCF4 ChIP libraries overview

Sample	Fragmentation method	Uniquely mapped reads (millions)	Number of peaks (0.1 FDR)	Number of peaks (0.01 FDR)	Peaks (0.1 FDR) overlapping with 6,868 high confidence ChIP-chip peaks (10)	Peaks (0.1FDR) overlapping with 11,912 ChIP-chip peaks (10)
TCF4 #1 (1st run)	sonication	1.2	1,127	948	829	851
TCF4 #1 (2nd run)	sonication	16.9	10,435	6,638	4,466	5,302
TCF4 #2	sonication	4.5	1,998	1,493	1,388	1,417
TCF4 #3	DNaseI	7.4	6,041	4,135	2,935	3,217
TBP	sonication	26.0	8,734	7,303	NA	NA
H3K4me3	sonication	9.2	15,671	15,411	NA	NA

NA – not applicable

smaller subset of peaks mapping elsewhere (mostly close to non-coding RNAs and/or possibly non-annotated transcripts), which is in line with published results (13,14).

Analysis of peak substructure

The double fragmentation ChIP-seq approach was found to provide a very high resolution with clear substructures in the larger peaks (Fig 2B). As the second fragmentation step using sonication could potentially introduce a shearing bias, we used a different method, employing partial DNase I digestion to exclude an artificial origin of the observed substructure. A near identical peak substructure was identified excluding fragmentation bias as the origin of the observed patterns and indicating that the observed substructure has a biological rather than technical origin (Fig 2B), however effects of PCR amplification and/or variation in context-dependent sequencing efficiency cannot be excluded as factors that contribute to the observed patterns.

Indeed, we found that the substructure pattern readily allows for the identification of TCF4 binding positions (Fig 2). Individual binding events of TCF4 as determined by the presence of the canonical binding motif (Fig 2A) were found in the tops of the peaks without the need for any computational deconvolution. In line with this and in contrast to existing protocols (4,5) virtually the same peak pattern is obtained when calling peaks separately from the negative and positive strands (Fig 2B). In addition, the distribution of sequencing reads around the TCF4 binding motifs of peaks with only one binding motif shows that the motif is present in the center of the peak with only a small shift of about 10 bp between the tops of the positive and negative strand peaks (Fig 2C). In contrast, for most commonly used approaches, only the ends of immunoprecipitation products are sequenced, resulting in sequencing coverage peaks flanking the real binding site. Typically, the tops of the + and - strand peaks are separated by up to 200 nt (5).

Biological relevance of binding sites found by double fragmentation method

Cisgenome (15) was used to identify overrepresented consensus motifs within the immunoprecipitated regions. The most common motif discovered (Fig 2A) was nearly identical to the consensus TCF4 binding motif that has been described previously (9,10). From 10,435 binding sites, 55.5% contain at least one TCF4 binding motif. Larger

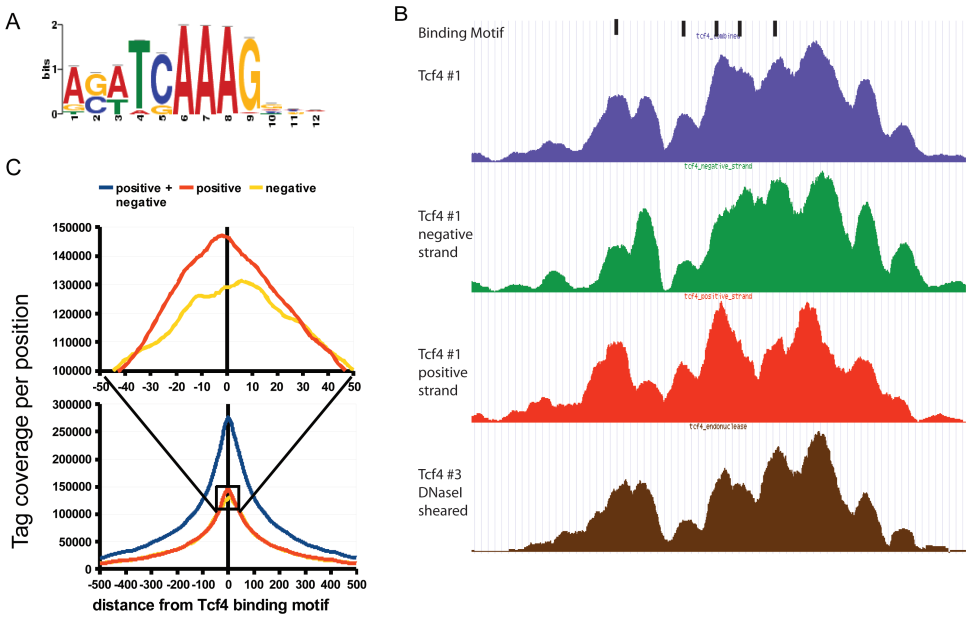


Fig 2: Substructure of binding regions. A) Consensus binding motif sequence logo as identified by Cisgenome from the ChIP-seq data. B) Comparison of Tcf4 binding regions reconstructed from the reads mapped to both strands (blue), the negative strand (green), the positive strand (red), all sub-fragmented using sonication, and reads derived from a library sub-fragmented with DNaseI (brown). The shape and structure of the binding region is highly similar for both strands and does not depend on the fragmentation method used. C) Distribution of sequencing tags from positive and negative strands around the consensus TCF4 binding motif. In contrast to existing protocols without additional fragmentation the maxima of the peaks called separately from the positive and the negative strand overlap with only minor shifting.

peaks, as defined by the number of reads in the peak, more often have TCF4 binding motifs compared to weaker ones (Fig 3A). In addition, 2,369 (22.7%) peaks contain two or more TCF4 binding motifs. This observation is in concordance with accepted models where the presence of several binding motifs close to each other increases the probability that the transcription factor spends more time bound to a particular region (16). In addition, by analyzing binding motifs found in TCF4 binding regions we were able to identify known and potentially novel transcription factors that interact with TCF4 (Figure S1 and Figure S2).

To explore the evolutionary conservation of the observed TCF4 binding sites we used the phastCons (17) scores for each position in the binding regions. Both 200-nt long neighboring flanking sequences and 12-nt long TCF4 binding motifs were more conserved compared to random genomic locations, where the conservation score of the TCF4 binding motif was on average higher than neighboring flanking regions (Fig 3B), indicating selective pressure on these motifs and pointing to functional relevance.

In contrast to ChIP-chip, ChIP-seq has a very high dynamic range and allows for semi-quantitative estimation of DNA-protein interaction strength (as a function of the number of sequencing reads which mapped to the binding region) (5). We divided peaks in bins according to the interaction strength and studied their characteristics. Interestingly,

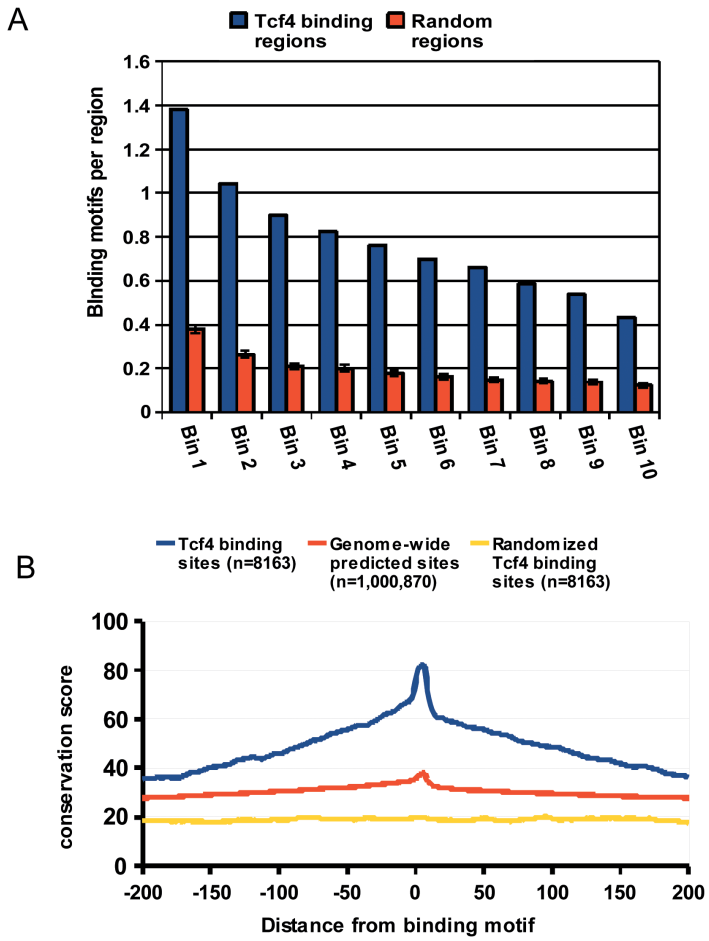


Fig 3: Characterization of TCF4 binding peaks. A) Number of peaks with at least one TCF4 binding motif in relation to the protein-DNA interaction strength. Peaks containing more reads (lower bin numbers) more often harbor a TCF4 binding motif compared to weaker ones. B) Conservation profile of experimentally identified TCF4 binding regions as well as all genomic regions containing the TCF4 consensus binding motif as compared to random regions. Experimentally identified binding regions were found to be more conserved than computationally predicted sites.

weaker peaks were found enriched towards transcription start sites (TSS). As these peaks were also found to more often lack a consensus TCF4 site, it is likely that these regions represent co-immunoprecipitated chromatin that interacts indirectly via DNA-looping with TCF4-containing protein-DNA complexes, similarly as described previously (18) (Figure S3 and Figure S4).

Correlation of sites found by double fragmentation method with differential gene expression

The genome-wide distribution of TCF4 binding sites with respect to TSS of the nearest gene shows a similar distribution as previously reported (10) (Fig 4). A substantial proportion of the peaks is located more than 10 kb from the closest TSS, supporting the model of long range

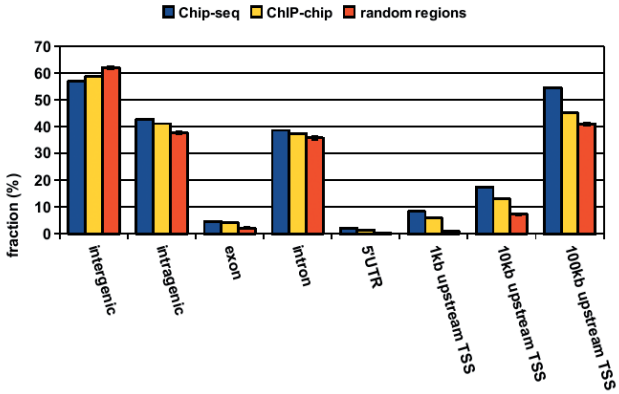


Fig 4: Distribution of TCF4 binding peaks. The distribution of the TCF4 ChIP-seq peaks was analyzed with respect to the closest gene and compared to the distribution of random regions. Genome-wide distribution of ChIP-seq peaks is similar to those identified previously by ChIP-chip with peaks predominantly located far from annotated transcription start sites. This is in line with the established role of TCF4 as a transcriptional enhancer. Error bars for random regions represent standard deviation of 100 randomized datasets.

regulation of gene expression by TCF4. This pattern is also in line with the distribution of other sequence-specific transcription factors such as estrogen receptor (19), STAT1 (2), Foxa2 (3) and p53 (20). To unravel functional TCF4 regulatory gene expression modules connected to Wnt signaling, we used microarray-based gene expression data from modified Ls174T colorectal cancer cell lines (Supplementary Table 1) (microarray data have been submitted to GEO database with accession number: GSE18560). Inducible overexpression of a dominant negative form of TCF4 or siRNA against β-catenin, as described previously (21,22) was used to conditionally turn Wnt signaling off. Genes were ranked according to induced expression changes after abrogation of Wnt signaling and analyzed for the presence of TCF4 immunoprecipitated binding regions. Genes with decreased expression after ablation of the Wnt pathway and thus positively regulated by Wnt and TCF4 had more peak-forming sequencing tags within 100kb from their transcription start site compared to genes with less prominent changes in expression or down-regulated genes (Fig 5A). Although sequencing reads were found to be enriched over the TSS of all 3 groups of genes – up-regulated, down-regulated and non-differentially regulated - enrichment was most prominent for actively up-regulated genes and comparable for negatively and non-regulated genes (Fig 5B).

CONCLUDING REMARKS

Starting from sub-nanogram amounts of immunoprecipitated chromatin we show that the double-fragmentation ChIP-seq protocol allows for the accurate determination of genome-wide binding patterns at high resolution. We show that the method is highly reproducible and versatile and can serve as an alternative for current ChIP-seq protocols especially when limited amounts of immunoprecipitated material are available. Although optimized for the AB/SOLiD platform, shearing settings could be

adjusted for the optimal size range for other platforms (e.g. Illumina/Solexa) as well. Most importantly, the biological relevance of the resulting datasets was firmly demonstrated by in-depth analysis of TCF4 binding regions.

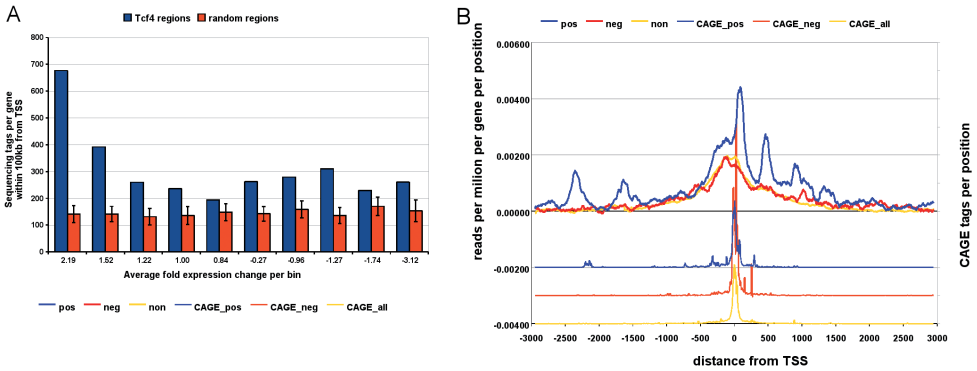


Fig 5: Distribution of binding regions with respect to Wnt regulated genes. A) Gene expression rank analysis. Genes positively regulated by Wnt contain more peak forming sequencing tags within 100kb from their transcription start sites B) Enrichment pattern of sequencing reads around TSS of up-, down-, and non-regulated genes. The observed pattern with additional maxima downstream and upstream of TSS could potentially be explained partially by the presence of alternative or non-annotated TSS, which is actually supported by the presence of CAGE tags in those regions.

MATERIALS AND METHODS

Cells

Ls174T human colon cancer cells carrying an activating point mutation in beta-catenin were used throughout this study. Ls174T-L8 cells carry a doxycyclin-inducible dominant-negative TCF4 transgene; Ls174T-pTER-β-catenin cells carry a doxycyclin-inducible shRNA against β-catenin; they allow for complete and specific blocking of the constitutively active Wnt pathway (21,22).

ChIP

Chromatin immunoprecipitation was performed as described previously (10). In brief, Ls174T cells were cross-linked with 1% formaldehyde for 20 min at room temperature. The reaction was quenched with glycine at a final concentration of 125 mM. The cells were successively washed with phosphate-buffered saline, buffer B (0.25% Triton-X 100, 10 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and buffer C (0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) at 4°C for 10 min each. The cells were then resuspended in ChIP incubation buffer (0.3% sodium dodecyl sulfate [SDS], 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and sheared using a Bioruptor sonicator (Cosmo Bio Co., Ltd.) with six pulses of 30 s each at the maximum setting (Library 1 and 2) or using a Covaris S2 (Covaris) for 8 minutes with the following settings: duty cycle: max, intensity: max, cycles/burst: max (Library 3, TBP, H3K4me3). Both approaches produced similar DNA fragment size range distributions. The sonicated chromatin was incubated for 12 h at 4°C with the appropriate antibody (polyclonal anti-TCF4 antibody, sc-8631; Santa Cruz Biotechnology, Inc.; polyclonal Anti-trimethyl-Histone H3 (Lys4) , 07-473, Millipore; TATA binding protein TBP antibody [1TB18] - ChIP Grade, ab12089, Abcam) at 1 µg of antibody per 10^6 cells with 150 µl of protein G beads (Upstate). The beads were successively washed 2 times with buffer 1 (0.1% SDS, 0.1% deoxycholate, 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 2 (0.1% SDS, 0.1% sodium deoxycholate, 1% Triton-X 100, 0.5 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 3 (0.25 M LiCl, 0.5%

sodium deoxycholate, 0.5% NP-40, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), and two times with buffer 4 (1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) for 5 min each at 4°C. The precipitated chromatin was eluted by incubation of the beads with elution buffer (1% SDS, 0.1 M $NaHCO_3$) at room temperature for 20 min, the eluted fraction was reconstituted to 0.3% SDS with ChIP incubation buffer and the immunoprecipitation repeated with half the amount of antibody. After washing and elution, the immunoprecipitated chromatin was de-cross-linked by incubation at 65°C for 5 h in the presence of 200 mM NaCl, extracted with phenol-chloroform, and ethanol precipitated. Measurement of chromatin concentration was done by high-sensitivity Qubit quantitation (Invitrogen).

Sequencing library preparation

Immunoprecipitated chromatin was dissolved in 100 µl of 10 mM Tris pH 8.0 buffer and sheared for a second time for 6 minutes using the Covaris sonicator (6 x 16 mm AFA fiber Tube, duty cycle: 20%, intensity: 5, cycles/burst: 200, frequency sweeping) to obtain suitable shorter fragments (75-125 bp). To exclude a shearing bias as a possible source of the observe binding site substructure, half of TCF4 Sample 3 was processed with partial digestion using DNaseI as an alternative to shorten the fragments for one control ChIP-seq library and one input library. The second half of the sample was processed without second fragmentation. For DNaseI treatment, chromatin was resuspended in 45 µl of freshly prepared reaction buffer (10mM $MnCl_2$, 0.1mM $CaCl_2$, 10mM Tris, pH 7.5) and digested with 0.5mU of DNaseI for 5 minutes at room temperature. The reaction was stopped by adding EDTA to a final concentration of 50mM. Chromatin was immediately extracted using phenol/chloroform and precipitated. After fragmentation, the fragments were blunt-ended and phosophorylated at the 5' end using the End-it Kit (Epicentre) according to the manufacturer's instructions. Ligation of double stranded adapters compatible with SOLiD sequencing was performed using Quick ligation kit (New England Biolabs) with 750 nM P1 and P2 double-stranded adaptors (Applied Biosystems), 11.7 µl of 2x Quick ligation buffer, 1 µl Quick Ligase in a total volume of 23.4 µl**.** Samples were purified using Ampure beads (Agencourt) and separated on a native 6% polyacrylamide gel. Fragments ranging from 140 to 190 bp were excised; the gel piece containing the selected DNA fragments was shredded and dispersed into 400 µl of Platinum PCR Supermix with 750 nM of each P1 and P2 PCR primer, 2.5 U of Pfu (Stratagene) and 5 U Taq (Bioline). In case of TBP and H3K4me3 samples, the acrylamide gel-based size selection step was skipped and the adapter-ligated library was directly further processed by PCR. Prior to ligation-mediated PCR the sample was incubated at 72˚C for 20 minutes in PCR mix to let the DNA diffuse out of the gel and to perform nick translation on non ligated 3'-ends of DNA fragments. After 17 cycles of amplification the library was purified using Ampure beads and was quality checked on 2100 Bioanalyzer (Agilent) for the absence of possible adapter dimers and heterodimers. Concentration of double-stranded DNA in the final sample was determined by Qubit fluorometer (Invitrogen).

SOLiD sequencing

To achieve clonal amplification of library fragments on the surface of sequencing beads, emulsion PCR (ePCR) was performed according to the manufacturer's instructions (Applied Biosystems). 600 pg of double stranded library DNA was added to 5.6 ml of PCR mix containing 1x PCR Gold Buffer (Applied Biosystems), 3000 U AmpliTaq Gold, 20nM ePCR primer 1, 3 µM of ePCR primer 2, 3.5 mM of each deoxynucleotide, 25mM $MgCl_2$ and 1.6 billion SOLiD sequencing beads (Applied Biosystems). PCR mix was added to SOLiD ePCR Tube containing 9 ml of oil phase and emulsified using ULTRA-TURRAX Tube Drive (IKA). Emulsion was dispensed into 96-well plate and cycled for 60 cycles. After amplification emulsion was broken with butanol, beads were enriched for template positive beads, 3'-end extended and covalently attached onto sequencing slides. Four physically separated samples were deposited on one sequencing slide and sequenced using standard settings on the SOLiD system version 2 to produce 35 nucleotide long reads. TBP and H3K4me3 libraries were sequenced using AB/Solid version 3 to produce 50bp long reads.

Mapping of sequencing data

Sequencing reads were quality trimmed by clipping at 3 consecutive nucleotides with quality score less than 10. Reads shorter than 18 nucleotides were discarded and the remaining reads were mapped against the human reference genome (hg18 assembly, NCBI build 36) using SHRiMP package (23) with default settings, which allows mapping in SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA. For analysis we used only uniquely mapped reads, which were defined as reads having at least two additional mismatches in the second best hit compared to the best hit. TBP and H3K4me3

libraries were mapped against the reference genome (hg18 assembly, NCBI build 36) using the Maq package (24), which allows mapping in SOLiD color space corresponding to dinucleotide encoding of the sequenced DNA with following settings: -n 3, -e 150. Reads with mapping quality zero were discarded.

Peak identification
The Cisgenome software package (15) was used for the identification of binding peaks from the ChIP-seq data. Two-sample analysis mode was used to compare samples with control input data. The parameters for peak discovery were set as follows: window size 100, step size 25, maximum gap 200, active single strand filtering, minimum peak with 225 and minimum reads per window was set to obtain < 0.1 or 0.01 false positive rate (FDR). Peaks called with 0.1 FDR from the second sequencing round of TCF4 Library 1 (n=10,435) were used as a final set for subsequent genome-wide analyses.

Characterization of TCF4 binding regions
The analysis of sequencing reads and TCF4 binding regions with respect to gene structure and distance to closest TSS, *de novo* binding motif discovery, evolutionary conservation analysis and known motif mapping was performed using Cisgenome software packages (15) and custom built Perl scripts. For rank analysis, peaks from the final dataset were sorted by peak rank (calculated by Cisgenome package and dependent on read count as a measure of interaction strength) and divided into 10 bins. Bin 1 contains the 10% largest peaks bin 2 contains the range between 10 and 20%, etc as determined by the number of reads per peak.

Microarray expression analysis of LS174T-L8 and Ls174T-pTER-β-catenin cells
Approximately 10^6 Ls174T-L8 or Ls174T-pTER-β-catenin cells were grown in the presence or absence of doxycycline for 24 hours for Ls174T-L8 or 72 hours for Ls174T-pTER-β-catenin cells. Total RNA was extracted using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. RNA concentration was determined using the NanoDrop ND-1000 and quality was determined using the RNA 6000 Nano assay on the Agilent 2100 Bioanalyzer (Agilent Technologies). For Affymetrix Microarray analysis, fragmentation of RNA, labeling, hybridization to HG-U133 Plus 2.0 microarrays, and scanning were carried out according to the manufacturer's protocol (Affymetrix Inc.). The expression data were normalized with the MAS5.0 algorithm within the GCOS program of Affymetrix. Target intensity was set to 100 ($\alpha1$ = 0.04 and $\alpha2$ = 0.06). Changes in the expression (logfolds and significance of change) for each of the comparisons were determined using the 'Comparison Analysis' from the GCOS program. All data were summarized using custom build Perl scripts. Genes were divided into three categories according to microarray expression data. Wnt upregulated genes - genes significantly down-regulated after suppressing Wnt dependent transcription by both ways in LS174T cells, in all 3 biological replicates (n = 647), ii) Wnt down-regulated genes - genes significantly up-regulated after suppressing Wnt dependent transcription by two ways in all 3 biological replicates (n = 576) and iii) non-Wnt-regulated genes - genes consistently without significant expression change after suppressing Wnt-dependent transcription (n = 3936) (Supplementary Table 1).

AUTHORS' CONTRIBUTION

MM, PH, HC and EC designed and coordinated the experiments. PH, MM, JS and MvdW performed all the experiments except microarrays. EdB performed SOLiD sequencing. JK, RV performed microarray experiments and analysis. MM, PH, VG and EC analyzed data. MM and EC wrote the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was financially supported by the Cancer Genomics Center (Netherlands Genomics Initiative). P.H. was supported by a Human Frontier Science Program Organization long-term fellowship.

REFERENCES

1. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science, 316, 1497-1502.
2. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A. et al. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods, 4, 651-657.
3. Wederell, E.D., Bilenky, M., Cullum, R., Thiessen, N., Dagpinar, M., Delaney, A., Varhol, R., Zhao, Y., Zeng, T., Bernier, B. et al. (2008) Global analysis of in vivo Foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res, 36, 4549-4564.
4. Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods, 5, 829-834.
5. Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res, 36, 5221-5231.
6. Teytelman, L., Ozaydin, B., Zill, O., Lefrancois, P., Snyder, M., Rine, J. and Eisen, M.B. (2009) Impact of chromatin structures on DNA processing for genomic analyses. PLoS ONE, 4, e6700.
7. Molenaar, M., van de Wetering, M., Oosterwegel, M., Peterson-Maduro, J., Godsave, S., Korinek, V., Roose, J., Destree, O. and Clevers, H. (1996) XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos. Cell, 86, 391-399.
8. Behrens, J., von Kries, J.P., Kuhl, M., Bruhn, L., Wedlich, D., Grosschedl, R. and Birchmeier, W. (1996) Functional interaction of beta-catenin with the transcription factor LEF-1. Nature, 382, 638-642.
9. van de Wetering, M., Cavallo, R., Dooijes, D., van Beest, M., van Es, J., Loureiro, J., Ypma, A., Hursh, D., Jones, T., Bejsovec, A. et al. (1997) Armadillo coactivates transcription driven by the product of the Drosophila segment polarity gene dTCF. Cell, 88, 789-799.
10. Hatzis, P., van der Flier, L.G., van Driel, M.A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W. et al. (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cell Biol, 28, 2732-2744.
11. Park, P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 10, 669-680.
12. Kharchenko, P.V., Tolstorukov, M.Y. and Park, P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol, 26, 1351-1359.
13. Santos-Rosa, H., Schneider, R., Bannister, A.J., Sherriff, J., Bernstein, B.E., Emre, N.C., Schreiber, S.L., Mellor, J. and Kouzarides, T. (2002) Active genes are tri-methylated at K4 of histone H3. Nature, 419, 407-411.
14. van Werven, F.J., van Bakel, H., van Teeffelen, H.A., Altelaar, A.F., Koerkamp, M.G., Heck, A.J., Holstege, F.C. and Timmers, H.T. (2008) Cooperative action of NC2 and Mot1p to regulate TATA-binding protein function across the genome. Genes Dev, 22, 2359-2369.
15. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M. and Wong, W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 26, 1293-1300.
16. Halford, S.E. and Marko, J.F. (2004) How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res, 32, 3040-3052.
17. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S. et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res, 15, 1034-1050.
18. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462, 58-64.
19. Carroll, J.S., Meyer, C.A., Song, J., Li, W., Geistlinger, T.R., Eeckhoute, J., Brodsky, A.S., Keeton, E.K., Fertuck, K.C., Hall, G.F. et al. (2006) Genome-wide analysis of estrogen receptor binding sites. Nat Genet, 38, 1289-1297.
20. Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z. et al. (2006) A global map of p53 transcription-factor binding sites in the human genome. Cell, 124, 207-219.
21. van de Wetering, M., Oving, I., Muncan, V., Pon Fong, M.T., Brantjes, H., van Leenen, D., Holstege, F.C., Brummelkamp, T.R., Agami, R. and Clevers, H. (2003) Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. EMBO Rep, 4, 609-615.
22. van de Wetering, M., Sancho, E., Verweij, C., de Lau, W., Oving, I., Hurlstone, A., van der Horn, K., Batlle, E., Coudreuse, D., Haramis, A.P. et al. (2002) The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. Cell, 111, 241-250.
23. Rumble, S.M., Lacroute, P., Dalca, A.V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol, 5, e1000386.
24. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling

variants using mapping quality scores. Genome Res, 18, 1851-1858.
25. Love, J.J., Li, X., Case, D.A., Giese, K., Grosschedl, R. and Wright, P.E. (1995) Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature, 376, 791-795.
26. Ross, E.D., Hardwidge, P.R. and Maher, L.J., 3rd. (2001) HMG proteins and DNA flexibility in transcription activation. Mol Cell Biol, 21, 6598-6605.
27. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. and Noble, W.S. (2007) Quantifying similarity between motifs. Genome Biol, 8, R24.
28. Rossi, A., Mukerjee, R., Ferrante, P., Khalili, K., Amini, S. and Sawaya, B.E. (2006) Human immunodeficiency virus type 1 Tat prevents dephosphorylation of Sp1 by TCF-4 in astrocytes. J Gen Virol, 87, 1613-1623.
29. Nateri, A.S., Spencer-Dene, B. and Behrens, A. (2005) Interaction of phosphorylated c-Jun with TCF4 regulates intestinal cancer development. Nature, 437, 281-285.
30. Jansson, E.A., Are, A., Greicius, G., Kuo, I.C., Kelly, D., Arulampalam, V. and Pettersson, S. (2005) The Wnt/beta-catenin signaling pathway targets PPARgamma activity in colon cancer cells. Proc Natl Acad Sci U S A, 102, 1460-1465.
31. Yang, M.Q. and Elnitski, L.L. (2008) Diversity of core promoter elements comprising human bidirectional promoters. BMC Genomics, 9 Suppl 2, S3.
32. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. et al. (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science, 321, 956-960.

SUPPLEMENTARY DATA FILES

Motif Logo	Enrichment ratio
	5.117172
	2.52351
	2.366461
	1.998872
	1.884736
	1.850934
	1.831706
	1.512527
	1.264676
	1.252181
	1.239546
	1.20281
	1.198533
	1.157542
	1.156647

Fig S1: Fifteen most overrepresented motifs found in Tcf4 binding regions. Tcf4 binding motif shows highest enrichment in binding sites compared to the other motifs.

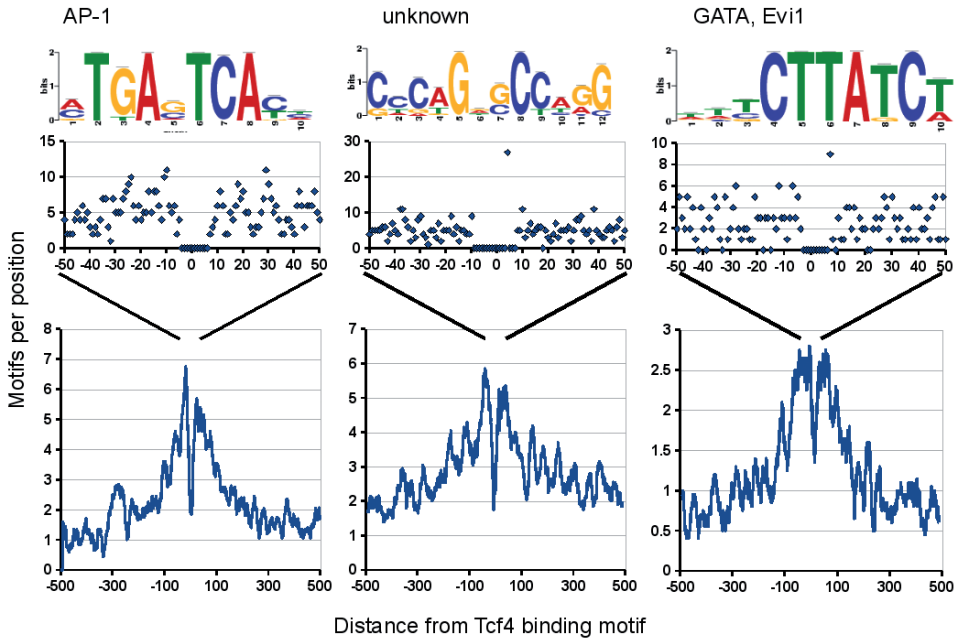


Figure S2: Examples of *de novo* called motifs enriched in peaks. Clustering of known protein binding motifs, such as AP-1 and GATA, close to the consensus Tcf4 binding motif supports the existence of possible functional interactions of these proteins with Tcf4. The middle motif is not associated with any known transcription factor and may represent the consensus binding site of a novel Tcf4-interacting partner. To identify DNA-binding proteins that could interact with Tcf4, we performed *de novo* identification of enriched sequence motifs in the immunoprecipitated regions and focused on the 15 most over-represented motifs. The majority of the *de novo* called motifs showed clear enrichment close to known Tcf4 binding motifs in Tcf4 binding regions, often within 10 bp, suggesting a potential direct interaction with Tcf4 as a heterodimer. Alternatively, since Tcf4 harbors a HMG-box domain, which can increase flexibility of DNA (25,26) Tcf4 binding may primarily function to facilitate binding of other transcription factors in its neighborhood. By comparing weighted matrices of *de novo* called motifs with binding motifs from the TRANSFAC database using the TOMTOM (27) web-based motif comparison tool, we identified binding motifs for several of the known Tcf4 interacting partners including SP1 (28), AP1 (29) and PPARγ (30). In addition, we discovered novel binding motifs of proteins that were not associated with Tcf4 before and could thus represent binding sites for yet unidentified or uncharacterized Tcf4 interacting partners.

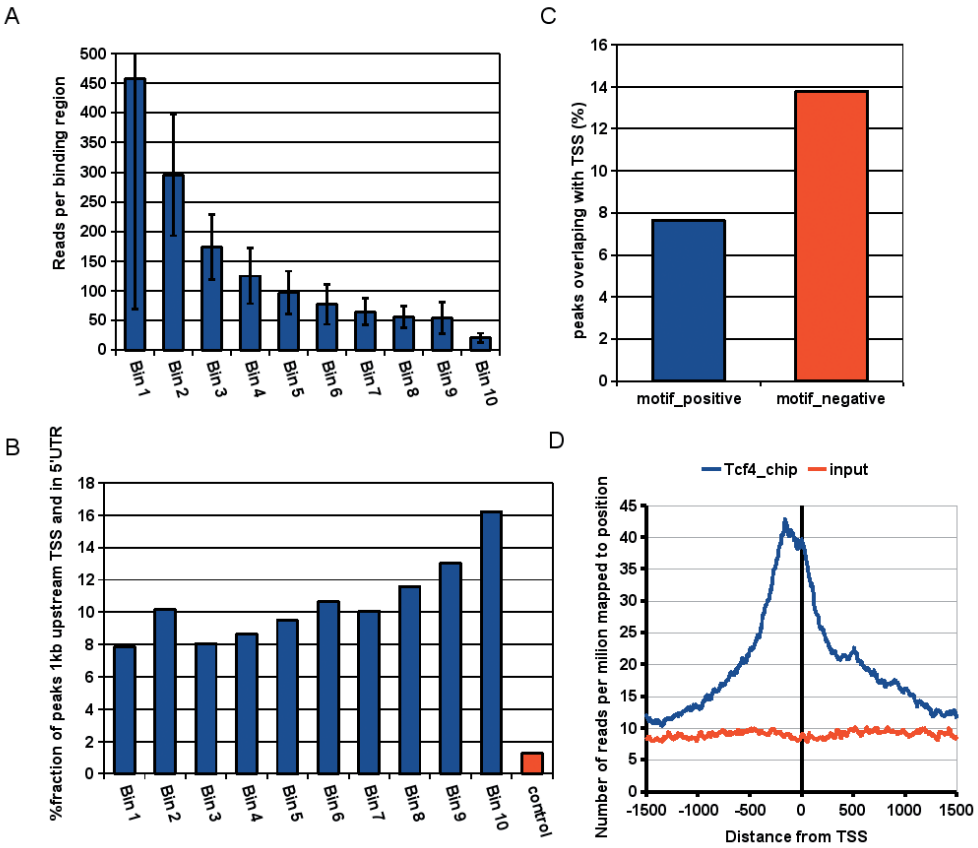

A
Reads per binding region
500
450
400
350
300
250
200
150
100
50
0
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
Bin 7
Bin 8
Bin 9
Bin 10
C
peaks overlaping with TSS (%)
16
14
12
10
8
6
4
2
0
motif_positive
motif_negative
B
%fraction of peaks 1kb upstream TSS and in 5'UTR
18
16
14
12
10
8
6
4
2
0
Bin 1
Bin 2
Bin 3
Bin 4
Bin 5
Bin 6
Bin 7
Bin 8
Bin 9
Bin 10
control
D
Tcf4_chip
input
Number of reads per milion mapped to position
45
40
35
30
25
20
15
10
5
0
-1500
-1000
-500
0
500
1000
1500
Distance from TSS

Figure S3: Indirect immunoprecipitation of interacting regions. A) The average number of sequencing reads mapped to the binding region with respect to protein-DNA interaction strength. Approximately 50% of the sequencing reads that map to binding regions map to 10% of the strongest ones (determined by the number of reads mapping to the binding region). B) Distribution of binding regions in relation to the closest protein-coding gene with respect to protein-DNA interaction strength. Weaker binding regions tend to overlap more often with transcription start sites (TSS). C) Percentage of binding regions mapped to TSS with respect to the presence of the Tcf4 consensus binding motif. D) Enrichment pattern of sequencing tags around TSS compared to the pattern of control sequencing tags from input DNA. In contrast to ChIP-chip, ChIP-seq has a very high dynamic range and allows for semi-quantitative estimation of DNA-protein interaction strength (as a function of the number of sequencing reads which mapped to the binding region) (5). Furthermore, lowly represented weak binding regions are more easily detected as the ChIP-seq output has a 'digital' nature (read counts) without the need for complicated background corrections. Extensive characterization of the distribution of Tcf4 binding regions revealed an unequal size distribution among peaks. The top 10 % of the strongest binding regions (as determined by the number of reads contributing to the peak) contain about ~50 % of all reads that map to peaks, suggesting that the majority of Tcf4 is bound to only a limited number of binding sites. The biological relevance of the remaining weaker binding peaks is difficult to interpret as i) the interaction itself could be weaker, potentially resulting from stabilized binding to random or suboptimal sites, ii) the interaction may only be present in a limited number of cells from the whole population, iii) Tcf4 can be less accessible by antibody on those regions or iv) they could represent biologically irrelevant noise. However, these weak binding sites could also be biologically relevant as they may represent other types of interactions, such as co-immunoprecipitated chromatin that interacts indirectly via DNA-looping with Tcf4-containing protein-DNA (19). To address this, we sorted the peaks according to the number of mapped sequencing reads and divided them into 10 bins. Sites covered by fewer reads (in the higher number bins) tend to occur more often close to transcription start site (TSS) positions than sites with more reads. As Tcf4 is considered to be primarily an enhancer-binding factor, the strong peaks are likely to reflect the real Tcf4 binding sites. Since indirect immunoprecipitation is expected to be less efficient, the frequency of reads representing such interactions can be expected to be relatively low. The observed increased number of weaker peaks overlapping with TSS supports the hypothesis that a substantial part of the identified Tcf4 binding regions actually originate from indirect immunoprecipitation of proteins interacting with Tcf4, similar to what was shown previously for the estrogen receptor (19). In support of this, a large proportion of weak binding regions (44.5%) was found to be devoid of consensus Tcf4 binding motifs. In general, Tcf4 motif-negative regions do overlap more often with TSS regions, compared to peaks with a clear Tcf4 binding motif. Other evidence that supports indirect immunoprecipitation of promoters of genes that are regulated by Tcf4 comes from the typical asymmetric overall substructure of these peaks. Interestingly, this pattern is remarkably similar to patterns observed for RNA polymerase II ChIP-seq in promoter regions of active genes (31,32), strongly suggesting functional physical interactions between Tcf4 enhancer complexes and the transcription machinery.

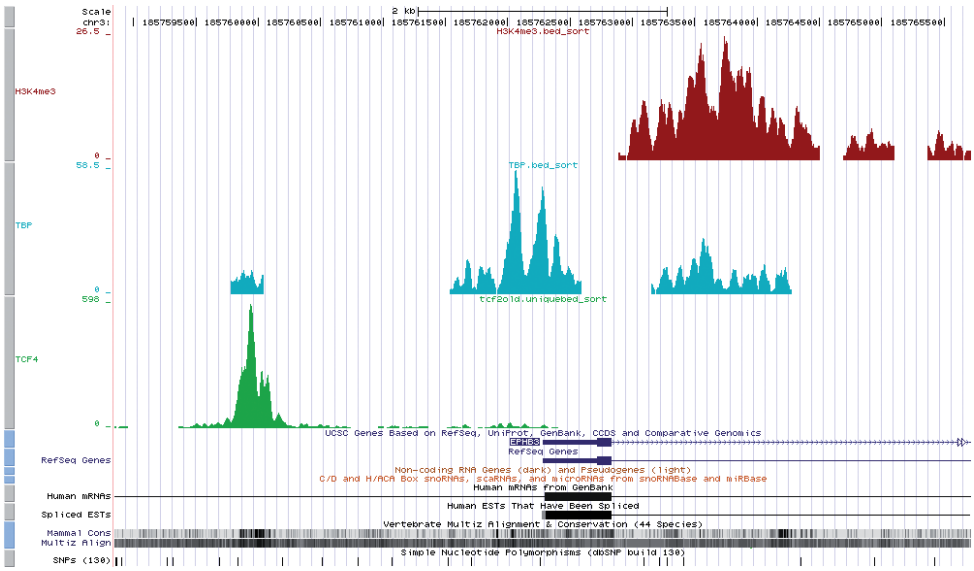


Figure S4: Example of peaks called from Tcf4, TBP and H3K4me3 dataset. Small but significant TBP peak is present on position of Tcf4 peak located upstream of expected transcription start site of EPHB3 gene. Additionally, small Tcf4 peak is present directly at transcription start site of EPHB3 gene on position where main TBP peak is present. This could suggest that these small peaks originate in indirect immunoprecipitation of TBP bound region interacting with Tcf4 and vice versa.

7

INTEGRATED GENOME-WIDE ANALYSIS OF TRANSCRIPTION FACTOR OCCUPANCY, RNA POLYMERASE II BINDING AND STEADY STATE RNA LEVELS IDENTIFIES DIFFERENTIALLY REGULATED FUNCTIONAL GENE CLASSES

Mokry M*, Hatzis P*, Schuijers J*, Lansu N, Ruzius FP, Clevers H, Cuppen E: **Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady state RNA levels identifies differentially regulated functional gene classes**. *Nucleic Acids Research* 2011, **doi: 10.1093/nar/gkr720**.
*equal contribution

ABSTRACT

Routine methods for assaying steady-state mRNA levels like RNA-seq and micro-arrays are commonly used as readouts to study the role of transcription factors (TFs) in gene expression regulation. However, cellular RNA levels do not solely depend on activity of transcription factors and subsequent transcription by RNA polymerase II (Pol II), but are also affected by RNA turnover rate. Here we demonstrate that integrated analysis of genome-wide TF occupancy, Pol II binding and steady state RNA levels provides important insights in gene regulatory mechanisms. Pol II occupancy, as detected by Pol II ChIP-seq, was found to correlate better with TF occupancy compared to steady state RNA levels and is thus a more precise readout for the primary transcriptional mechanisms that are triggered by signal transduction. Furthermore, analysis of differential Pol II occupancy and RNA-seq levels identified genes with high Pol II occupancy and relatively low RNA levels and vice-versa. These categories were found to be strongly enriched for genes from different functional classes. Our results demonstrate a complementary value in Pol II chip-seq and RNA-seq approaches for better understanding of gene expression regulation.

INTRODUCTION

Extrapolation of transcriptional changes in response to signal transduction to molecular mechanisms and regulatory networks remains a major challenge. Over the past years, the increase in microarray densities and quality and the development of massively parallel sequencing of transcriptomes (RNA-seq) allowed affordable genome-wide readout of gene expression over multiple samples with high accuracy and reproducibility. These techniques have proven to be very useful for studying and understanding regulatory networks controlled by different transcriptional programs. However, the above mentioned techniques measure accumulated levels of RNA that do not necessarily fully reflect transcriptional status of a gene under the given conditions, because steady-state RNA levels are the result of a tightly regulated balance between RNA synthesis and degradation rate (1) with certain classes of genes having different rates of mRNA degradation (2-4).

Other more direct alternatives for measuring transcription are based on "nuclear run-on" (5), dynamic transcriptome analysis (6) or sequencing of nascent transcripts from immunoprecipitated RNA polymerase II (7). However, these techniques require a relatively laborious experimental setup (e.g. metabolic RNA labeling with 4-thiouridine in living cells) or rely on expression of tagged versions of proteins, making it difficult to use in organism-based studies. Other methods like GRO-seq (8) require the isolation of viable nuclei, which may affect the transcriptional programs in response to stimuli that would normally not occur in intact cells. In addition, these techniques are not compatible with frozen or formalin fixed paraffin embedded (FFPE) archived material.

To address these issues and to obtain a more direct readout of gene expression in a simple and unbiased way, we applied RNA Polymerase II (Pol II) ChIP-seq (9) as a versatile complementary approach to RNA-seq and microarrays. We demonstrate that this approach, which is based on commonly used ChIP-seq and RNA-seq protocols, provides detailed insight in transcriptional processes. While we demonstrated utility in cultured cells, ChIP-seq and RNA-seq have been shown to work on frozen or FFPE archived material as well as on very small numbers of cells (10-14), providing unique opportunities for studying transcriptional processes where other methods that more directly measure transcriptional rates have limitations or are even impossible.

Applied to the colon cancer model system used here, we were able to identify subclasses of genes that appear regulated by different mechanisms upon WNT-induced signal transduction. These findings illustrate the complementarity of techniques in further dissecting gene regulatory networks.

MATERIALS AND METHODS

Cells

We used Ls174T human colon cancer cells carrying an activating point mutation in beta-catenin and Ls174T-pTER-β-catenin cell line carrying a doxycyclin-inducible shRNA against β-catenin (15). Cells were grown in the presence or absence of doxycycline (1μg/ml) for 72 hours.

Microarray analyses

We used publicly available data of doxycyclin treated and untreated Ls174T-pTER-β-catenin performed on HG-U133 Plus 2.0 microarrays (Affymetrix) (9). CEL files (GEO accession number: GSE18560) were processed by RMA method (16) using rma() function from Bioconductor affy library with standard settings. Gene expression is defined as direct value from RMA analysis. Expressed gene is gene with expression higher than 16. Differentially transcribed genes were set as genes with at least 2-fold intensity change in all 3 biological replicates with normalized intensity higher than 16 in all 6 samples.

RNA-seq

Total RNA was extracted using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. To deplete for non-informative ribosomal RNA 5µg of total RNA were purified using Ribominus kit (Invitrogen) according to manufacturer's instructions. Ribosome depleted RNA was resuspended in 50µl of DEPC treated water and fragmented for 60s using the Covaris sonicator (6 x 16 mm AFA fiber Tube, duty cycle: 10%, intensity: 5, cycles/burst: 200, frequency sweeping). Sheared RNA fragments were phosphorylated using 30U of Polynucleotide Kinase (Promega) with 0.5mM ATP for 30 minutes at 37ºC. Phosphorylated RNA was purified using TRIzol according manufacturer's instruction and resuspended in 1.5µl of DEPC treated water with 1µl of Adaptor mix A and 1.5µl of Hybridization solution from SOLiD Small RNA Expression Kit (SREK) (Ambion). The mixture was incubated at 65ºC in a thermocycler for 5 minutes and quickly cooled on ice. RNA with hybridized adaptors was mixed with 5µl of SREK ligation buffer and 1 µl of SREK ligation enzyme and incubated at room temperature for 4 hours. Ligated sample was mixed with 10 µl of denaturing buffer (90% formamide, bromphenol blue, crysol red) and size selected on 10% denaturing urea gel for the appropriate size fraction (150-300bp). The piece of gel containing selected fragments was shredded and RNA was eluted in 300 µl 300mM NaCl with gentle agitation for 4 hours at RT. The eluate was separated from gel debris using SPIN-X centrifuge tube filters (Costar), the RNA was precipitated by isopropanol and resuspended in 5µl of DEPC treated water. Size selected RNA was mixed with 2 µl of reverse transcription (RT) buffer, 1.5 µl of dNTPs (10mMeach), 0.5 µl of barcode RT primer (10µM), incubated in a thermocycler at 72ºC for 4 minutes, 62ºC for 2 minutes and then put on ice. The sample with hybridized barcode RT primer was mixed with MMLV- RT enzyme (Promega) and incubated at 37ºC for 30 minutes. The library was amplified by ligation mediated PCR (LM-PCR) by adding 2 µl of RT mix from the previous reaction into 100 µl of PCR mix from the SREK kit (Ambion) with P1 and P2 primer compatible with SOLiD/AB sequencing and cycled in a thermocycler with the following program: 95ºC for 2min; 15 cycles of 95ºC for 30s, 62ºC for 30s, 72ºC for 30s; 72ºC for 7min. The library was size selected on a 2% agarose gel for 150-400 bp long fragments and sequenced on SOLiD/AB sequencer in a multiplexed way to produce 50bp long reads.

Sequencing reads were mapped against the reference genome (hg18 assembly, NCBI build 36) with Maq package (17), with following settings: -c -n 3, -e 170. Reads with mapping quality zero were discarded. To set gene expression from RNA-seq data we counted the number of the sequencing tags aligned to exons and UTRs with the same strand orientation as the annotated transcripts. To avoid transcripts with zero mapped tags to interfere with logarithmic transformation of read counts, one read per every 10 millions sequencing tags was added to each transcript. Raw read counts were normalized to the length of mature transcript RNA and sequencing depth. All six samples (three biological replicates of two experimental conditions) were quantile normalized using normalizeQuantiles() function (17) from limma (18) and are presented as normalized read counts per transcript per 10kb of transcript per million sequencing tags (NRP10KM). Expressed gene is gene with expression higher than 4 NRP10KM. Differentially transcribed genes were set as genes with at least 2-fold NRP10KM change in all 3 biological replicates with absolute NRP10KM higher than 4 in all 6 samples.

Pol II ChIP-seq

Approximately 30 $x.10^6$ Ls174T-pTER-β-catenin cells grown 72 hours in the presence or absence of doxycycline (1µg/ml) were used for ChIP-seq procedure. Chromatin immunoprecipitation (9) (9,19). In brief: cells were crosslinked with 1% formaldehyde for 20 min at room temperature. The reaction was quenched with glycine and the cells were successively washed with phosphate-buffered saline, buffer B (0.25% Triton-X 100, 10 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and buffer C (0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]). The cells were then resuspended in ChIP incubation buffer (0.3% sodium dodecyl sulfate [SDS], 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and sheared using Covaris S2 (Covaris) for 8 minutes with the following settings: duty cycle: max, intensity: max, cycles/burst: max, mode: Power Tracking. The sonicated chromatin was diluted to 0.15 SDS, incubated for 12 h at 4°C with 10 µl of the anti RBP1 (PB-7G5) antibody (Euromedex) per IP with 100 µl of protein G beads (Upstate). The beads were successively washed 2 times with buffer 1 (0.1% SDS, 0.1% deoxycholate, 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 2 (0.1% SDS, 0.1% sodium deoxycholate, 1% Triton-X 100, 0.5 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 3 (0.25 M LiCl, 0.5% sodium deoxycholate, 0.5% NP-40,

1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), and two times with buffer 4 (1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) for 5 min each at 4°C. Chromatin was eluted by incubation of the beads with elution buffer (1% SDS, 0.1 M $NaHCO_3$), the eluted fraction was reconstituted to 0.15% SDS with ChIP incubation buffer and the immunoprecipitation repeated for second time with half the amount of antibody. After washing and elution, the immunoprecipitated chromatin was de-cross-linked by incubation at 65°C for 5 h in the presence of 200 mM NaCl, extracted with phenol-chloroform, and ethanol precipitated. Immunoprecipitated chromatin was additionally sheared, end-repaired, sequencing adaptors were ligated and the library was amplified by ligation mediated PCR (LMPCR). After LMPCR, the library was purified and checked for the proper size range and for the absence of adaptor dimers on a 2% agarose gel and sequenced on SOLiD/AB sequencer to produce 50 basepair long reads. Sequencing reads were mapped against the reference genome (hg18 assembly, NCBI build 36) using the Maq package (17), with following settings: -c -n 3, -e 170. Reads with mapping quality zero were discarded. To set gene expression from Pol II ChIP-seq data, we counted the number of the sequencing tags aligned to annotated transcript coordinates. To avoid transcripts with zero mapped tags to interfere with logarithmic transformation of read counts, one read per every 10 millions sequencing tags was added to each transcript. Raw read counts were normalized to the transcript length (from TSS to TES) and sequencing depth. All six samples (three biological replicates of two experimental conditions) were quantile normalized using normalizeQuantiles() function (17) from limma (18) and are presented as normalized read counts per transcript per 100kb of transcript per million sequencing tags (NRP100KM). Expressed gene is gene with expression higher than 16 NRP100KM. Differentially transcribed genes were set as genes with at least 2-fold NRP100KM change in all 3 biological replicates with absolute NRP100KM higher than 16 in all 6 samples.

ChIP-seq of transcription factors

We used publically available datasets for TCF4 and TBP (GEO database accession number: GSE18481) produced in our lab (9) in Ls174T cells. Chromatin immunoprecipitation, library preparation and sequencing of other transcription factors (Supplementary Material I) in Ls174T cells were performed as for the Pol II ChIP-seq with modifications: Approximately 50.10^6 cells were used per IP. For β-catenin ChIP-seq, cells were crosslinked for 40 minutes using ethylene glycol-bis(succinimidylsuccinate) (Thermo scientific) at 12.5mM final concentration, with addition of formaldehyde (1% final concentration) after 20 minutes of incubation. Cisgenome software package (20) was used for the identification of binding peaks from the ChIP-seq data.

Data files from ENCODE project

ChIP-seq data of 21 transcription factors and Pol II (Supplementary Material II) were produced in by the Myers Lab at the HudsonAlpha Institute for Biotechnology and downloaded from http://genome.ucsc.edu/ENCODE/downloads.html website (21). RNA-seq data (Supplementary Material II) were produced by the Wold Group at the California Institute of Technology and downloaded from http://genome.ucsc.edu/ENCODE/downloads.html website (21).

Calculation of TTAS

TTAS represent the relative likelihood of transcript *j* to be regulated by transcription factor *i*. To calculate $TTAS_{ij}$ we first calculated raw score (t_{kj}) which reflects likelihood of transcript *j* being regulated by transcription factor *i* via binding site *k*. We adapted published method used previously to calculate TF-Gene association score (22) to calculate raw score for each transcript – binding site t_{kj} separately. We first calculated the distribution of binding sites *k* of transcription factor *i* with closest transcriptional start sites *g* and created histograms *Hist* of distances *l*(k,g) consisting of 18 location bins separated by {- 200k bp, -100k bp, -50k bp, -20k bp, -10k bp, -5k bp, -2k bp, -1k bp, 0 bp, 1k bp, 2k bp, 5k bp, 10k bp, 20k bp, 50k bp, 100k bp, 200k bp}. Next, we randomized positions of TF binding sites *k* and calculated distribution of random sites with respect to transcriptional start sites in the same way as distribution of real sites. Let *m* be the index of bin corresponding to *l(k,g)* the raw score *t* (Supplementary Fig 1) for binding site *k* and transcript *j* is calculated by:

$$t_{kj} = \begin{cases} 0, & if\ Hist_{real}(m) \le Hist_{rand}(m) \\ (Hist_{real}(m) - Hist_{rand}(m))/Hist_{real}(m), & if\ Hist_{real}(m) > Hist_{rand}(m) \end{cases}$$

Next, since peak intensity was shown to be factor that helps in predicting expression (23) we included this principle in our final TTAS. Final TTAS of transcript-transcription factor pair is then calculated as log2 value of the sum of all above zero raw scores multiplied by number of tags mapped to peak coordinates:

$$TTAS_{ij} = log_2 \sum_k t_k n_k$$

where t_k is the raw score of the kth binding site of transcription factor i in the vicinity of transcript j and n_k is number of sequencing tags mapped to the kth peak coordinates. By this approach we take into an account TF-DNA interaction strength (reflected by number of reads in peak), distance from TSS, number of individual peaks in vicinity of TSS and distribution of binding sites. To extract principal components from TTAS we used R language princomp() (24) command.

Classification trees

We used individual principal components as inputs to train the classification tree to distinguish between expressed and non-expressed genes. Training was performed by the CART algorithm (25) using the rpart() command from R package rpart (26) (http://CRAN.R-project.org/package=rpart). To avoid over-fitting of the data, we prune back the tree to select the tree size with complexity parameter that associates with the smallest 10-cross validation error.

Calculation of TAS

TAS of represents normalized enrichment of Pol II tags mapped within 300 bp from TSS to tags mapped to the transcript body (excluding 3'UTR) of transcript i:

$$TAS_i = \frac{\left(\frac{t_i}{600}\right)}{\left(\frac{b_i}{l_i}\right)}$$

Where t_i represents number of tags mapped within 300 bp from TSS of transcript i, b_i represents number of tags mapped to transcript i excluding first 300 bp and 3'UTR and l_i represent length of transcript i excluding first 300 bp and 3'UTR. Transcripts with change in TAS after doxycycline induction are transcripts with at least 0.6 log2 fold change of TAS in all 3 biological replicates. We assayed only transcripts with at least 50 aligned reads combined from all 3 replicates and with gene expression defined by POL II higher than 16 in all 3 replicates.

RESULTS AND DISCUSSION

Model system

For studying molecular mechanisms downstream of a defined signal transduction pathway, we used the human colon cancer cell line Ls174T-pTER-β-catenin (15). These cells carry a doxycyclin-inducible shRNA targeting β-catenin, which allows for complete and specific blocking of the - in these cells constitutively active - Wnt pathway, causing major changes in the expression of many genes; including direct Wnt target genes (15). We performed microarray expression analysis, RNA-seq and Pol II ChIP-seq (Supplementary Table) in triplicate on untreated cells and cells 72 hours after doxycyclin induction to determine which measurement correlates best with the activity of transcription factors that are associated with the Wnt signal transduction pathway or the core transcriptional machinery.

Pol II ChIP-seq

We used the well-characterized monoclonal antibody 1PB-7G5, raised against the heptad repeat CTD-containing peptide of the RPB1 subunit, which recognizes both hyper- and hypo-phosphorylated forms of Pol II with high affinity (27). Although the use of antibodies specific for hyper-phosphorylated forms of actively transcribing polymerase could potentially further improve on the specificity of the method, the antibody used here has the advantage of a very high affinity, which results in high yields of chromatin and a very robust and reproducible procedure for routine operation what is crucial especially when source of material is limited. To calculate the Pol II DNA occupancy based on Pol II ChIP-seq data, we counted all sequencing tags aligned between the annotated transcription start sites and the transcription termination sites.

Read counts were normalized for transcript gene length and sequencing depth and quantile normalization was performed across all 6 samples (three biological replicates of two conditions).

Correlation of Pol II ChIP-seq with RNA levels

To determine the RNA levels from the RNA-seq data we used a comparable approach as for Pol II ChIP-seq. However, we only counted sequencing tags aligned to exons and UTRs with the same strand orientation as the annotated transcript and normalized the read count to the length of mature transcript mRNA. Pol II occupancy and RNA levels assayed by RNA-seq and microarrays of 21,854 RefSeq genes showed good correlation among biological replicates within the same method (Fig 1, Supplementary Fig 2) although a significant and reproducible lower correlation was observed between different methods as compared to triplicates. These results show that the reproducibility of Pol II ChIP-seq is at least comparable to other methods for measuring gene expression, but does not reveal which method best reflects the transcriptional processes. Interestingly, the Pol II Chip-seq results cluster separately from the microarrays and RNA-seq results (Fig 1A). This indicates that Pol II occupancy and RNA-seq measurements have a different information content that could be exploited to obtain mechanistic and biological insights.

Identification of differentially expressed genes

Next, to explore the ability of Pol II ChIP-seq to identify differentially expressed genes we compared changes in Pol II occupancy in Wnt-active and non-active cells with differences of RNA levels assayed by RNA-seq and microarrays. Based on Pol II Chip-seq, we identified 345 differentially expressed genes (157 up and 188 down-regulated) with at least 2-fold change in Pol II occupancy in all 3 replicates. These genes show overlap with genes that are differentially expressed (> 2-fold) as detected by microarrays (110 out of 505) and RNA-seq (128 out of 816). Overlap for down-regulated genes, which represent possible Wnt target genes, is better than for up-regulated genes (Fig 1B). A total of 77 genes were identified by all three methods, but a set of 152 genes commonly identified by RNA-seq and microarrays, was not detected as differentially expressed by Pol II Chip-seq. These discrepancies could reflect different mechanisms of RNA level regulation. For example, induced changes in RNA stability will results in changed RNA levels as measured by RNA-seq or microarrays, but with smaller or no differences in Pol II occupancy (Supplementary Figure 3). However, these features do allow us to distinguish the regulatory mechanisms that act through production of RNA from those regulating RNA stability. Interestingly, gene ontology analysis revealed different gene classes in down-regulated genes (Wnt-target genes) identified by POL II as compared to methods measuring changes in RNA levels, indicating different types of regulation for different functional gene categories (Fig 1C, Supplementary Material III).

Correlation of Pol II ChIP-seq with the presence of transcriptional regulators

To determine which method correlates best with transcription factor mediated regulatory processes, we compared the Pol II DNA occupancy and RNA levels with transcription factor (TF) presence. We used ChIP-seq profiles of TCF4 and TBP (9), and generated additional ChIP-seq profiles of 3 other transcription factors (β-catenin, E2A and c-Myc) in LS174 cells (Supplementary Material I). These transcription factors are part of, or are related to the Wnt-pathway (28-30), which is constitutively active in LS174 cells, or are part of the basic transcriptional machinery. To link binding sites to individual genes, we determined the likelihood of a gene being

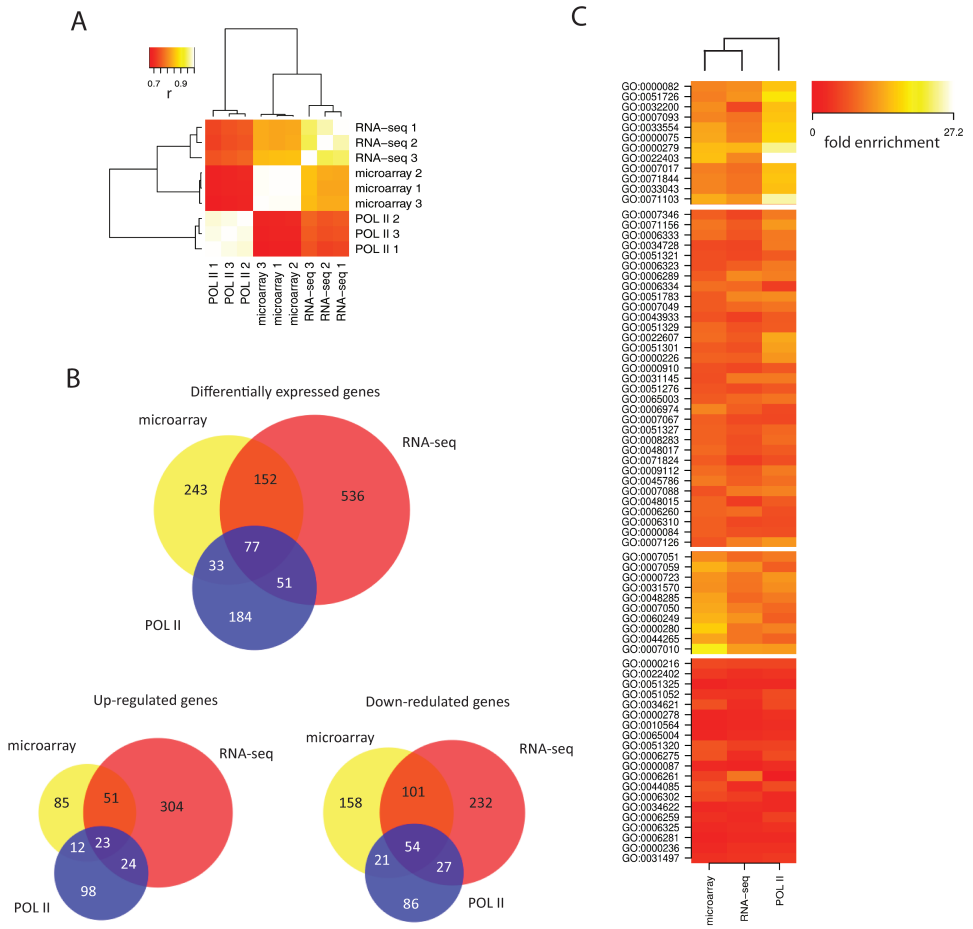


Figure 1: Comparison of Pol II ChIP-seq, RNA-seq and microarrays. A) Clustering of 3 different methods (RNA-seq, gene-expression microarray, Pol-II ChIP-seq) and biological replicates based on correlations of absolute gene-expression levels as measured by Pol II occupancy, RNA-seq, and normalized probe intensities from microarrays. B) Overlap between differentially expressed genes as determined by POL II occupancy, RNA-seq, and microarrays. C) Gene ontology analysis of down-regulated genes (Wnt target genes) identified by the same three methods. Each displayed term was found significantly enriched by at least one method. Functional classes in first cluster are found less enriched in genes identified by methods based on measuring RNA-levels, compared to genes identified based on POL II occupancy.

regulated by a particular transcription factor based on genome-wide TF-gene distribution patterns. Without the information about long-range chromatin interactions, this score only has a probabilistic character, since many TFs can regulate genes that are up to several hundreds of kilobases away from a binding site, while on the other hand neighboring genes do not necessarily have to be regulated (31,32). We therefore defined a TF-transcript association score (TTAS) in which we combined previously described scores based on TF-DNA interaction strength, distance to transcription start site and genome-wide binding site distribution with respect to transcription start sites (22,23) (Supplementary

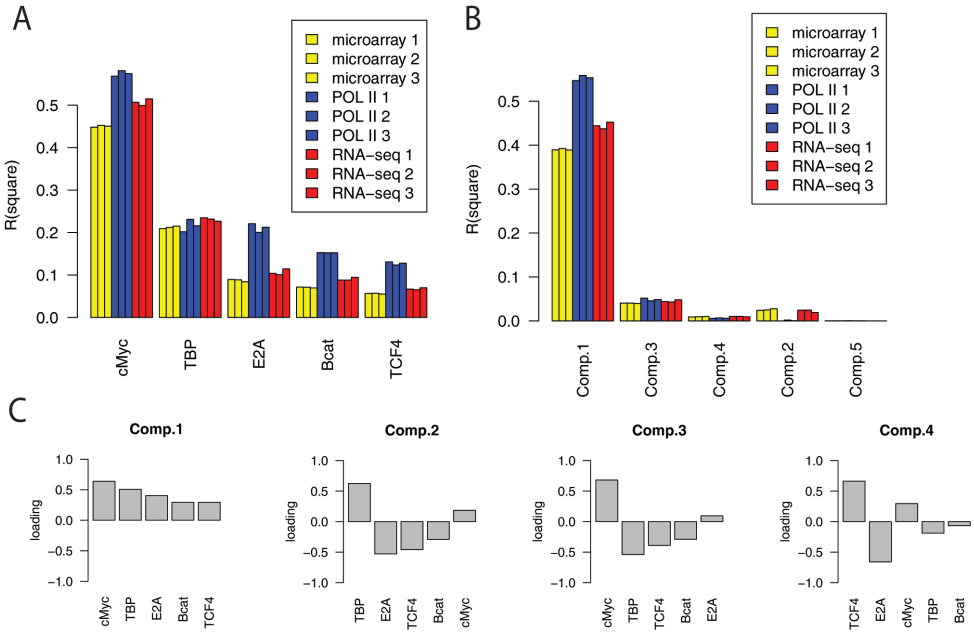
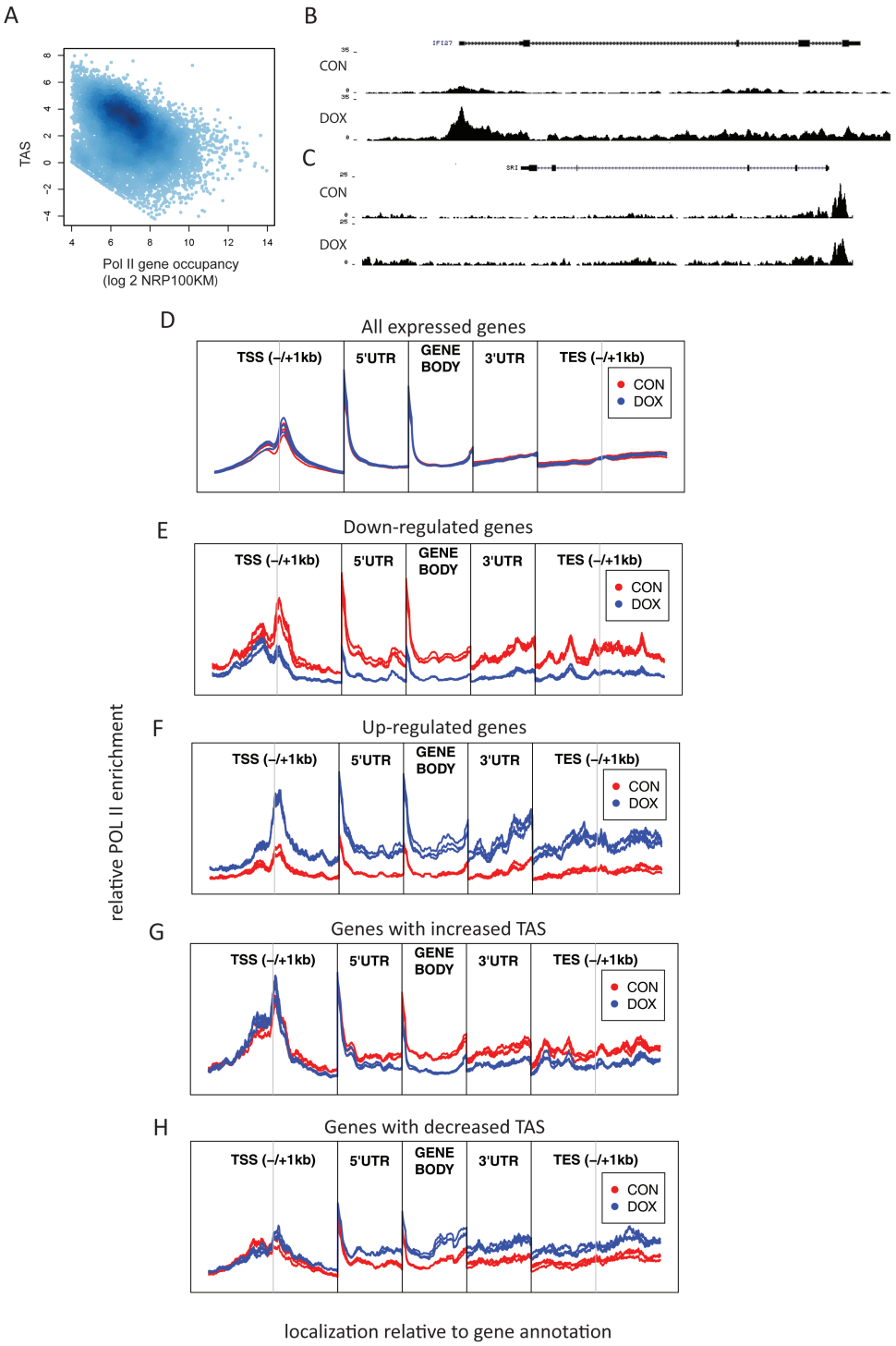


Figure 2: Linear regression analysis of Pol II occupancy, RNA-seq and microarray probe intensities with A) transcription factor occupancy represented by individual TTAS scores separately for each transcription factor and with B) individual principal components extracted from TTAS of 5 transcription factors. TTAS scores reflect the likelihood of a gene being regulated by a given transcription factor. C) Loading s of TTAS scores into four major principal components which explain most of variability in Pol II occupancy. Individual loadings are multiplied by -1 when correlation of principal component with Pol II occupancy is negative.

Figure 1). Linear regression analysis of individual TTAS together with Pol II occupancy or RNA levels and principal component regression analysis (PCRA) of individual principal components extracted from TTAS (Fig 2A, 2B) shows a better correlation of transcription factor occupancy with Pol II occupancy as compared to RNA-seq and microarray-based RNA level measurements.

More than half of the variation in Pol II occupancy can be explained by first principal component. In this principal component all five TFs have positive loadings (Fig 2C), indicating their major role in activation of gene expression (23). More importantly, high degree of co-linearity also explains why first principal component does not reflect more of the Pol II occupancy variation than cMyc alone. The other principal components explain only minority of variation in Pol II occupancy and RNA levels, however they may reveal existence of genes with different mechanisms of regulation, for example. in subgroup of genes with high third principal component values, the presence of TCF4, β-catenin and even TBP may lead to transcriptional repression, while second principal component suggests existence of mechanisms where TF regulate RNA levels without changing Pol II occupancy. Positive loadings of cMyc in all five principal components support its role as pure transcriptional activator in cancer cells what supports his role in regulation of transcriptional pause release (33).


A
TAS
Pol II gene occupancy
(log 2 NRP100KM)
B
IFI27
CON
DOX
C
SR1
CON
DOX
D
All expressed genes
TSS (–/+1kb)
5'UTR
GENE BODY
3'UTR
TES (–/+1kb)
CON
DOX
E
Down-regulated genes
F
Up-regulated genes
G
Genes with increased TAS
H
Genes with decreased TAS
relative POL II enrichment
localization relative to gene annotation

ENCODE datasets

To strengthen these findings, we repeated the analysis using publically available ChIP-seq profiles of 21 different TFs performed in duplicates, Pol II ChIP-seq and RNA-seq datasets from a lymphoblastoid cell line (GM12878) that was generated as a part of the ENCODE project (21,34) (Supplementary Material II). Both linear regression analysis of individual TTAS and PCRA showed very similar results as for our datasets, with Pol II ChIP-seq correlating better with TF occupancy than cellular RNA levels (Supplementary Figure 4).

Poised polymerase

Since our model for readout of gene expression expects all DNA bound Pol II to be processive, we next evaluated the potential effects of poised polymerase (35) and polymerase that is accumulated at 3'-UTRs on the readout and its correlation with TFs. We repeated the analysis while excluding the information from sequencing tags mapped close to 3'-UTRs and close to TSS where the majority of the poised polymerase is located. However, similar correlations with transcription factor occupancy as for total Pol II occupancy were obtained (Supplementary Figure 5), suggesting that non-processive poised polymerase and polymerase accumulated at 3'-UTR reflect only a minor fraction of DNA-bound Pol II compared to processive polymerase and thus does not have a major influence on the overall readout of gene expression based on Pol II ChIP-seq.

Nevertheless, quantification of changes in poised polymerase levels could be useful for dissection of regulatory mechanisms of individual genes as regulation of gene transcription after recruitment of RNA polymerase II to transcription start sites is a widely accepted mechanism of regulation of mRNA levels (35). In a simplified model, lowly expressed genes have a relatively high accumulation of poised Pol II around their transcription start sites compared to actively transcribed genes, where Pol II is spread over the complete genomic region. Thus, a change in Pol II accumulation close to the transcription start site (TSS) compared to accumulation in gene body can reflect changes in transcription. We therefore explored the possibility of using changes in TSS accumulation scores (TAS) to determine differentially expressed genes. The TAS score represents the relative enrichment of POL II at the promoter compared to the gene body. Indeed, the reduction of Pol II in TSS reflected by decreased TAS correlates (r = -0.366) with an increase in Pol II occupancy of the gene (Fig 3A). Next, we identified 481 genes that reproducibly changed TAS after WNT signaling inhibition by doxycyclin treatment in all 3 replicates. 78 (23%) of the genes that were identified by Pol II ChIP-seq as differentially expressed based on a change in the number of tags aligned to the whole transcript overlap with the set of genes with change in TAS score. This suggests that even though regulation of gene expression is typically

◁ **Figure 3: Poised polymerase and metagene analysis of Pol II occupancy.** A) Correlation of total Pol II occupancy with TAS score. The TAS score represents the relative enrichment of POL II at the promoter compared to the gene body and reflects the fraction of poised polymerase. Genes with lower expression have more Pol II deposited on their transcription start site and less processing polymerase in gene body compared to genes with higher expression (r = -366). B)C) Pol II coverage over up-regulated gene with B) increased density over TSS and gene body and C) increased density in gene body without change in TSS. D) Relative enrichment of POL II sequencing tags in Wnt plus (CON) and Wnt minus (DOX) samples with respect to gene annotation in D) all annotated genes E) down-regulated and F) up-regulated genes. POL II enrichment changes simultaneously in TSS and gene body, suggesting that a substantial proportion of transcription regulation is mediated by changes in POL II recruitment. In subclass of genes, with increase G) or decrease H) in relative enrichment of POL II at the promoter compared to gene body; difference of POL II accumulation in downstream part of genes is not accompanied by change in enrichment on TSS. Every individual region is normalized separately against input and average enrichment of CON samples.

mediated by an increase or decrease in POL II recruitment to TSS, which leads to change in Pol II occupancy in downstream parts of genes (Fig 3D-F), in some genes changed Pol II occupation in gene body is not accompanied by changed Pol II occupancy at TSS and therefore in these genes major regulation is mediated via changed frequency of releasing paused polymerase without a need for regulation of POL II recruitment (Fig 3G 3H). This mechanism has been previously shown as major regulator mechanisms of cMyc (33); transcription factor which is also active in our model system. In conclusion, additionally to the quantitative aspect of Pol II ChIP-seq, the spatial distribution information of Pol II can be used as additional information to develop more reliable models for the identification of specific subsets of differentially expressed genes that are predominantly regulated by releasing paused polymerase (33).

Bimodal distribution of Pol II occupancy

Pol II DNA occupancy and RNA levels measured by RNA-seq and microarrays show a bimodal distribution pattern that is specific and reproducible for every method (Fig 4A). A similar picture is observed at the genome-wide level of Pol II occupancy where more than half of the genome is occupied by Pol II with a domain-like distribution (Supplementary Figure 6), most likely reflecting open and closed chromatin regions. This pattern can be used to split genes into expressed and non-expressed groups(Fig 4B). Pol II ChIP-seq was found to classify more genes as expressed compared to the other methods. Many genes that are defined as expressed by Pol II ChIP-seq appear as non-expressed by the other two methods (Fig 4C, 4D). This is, however, not true in the opposite direction: genes defined as expressed by RNA-seq or microarray are virtually always categorized

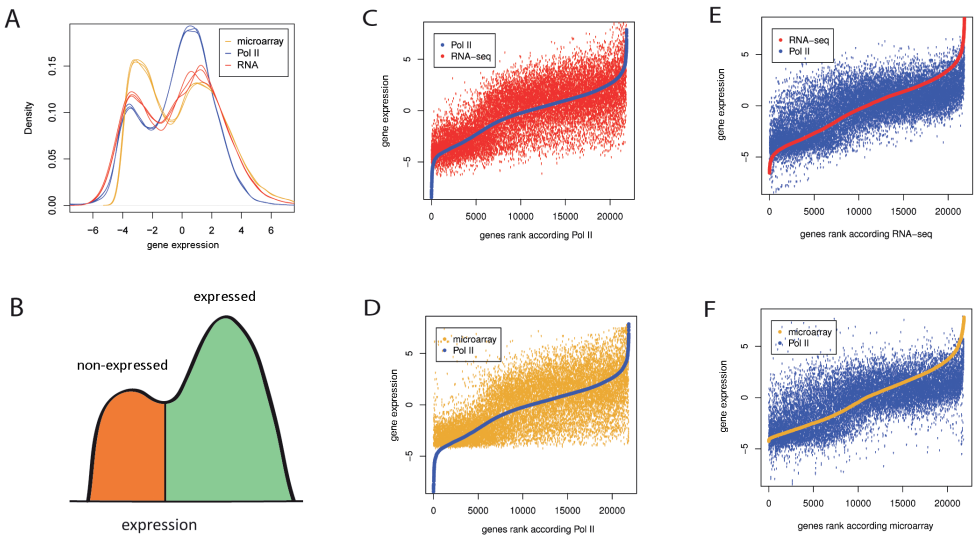


Figure 4: Bimodal distribution of Pol II occupancy, RNA-seq and microarray probe intensities A) All methods reveal a bimodal pattern of gene expression indicative of B) expressed (green) and non-expressed (red) genes. C) to F) Rank analysis of Pol II occupancy, RNA-seq and microarray probe intensities. Genes are ranked according to Pol II ChIP-seq results C) D), according to RNA-seq E) or microarrays F). Many transcripts classified as expressed according the Pol II ChIP-seq, are called as not expressed according to the RNA-seq and microarray data. In contrast, only a very limited number of transcripts that are called transcribed by RNA-seq and microarrays are called as non-transcribed by Pol II ChIP-seq. All expression values represent median centered and log2 transformed NR100KM (Pol II), NR10KM (RNA-seq) and normalized microarray probe intensity.

as transcribed by Pol II ChIP-seq (Fig 4E, 4F). In line with this, expressed and non-expressed genes form more distinct clusters in a principle component analysis (Fig 5) when these classes are defined from Pol II ChIP-seq than from RNA levels. Furthermore, expression classes defined from Pol II occupancy are more accurately predicted by classification and regression tree (CART) algorithm (25), when using principle components extracted from TTAS scores as input to predict expression status of all genes. In our dataset, only 11.8% of genes were wrongly predicted when Pol II was used to define expression status of a gene, with 14.8% and 17.6% percent of wrongly predicted genes when RNA-seq and microarrays were used to determine expression status.

Gene ontology analysis

To determine if the observed differences in RNA levels and POL II occupancy reflect functional gene classes, we performed a gene ontology term (GO) enrichment analysis (36). We empirically defined three classes containing genes that are expressed as defined by Pol II ChIP-seq but as non-expressed (class I), lowly expressed (class II) and very highly expressed (class III) by RNA-seq (Supplementary Figure 7). Interestingly, a significant enrichment of particular GO terms in different gene classes is observed (Supplementary Material IV), which overlaps with previously described gene classes that are characterized by particular low or high mRNA turnover rates (2-4). Genes in class I are enriched in secreted proteins and plasma membrane receptors and reflect regulated genes for which mRNA is synthesized but rapidly degraded. Class II genes are enriched in transcriptional regulators, while the third class includes genes that are involved in basal translational homeostasis and energy metabolisms, characterized by high mRNA levels. These analyses indicate that gene expression of particular gene classes with short living mRNA can be underestimated by methods measuring solely RNA levels while the expression of gene classes with more stable transcripts can be systematically overestimated.

Concluding remarks

Taken together, we have shown that routine Pol II ChIP-seq method provides valuable information that is complementary to total mRNA level measurements and allows us to independently dissect separate mechanisms involved in gene expression i.e. allow us to separately study changes in production of

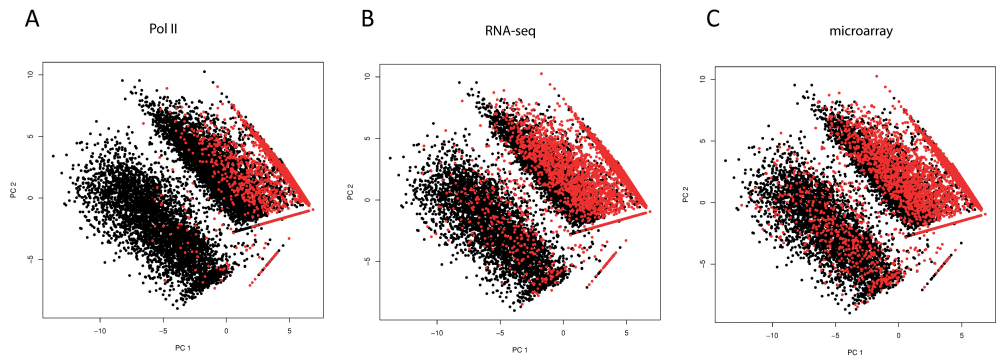


Figure 5: Clustering of expressed (black) and non-expressed (red) genes defined by POL II chip-seq, RNA-seq and microarrays. Genes are plotted according to principal components extracted from TTAS scores, which reflect the likelihood of a gene being regulated by a given transcription factor. Genes categorized into expressed and non expressed according bimodal distribution of Pol II results A) form more defined clusters compared to genes categorized according to RNA levels B) C).

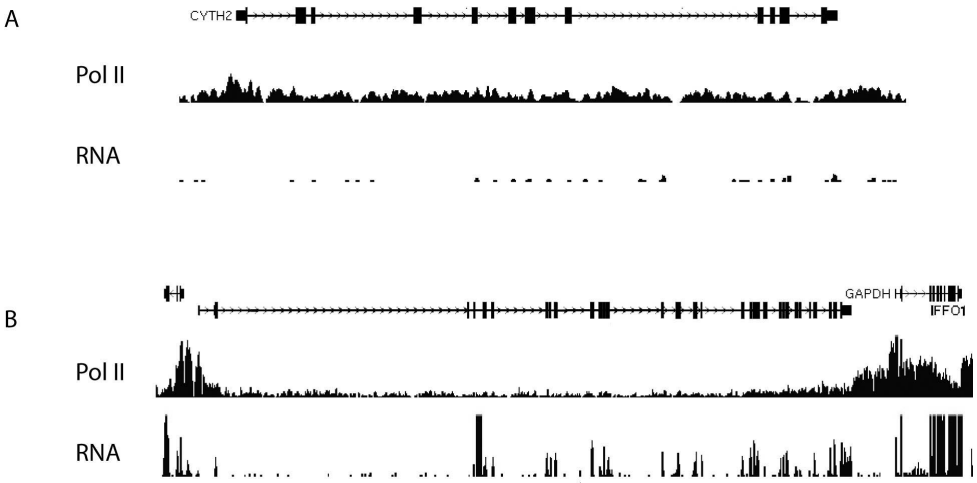


Figure 6 Examples Pol II and RNA-seq coverage over of genes with low (A) and high (B) RNA stability. Vertical axis represents sequencing relative tags coverage per position A) Gene with high density of sequencing tags over gene body with very low levels of RNA. Pol II ChIP-seq classifies depicted gene as expressed, however both RNA-seq and microarrays classify the gene as non-expressed. (Class I gene) B) Genes with high density of Pol II ChIP-seq tags mapping to gene body with high numbers of sequencing tags from RNA library mapping to annotated exons.

RNA and mechanisms responsible for RNA stability. In addition, we showed that due to highly variable RNA stability, gene expression levels and dynamics of particular gene classes are systematically over- or under-estimated when solely RNA levels are used to determine gene expression.

PCRA revealed that Pol II ChIP-seq correlates better with transcriptional regulators and also showed that in limited number of genes, transcription factors may surprisingly regulate RNA levels without affecting Pol II occupancy. We were able to recapitulate these findings using publicly available datasets when different combination of TFs was used, supporting universal nature of our findings.

In addition to quantitative aspect of Pol II ChIP-seq, we showed that spatial distribution of Pol II and its changes can be used as additional information to identify of specific subsets of differentially expressed genes that are predominantly regulated by releasing paused polymerase instead of increasing rate of polymerase recruitment.

Altogether Pol II ChIP-seq integrated with methods determining RNA levels has potential to serve as versatile tool to dissect individual mechanisms regulating gene expression, applicable in variety of model systems.

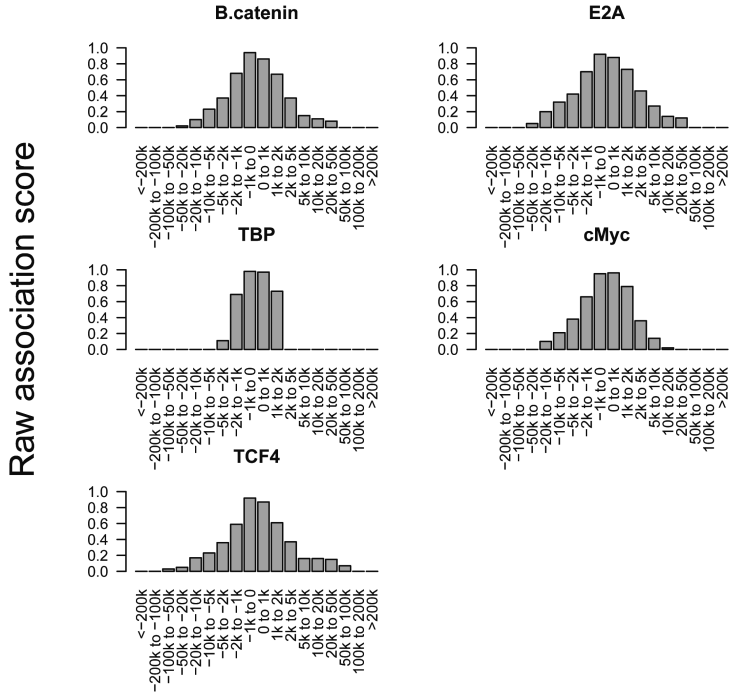
REFERENCES

1. Shyu, A.B., Wilkinson, M.F. and van Hoof, A. (2008) Messenger RNA regulation: to translate or to degrade. EMBO J, 27, 471-481.
2. Raghavan, A., Ogilvie, R.L., Reilly, C., Abelson, M.L., Raghavan, S., Vasdewani, J., Krathwohl, M. and Bohjanen, P.R. (2002) Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. Nucleic Acids Res, 30, 5529-5538.
3. Sharova, L.V., Sharov, A.A., Nedorezov, T., Piao, Y., Shaik, N. and Ko, M.S. (2009) Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. DNA Res, 16, 45-58.
4. Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D. and Brown, P.O. (2002) Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A, 99, 5860-5865.
5. Garcia-Martinez, J., Aranda, A. and Perez-Ortin, J.E. (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Mol Cell, 15, 303-313.
6. Miller, C., Schwalb, B., Maier, K., Schulz, D., Dumcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dolken, L. et al. (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. Mol Syst Biol, 7, 458.
7. Churchman, L.S. and Weissman, J.S. (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature, 469, 368-373.
8. Core, L.J., Waterfall, J.J. and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322, 1845-1848.
9. Mokry, M., Hatzis, P., de Bruijn, E., Koster, J., Versteeg, R., Schuijers, J., van de Wetering, M., Guryev, V., Clevers, H. and Cuppen, E. (2010) Efficient Double Fragmentation ChIP-seq Provides Nucleotide Resolution Protein-DNA Binding Profiles. PLoS ONE, 5, e15092.
10. Adli, M., Zhu, J. and Bernstein, B.E. (2010) Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat Methods, 7, 615-618.
11. Dahl, J.A. and Collas, P. (2008) A rapid micro chromatin immunoprecipitation assay (microChIP). Nat Protoc, 3, 1032-1045.
12. Dahl, J.A. and Collas, P. (2008) MicroChIP--a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. Nucleic Acids Res, 36, e15.
13. Dahl, J.A., Reiner, A.H. and Collas, P. (2009) Fast genomic muChIP-chip from 1,000 cells. Genome Biol, 10, R13.
14. Fanelli, M., Amatori, S., Barozzi, I., Soncini, M., Dal Zuffo, R., Bucci, G., Capra, M., Quarto, M., Dellino, G.I., Mercurio, C. et al. (2010) Pathology tissue-chromatin immunoprecipitation, coupled with high-throughput sequencing, allows the epigenetic profiling of patient samples. Proc Natl Acad Sci U S A, 107, 21535-21540.
15. van de Wetering, M., Oving, I., Muncan, V., Pon Fong, M.T., Brantjes, H., van Leenen, D., Holstege, F.C., Brummelkamp, T.R., Agami, R. and Clevers, H. (2003) Specific inhibition of gene expression using a stably integrated, inducible small-interfering-RNA vector. EMBO Rep, 4, 609-615.
16. Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res, 31, e15.
17. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 18, 1851-1858.
18. Smyth, G. (2005) In Gentleman, R., Carey, V., Huber, W., Irizarry, R. and Dudoit, S. (eds.), Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer-Verlag, pp. 397-420.
19. Hatzis, P., van der Flier, L.G., van Driel, M.A., Guryev, V., Nielsen, F., Denissov, S., Nijman, I.J., Koster, J., Santo, E.E., Welboren, W. et al. (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. Mol Cell Biol, 28, 2732-2744.
20. Ji, H., Jiang, H., Ma, W., Johnson, D.S., Myers, R.M. and Wong, W.H. (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol, 26, 1293-1300.
21. Myers, R.M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R.C., Bernstein, B.E., Gingeras, T.R., Kent, W.J., Birney, E., Wold, B. et al. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol, 9, e1001046.
22. Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J. et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 133, 1106-1117.
23. Ouyang, Z., Zhou, Q. and Wong, W.H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. Proc Natl Acad Sci U S A, 106, 21521-21526.
24. Development Core Team. (2009) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.
25. Hastie, T., Tibshirani, R. and Friedman, J. (2003) The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
26. Therneau, T.M. and Atkinson, B. (2009) rpart: Recursive Partitioning.

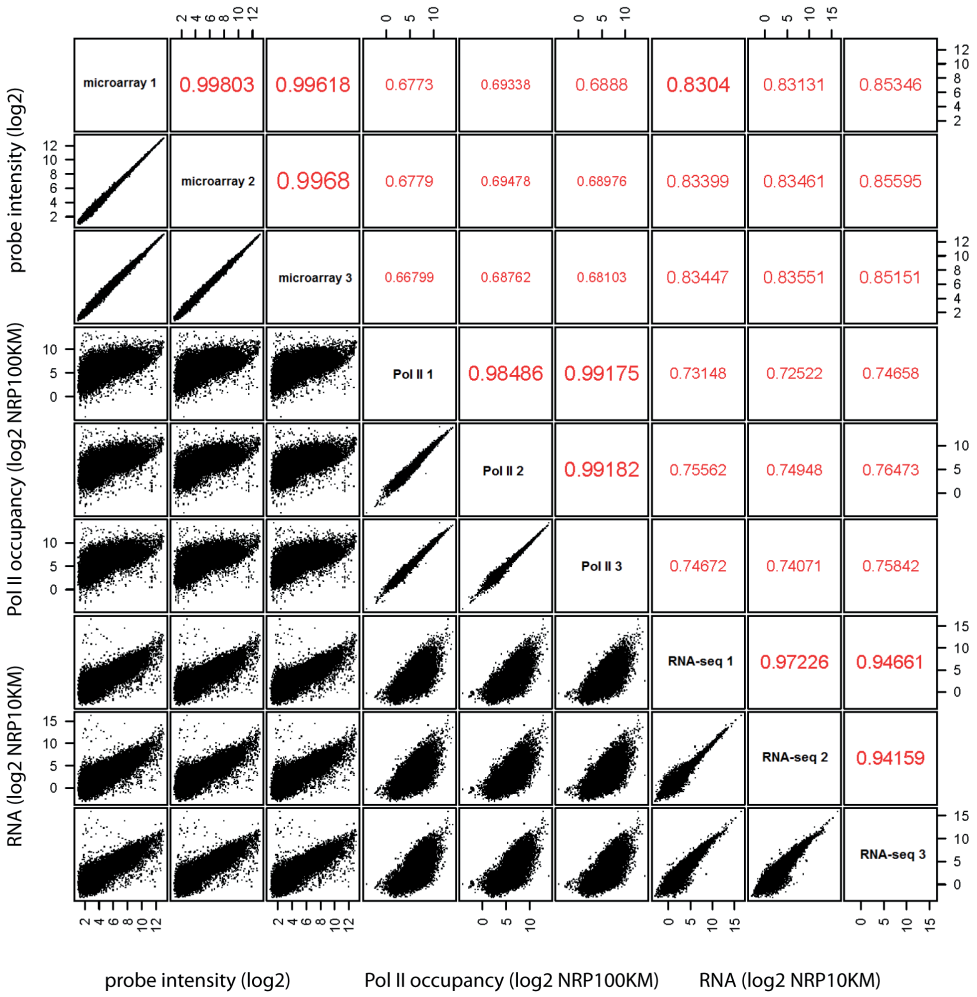
27. Besse, S., Vigneron, M., Pichard, E. and Puvion-Dutilleul, F. (1995) Synthesis and maturation of viral transcripts in herpes simplex virus type 1 infected HeLa cells: the role of interchromatin granules. Gene Expr, 4, 143-161.
28. Korinek, V., Barker, N., Morin, P.J., van Wichen, D., de Weger, R., Kinzler, K.W., Vogelstein, B. and Clevers, H. (1997) Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. Science, 275, 1784-1787.
29. Graham, T.A., Weaver, C., Mao, F., Kimelman, D. and Xu, W. (2000) Crystal structure of a beta-catenin/Tcf complex. Cell, 103, 885-896.
30. He, T.C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., da Costa, L.T., Morin, P.J., Vogelstein, B. and Kinzler, K.W. (1998) Identification of c-MYC as a target of the APC pathway. Science, 281, 1509-1512.
31. Kooren, J., Palstra, R.J., Klous, P., Splinter, E., von Lindern, M., Grosveld, F. and de Laat, W. (2007) Beta-globin active chromatin Hub formation in differentiating erythroid cells and in p45 NF-E2 knock-out mice. J Biol Chem, 282, 16544-16552.
32. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature, 462, 58-64.
33. Rahl, P.B., Lin, C.Y., Seila, A.C., Flynn, R.A., McCuine, S., Burge, C.B., Sharp, P.A. and Young, R.A. (2010) c-Myc regulates transcriptional pause release. Cell, 141, 432-445.
34. Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M. et al. (2009) Unlocking the secrets of the genome. Nature, 459, 927-930.
35. Margaritis, T. and Holstege, F.C. (2008) Poised RNA polymerase II gives pause for thought. Cell, 133, 581-584.
36. Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48.
37. Anders, S. (2009) Visualization of genomic data with the Hilbert curve. Bioinformatics, 25, 1231-1235.

Supplementary Table, Sequencing and mapping statistics.

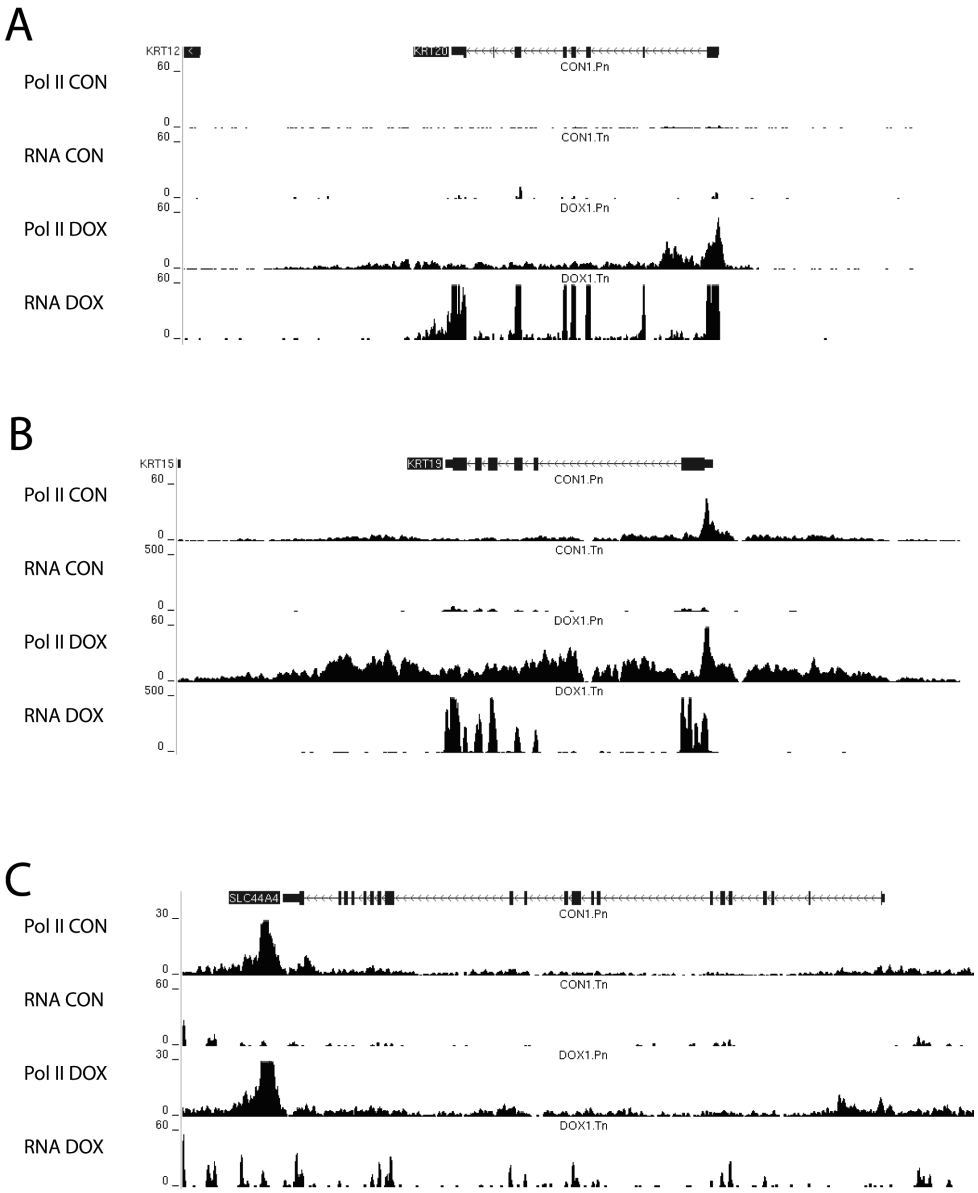
	Sample	Raw reads	Mapped reads	Mapped reads (q>0)	Percentage mapped	Percentage with q>0
RNA sequencing	CON1	43,519,947	19,409,978	11,469,611	44.6	59.1
	CON2	32,962,479	15,120,183	8,890,432	45.9	58.8
	CON3	46,869,848	20,349,215	11,451,852	43.4	56.3
	DOX1	31,602,468	13,622,531	7,625,974	43.1	56.0
	DOX2	37,447,267	15,428,604	9,565,192	41.2	62.0
	DOX3	42,341,072	15,213,191	8,315,499	35.9	54.7
Pol II ChIP sequencing	CON1	65,455,910	41,065,573	36,186,881	62.7	88.1
	CON2	63,256,626	41,829,937	37,054,927	66.1	88.6
	CON3	52,403,736	33,479,940	29,424,271	63.9	87.9
	DOX1	69,649,772	35,978,193	31,798,338	51.7	88.4
	DOX2	57,203,495	37,304,546	33,257,995	65.2	89.2
	DOX3	60,205,429	34,741,976	30,692,140	57.7	88.3



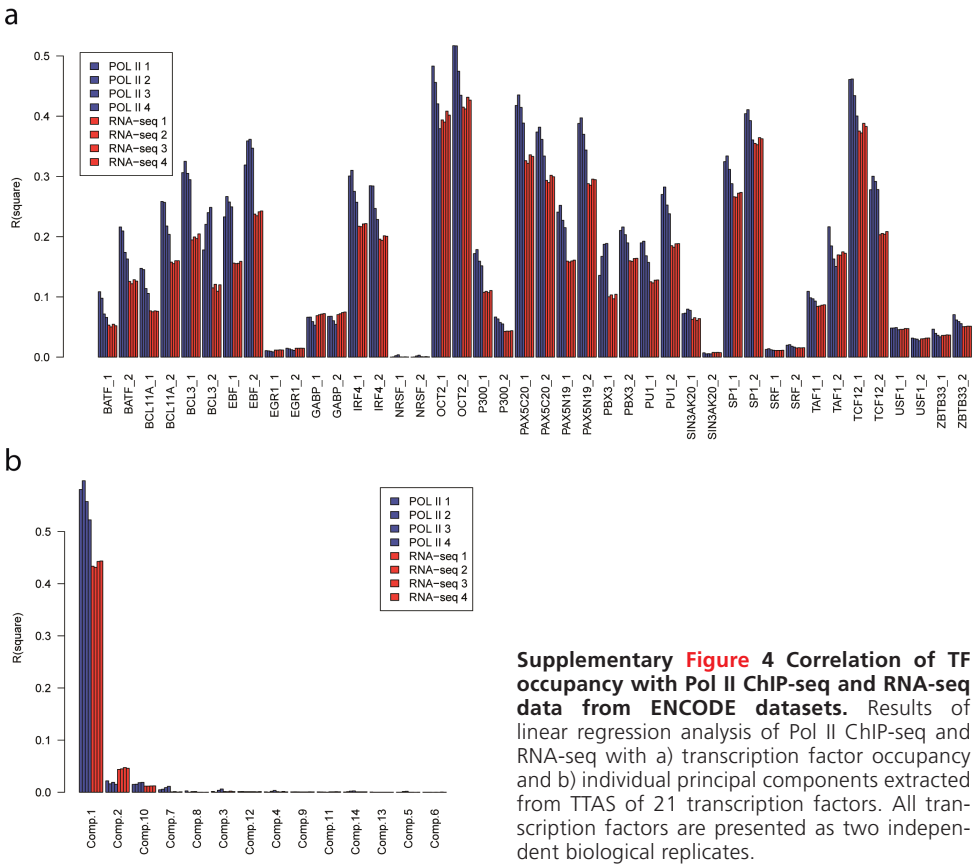
Supplementary Figure 1: Raw scores used for calculation of TTAS for five transcription factors depending on the distance from transcription start site of the associated gene. TTAS scores reflect the likelihood of a gene being regulated by a given transcription factor within set distance from gene's TSS. Some transcription factor binding sites are distributed exclusively close to transcription start sites. Others can regulate more distant genes.



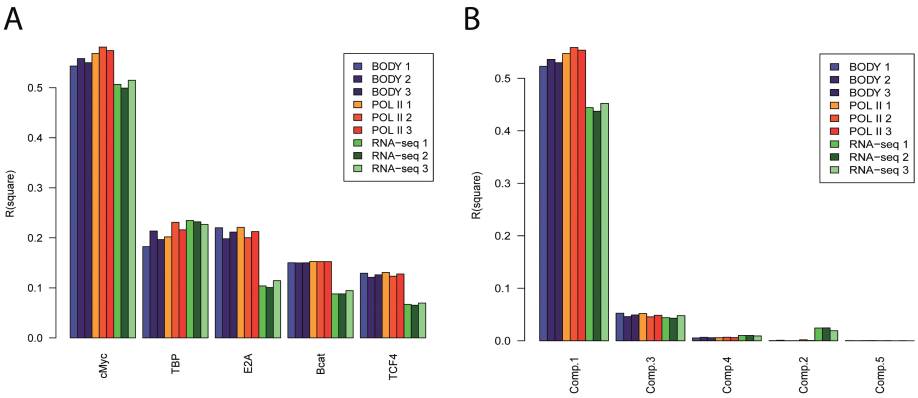
Supplementary Figure 2: Comparison of of Pol II occupancy and RNA levels as measured by RNA--seq and microarrays and their reproducibility over 21,854 RefSeq genes. Each method shows good reproducibility between all 3 replicates. Correlations between different methods are decreased**.** Red numbers represent pair-wise Pearson's correlation coefficients.



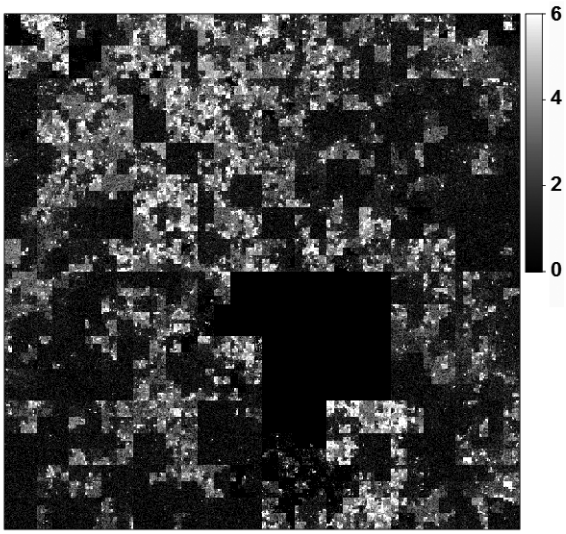
Supplementary figure 3: Examples of Pol II and RNA-seq coverage over differentially expressed genes under two experimental conditions. Vertical axis represents sequencing tags coverage per position A)B) Upregulated genes shows increase in tag density for both Pol II and RNA-seq. C) Pol II tag density over gene body is increased less extensively after doxycycline treatment compared to RNA levels of this gene. Up-regulation of gene comes with increased mRNA stability. Control sample - CON and doxycyclin treated sample - DOX.



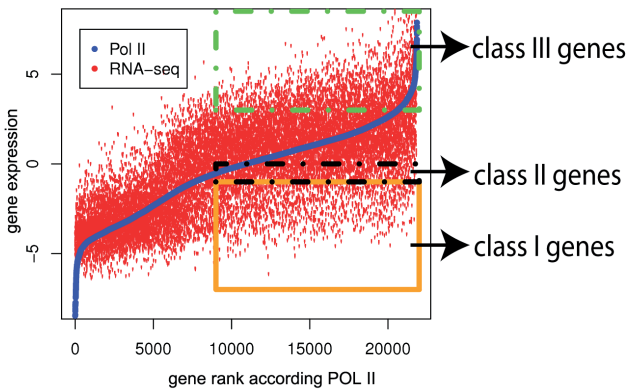
Supplementary Figure 4 Correlation of TF occupancy with Pol II ChIP-seq and RNA-seq data from ENCODE datasets. Results of linear regression analysis of Pol II ChIP-seq and RNA-seq with a) transcription factor occupancy and b) individual principal components extracted from TTAS of 21 transcription factors. All transcription factors are presented as two independent biological replicates.



Supplementary Figure 5: Effect of poised polymerase and polymerase accumulated on 3'-UTRs on correlation of Pol II ChIP-seq with transcription factor occupancy A) Correlation of Pol II presence in gene body excluding 3'UTR and first 300 bp ("BODY"), Pol II presence in whole gene transcript ("POL II") and by RNA-seq with transcription factor occupancy B) Correlation of Pol II occupancy with individual principal components extracted from TTAS of 5 transcription factors. Excluding pol II tags bound close to transcription start sites and 3'-UTRs has only a minor influence on the correlation of Pol II ChIP-seq data with TF occupancy.



Supplementary Figure 6 Pol II occupancy on chromosome 1 reveals that major part of chromosome 1 is occupied by Pol II. Pol II occupation is restricted to distinguishable domains. For better visualization of domains, chromosome 1 coordinates are represented as continuous space filling curve (Hilbert curve). Color log2 scale represents quantile normalized number of Pol II tags mapped to 500bp bins. Large area without coverage of POL II sequencing tags is represents centromere.



Supplementary Figure 7 Rank analysis of Pol II occupancy with empirically defined gene classes (I, II, III) used for gene ontology term enrichment analysis. Genes are ranked according to Pol II ChIP-seq results. Genes in all classes are expressed as defined by Pol II ChIP-seq and seem to be non-expressed (class I), lowly expressed (class II) and very highly expressed (class III) by RNA-seq. All expression values represent median centered and log2 transformed NR100KM (Pol II) and NR10KM (RNA-seq).

8

TCF4 PRESENCE DEFINES TWO CLASSES OF β-CATENIN BOUND REGIONS WITH DIFFERENT TRANSCRIPTIONAL OUTCOME

ABSTRACT

Upon activation of Wnt signaling β-catenin is recruited to enhancer sites by DNA-binding T cell factor/lymphoid enhancer factor (TCF/LEF) family members, thereby inducing transcription of Wnt target genes. Over the past years several other factors with the ability to recruit β-catenin independent of TFC/LEF were reported, suggesting direct interplays between the canonical Wnt pathway and other signaling pathways. To study the extent and role of TCF/LEF-independent β-catenin recruitment (TIBR) we used ChIP-seq to identify genomic elements occupied by β-catenin but not TCF4. We identified thousands of putative TIBR sites in two different model systems, although these regions generally represent only weaker β-catenin binding sites. These sites harbor only a minority of DNA-bound β-catenin, they are associated with different histone marks and chromatin status and their presence correlates with negatively regulated Wnt-target genes. Surprisingly, over-expression of a dominant negative form of TCF4 (dnTCF4) that lacks the β-catenin binding domain results in depletion of β-catenin at these sites. Furthermore, dnTCF4 can bind to these sites, suggesting that all DNA-bound β-catenin involves recruitment through TCF familiy members.

INTRODUCTION

In canonical Wnt signaling the Wnt ligand binds to the Frizzled (Fz) receptor and forms a complex with the LRP protein, which after a series of events competes away Axin from the cytoplasmic destruction complex (CDC) and refrains CDC from phosphorylating β-catenin and subsequent targeting for ubiquitination. Cytoplasmic β-catenin is therefore stabilized and is translocated into the cell nucleus to drive specific transcriptional programs [1]. Alternatively, in pathological conditions β-catenin becomes constitutively stable when acquired mutations in β-catenin make it resistant to phosphorylation and degradation or one of the components of CDC is inactivated by mutations. These mechanisms play a major role in sporadic colorectal cancer [2].

Once in the nucleus, β-catenin is recruited to enhancer regions by TCF/LEF [3-6] where it acts as transcriptional activator of various target genes after displacing Groucho. At the chromatin level, many interacting partners were identified in genetic screens suggesting complex mechanisms of transcriptional regulation of Wnt target genes [7-11].

Besides TCF/LEF mediated recruitment of β-catenin to enhancer sites, numerous other transcription factors have been identified as capable to recruit β-catenin independently to TCF/LEF: nuclear receptors [8, 12], OCT4 [13], and others [2, 6]. However, none of these studies provide systematic approaches to substantiate the significance of the proportion of TCF/LEF-independent β-catenin recruitment. In this study we performed genome-wide systematic analyses of potential TIBR sites and provide supporting evidence that all nuclear β-catenin, also at TIBR sites, is mediated via TCF/LEF recruitment.

RESULTS AND DISCUSSION

Identification of β-catenin bound regions

To obtain systematic information about the genomic loci to which β-catenin is recruited we performed β-catenin ChIP-seq on Ls174T-L8 cells carrying a doxycyclin-inducible dominant-negative TCF4 (dnTCF4) transgene [14]. These cells carry an activating point mutation in β-catenin, which results in constitutively active canonical Wnt pathway, unless dnTCF4 is induced. We identified 2,295 high confidence β-catenin bound regions. In these regions, the most overrepresented sequence motif found is identical to the TCF4 binding motif that was identified in previous studies [15, 16] (Figure 1A). In addition, the presence of peaks correlates with transcriptional changes induced by over-expression of dnTCF4 (Figure 1B) supporting the biological relevance of the identified sites.

Identification of TIBR sites

As a next step, to identify β-catenin sites without the presence of TCF/LEF family member we preformed TCF4 ChIP-seq in parallel on the same chromatin sample as used for β-catenin ChIP-seq and explored the presence/absence of sequencing tags derived from Tcf4 ChIP-seq over β-catenin sites. In total, 814 β-catenin bound regions without presence of Tcf4 were identified (Figure 1C 1D), suggesting the existence of other recruiter than TCF4. This can be either another member of TCF/LEF family or another alternative recruiter.

To explore if these findings are consistent for different model systems we performed β-catenin ChIP-seq on HEK 293T cells after Wnt stimulation. While Ls174T-L8 cells represent a system with constitutively active Wnt signaling, in HEK cells Wnt signaling is not active until stimulation. We identified

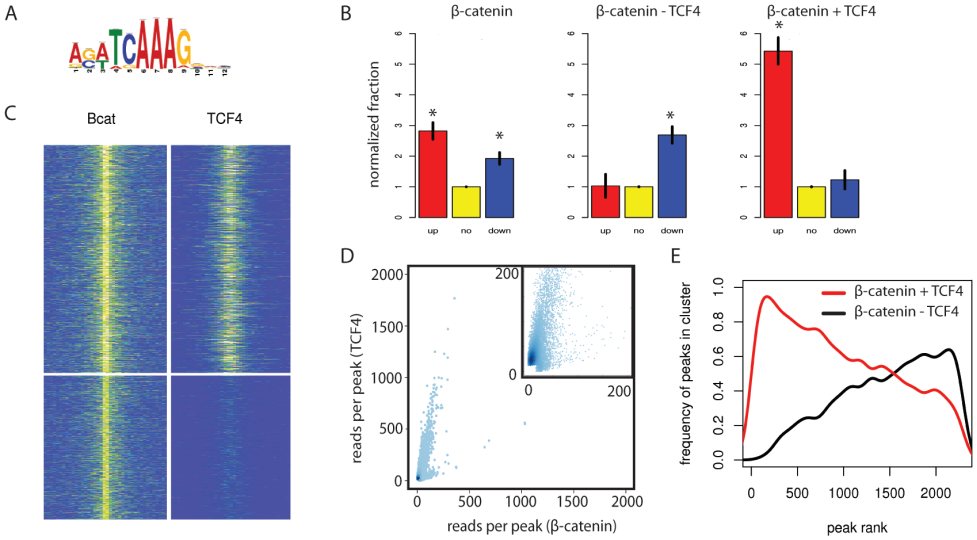


Figure 1: Two distinct classes of beta catenin bound regions in LS174T-L8 cells. A) The most over-represented sequence motif found in β-catenin peaks is similar to the canonical TCF4 binding motif. B) β-catenin peaks are associated with genes that are actively regulated by constitutively active Wnt-signaling. β-catenin sites without significant presence of TCF4 (TIBR) are more enriched for down-regulated genes compared to shared β-catenin/TCF4 sites, which are enriched in up-regulated genes. Error bars represent standard deviation of 10,000 randomized datasets. *$p<10^{-4}$. C) Heatmap presenting results of k-means clustering of β-catenin bound regions according to the presence of TCF4. D) Occupation of sites bound by either TCF4 or β-catenin by sequencing tags derived from β-catenin or Tcf4 ChIP-seq. E) Representation of different peak ranks among two β-catenin peak classes.

5,020 β-catenin bound regions (Supplementary Figure 1) and 1,770 TIBR sites, with only 285 and 14 overlapping with regions found in LS cells. Small overlap of β-catenin bound regions in two different cell lines is in line with high dynamics of distant enhancer regions [17] with more variable TIBR sites compared to TCF4 mediated regions.

Two distinct classes of β-catenin bound regions

A high proportion of nuclear β-catenin sites was identified as potential TIBR sites (814 out of 2295). However, these clearly represent weaker (lower ranked) β-catenin peaks (Figure 1E), indicating that that the majority of β-catenin binding is mediated directly by TCF4 (Supplementary Figure 2). Both the presence of TIBR sites as well as proper β-catenin/TCF4 sites correlate with transcriptional changes induced by dnTCF4 (Figure 1B). However TIBR sites are more frequently found close to down-regulated genes compared to β-catenin/TCF4 sites, which associate with up-regulated genes. These findings confirm the established role of β-catenin/TCF4 sites as enhancer, but suggest a repressor role for TIBR sites.

Surprisingly, TIBR sites in LS cells have a different distribution with respect to transcriptional start sites (TSS). They are more often found close to TSS compared to β-catenin/TCF4 sites (Figure 2A). TIBR sites also are more frequently co-occupied by RNA Polymerase II (Pol II) (Figure 2B) compared to β-catenin/TCF4 sites. This may suggest that part of these sites are not distant enhancers but are indirectly co-precipitated proximal promoters due to chromatin interactions with the enhancer site after β-catenin

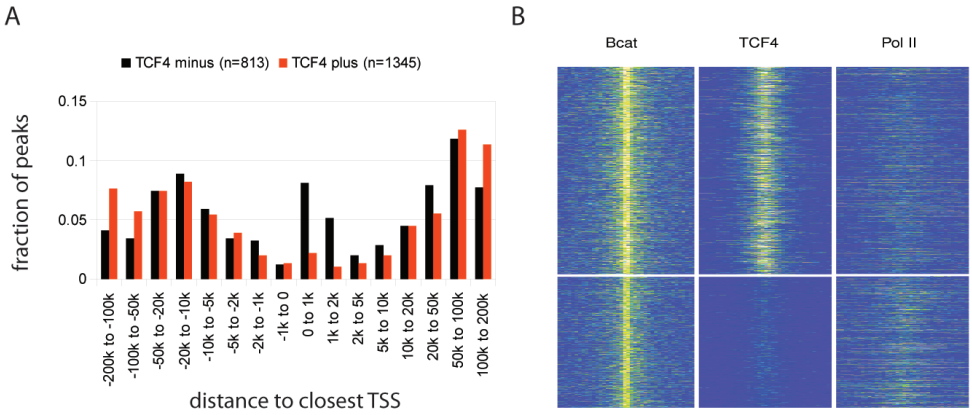


Figure 2: Distribution of β-catenin peaks with respect to transcription start sites (TSS). A) β-catenin sites without significant TCF4 presence are closer to TSS compared to β-catenin sites with TCF4. B) Heatmap showing presence of β-catenin, TCF4 and RNA Polymerase II in β-catenin bound regions.

recruitment [18]. However, in HEK cells distribution of TIBR sites and β-catenin/TCF4 sites with respect to transcriptional start sites appears similar, pointing towards different nature of TIBR sites in system with transient Wnt activation compared to system with constitutively active of Wnt signaling.

Even though TIBR sites are enriched for the canonical TCF4 binding motif compared to control regions, this enrichment is weaker compared to β-catenin/TCF4 sites (Figure 3). On the other hand some other binding motifs tend to be more enriched on TIBR sites (Figure 3), suggesting the possible presence of alternative recruiters. However, these motifs are also found for β-catenin/TCF4 sites, suggesting a shared mechanism.

TIBR sites are more frequently associated with inactive chromatin signature

As a next step we explored the presence of 8 histone marks and CTCF close to β-catenin bound regions (Figure 4). We used publicly available genome-wide maps from ENCODE project from HEK 293T cells [17]. Interestingly, even though association of TIBR sites with different histone marks are grossly similar as for TCF4 mediated β-catenin bound regions, TIBR sites are more frequently associated with repressive chromatin stage and insulators. In line with this, TIBR sites are more likely associated with less active enhancers and insulators compared to β-catenin/TCF4 sites.

Functional analysis of TIBR sites

We next explored whether identified TIBR sites have WNT dependent transcriptional regulatory activity. From 8 candidate TIBR sites that were cloned in front of a minimal promoter, only one was capable to slightly enhance luciferase reporter activity with the other 7 having no or even repressive activity. In comparison, similar experiments for TCF4 bound regions typically lead to 50% success rate [15] suggesting different nature of TIBR sites.

dnTCF4 binds to both TIBR and B-catenin/TCF sites

To evaluate the possible role of other TCF/LEF members than TCF4 and that act in a redundant way in recruitment of β-catenin in TIBR sites we over-expressed dnTCF4 in Ls174T-L8 cells. dnTCF4 lacks the β-catenin binding domain but retains the ability to

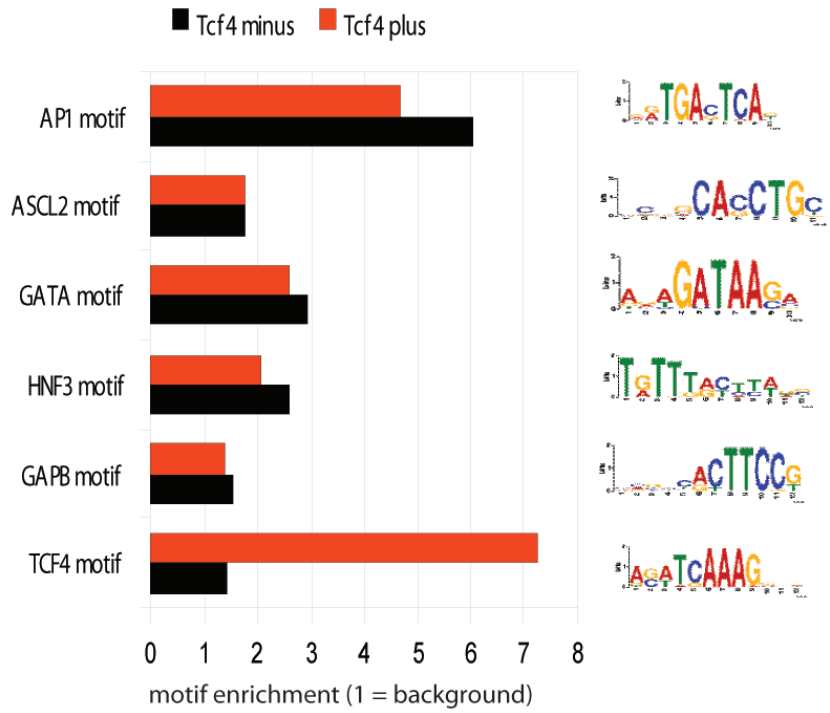


Figure 3: Enrichment of various sequence motifs in two classes of β-catenin peaks. Six most over-represented motifs found in β-catenin bound regions with and without presence of TCF4 in Ls174-L8 cells.

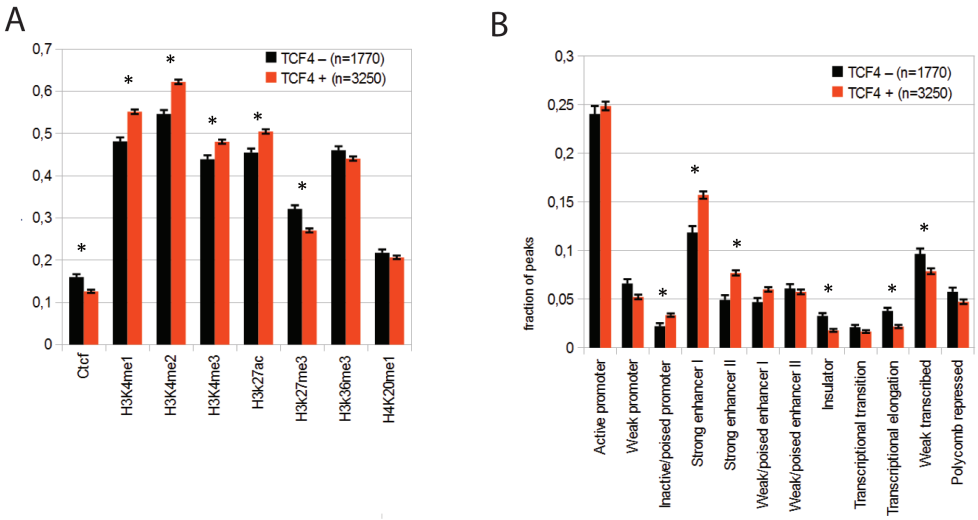


Figure 4 Association of A) histone marks, CTCF and B) different chromatin types from ENCODE dataset with TCF4-containing and TCF4-depleted classes of β-catenin bound genomic regions. Error-bars represent standard deviation of 1000 randomized datasets. * depicts significantly ($p<0.05$) different association of chromatin features between Tcf4 positive and TCF4 negative β-catenin peaks according chi-square test.

bind DNA enhancer sites. Because the consensus binding site is shared among TCF family members (reference), dnTCF4 can compete for binding with any wild-type TCF/LEF member would be good to have a reference for this part as well! is pretty crucial for interpretation. We performed ChIP for TCF4, β-catenin and FLAG (dnTCF4) in cells with and without over-expressed dnTCF4. Surprisingly, after deep sequencing of immunoprecipitated DNA fragments we found that dnTCF4 was recruited to TIBR sites and was capable to repress β-catenin binding (Figure 6). This suggests that TIBR sites are most likely occupied by a TCF4/LEF family member that is able to recruit β-catenin while the other possible models explaining existence of TIBR sites (Figure 7) are less likely. However, we cannot exclude the possibility that β-catenin is recruited via TCF4, while recruitment of other factors prevents efficient immunoprecipitation of TCF4 by masking it epitopes. Especially, when polyclonal antibody we used in this study is raised against short N-terminal part of TCF4, making full epitope masking more likely. Yet another model would be that β-catenin is recruited through TCF4 together, potentially together with other transcription factors and that TCF4 is released from this complex after stabilization, leaving b-catenin behind. However, it does not explain why we cannot detect proportion of TCF4 before it recruits β-catenin. Other mechanisms suggests co-immunoprecipitation of genomic regions which are spatially close to β-catenin bound regions due to enhancer looping; however one would expect similar co-immunoprecipitation for both TCF4 and β-catenin-mediated pull-downs.

Concluding remarks

Taken together, we observed two distinct functional classes of β-catenin peaks. The first class of β-catenin peaks is accompanied by the presence of TCF4, correlates with positively regulated Wnt-target genes, harbors the majority of DNA recruited β-catenin and represents classical distant enhancers. The second class of β-catenin-bound regions lacks significant amounts of TCF4 as detected by ChIP-seq assays, correlates with negatively regulated Wnt-target genes, harbors only a minority of DNA-bound

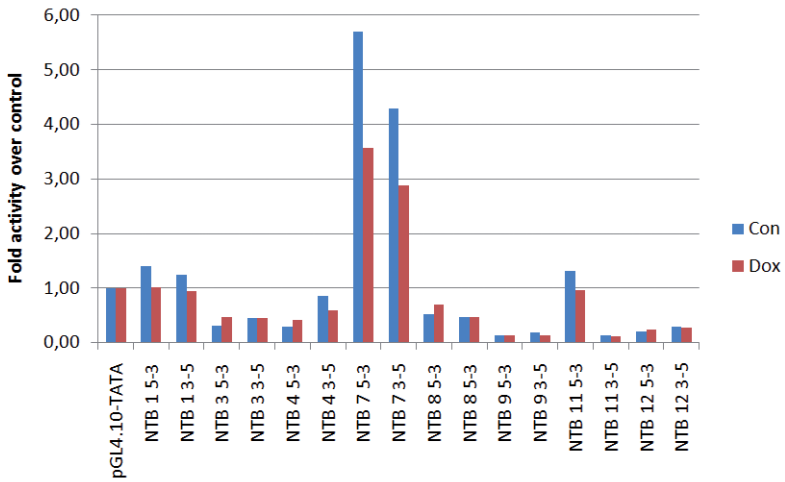


Figure 5: Enhancer activity of TIBR sites: TIBR sites were cloned in a firefly luciferase expression vector. The reporters were transfected into Ls174T-L8 cells with *Renilla* luciferase as transfection control. Values represent luciferase activity relative to empty pGL4.10 vector.

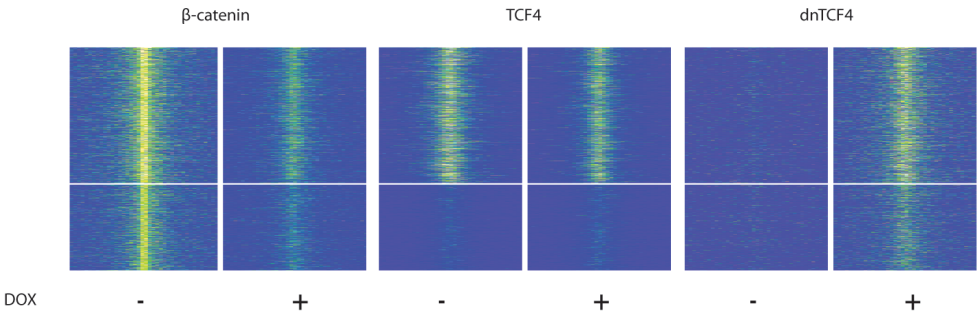


Figure 6: K-means clustering results of β-catenin bound regions according to the presence of TCF4. Signal from all individual ChIPs is arranged in the same order as β-catenin bound regions. DOX –/+ represents conditions without/with doxycyclin induction of dnTCF4.

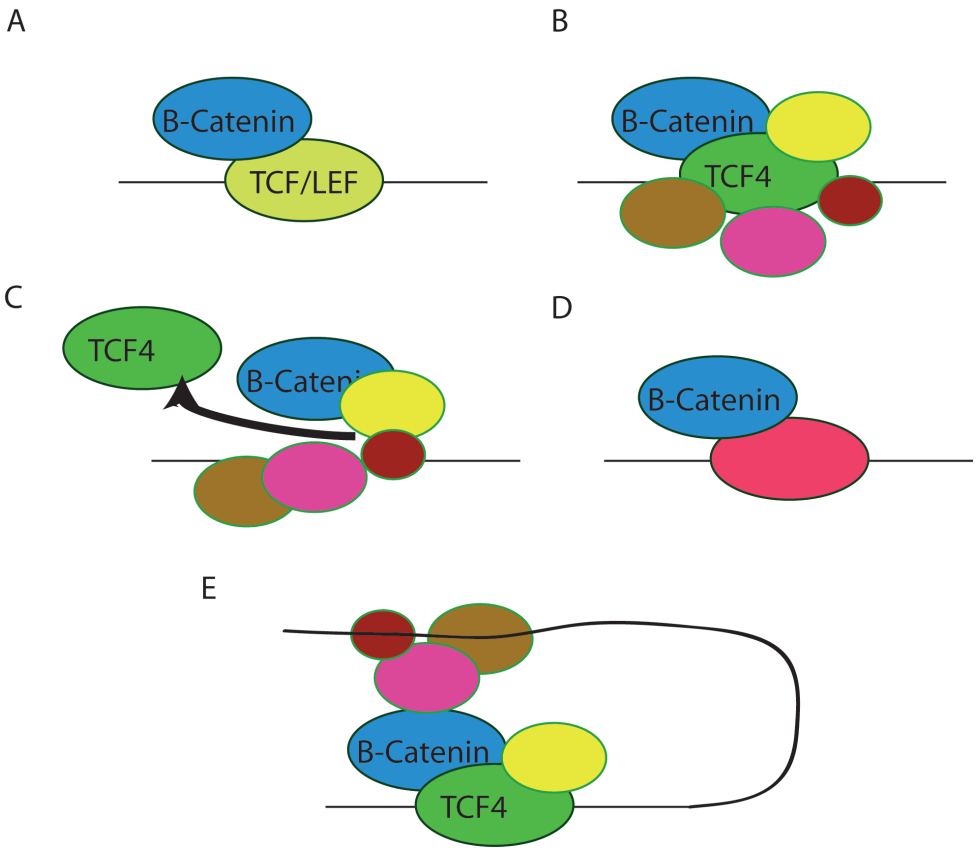


Figure 7: Models explaining TIBR sites. A) β-catenin is recruited via TCF/LEF1 member other than β-catenin. B) β-catenin is recruited via TCF4, while recruitment of other factors prevents efficient immunoprecipitation of TCF4. C) β-catenin is recruited via TCF4 together with other transcription factors. After stabilization of the protein complex, TCF4 is released. D) β-catenin is recruited via other recruiter than TCF/LEF1. E) β-catenin is recruited via TCF4 and afterwards loops to another genomic location that co-immunoprecipitates together with the original binding TCF-mediated site.

β-catenin, and is more likely associated with less active chromatin. However, we provide evidence that for both classes β-catenin recruitment is mediated through a TCF/LEF member, indicating that there are no alternative routes of β-catenin recruitement in the systems studied here.

MATERIALS AND METHODS

ChIP-seq

Chromatin immunoprecipitation and sequencing was performed on Ls174T-L8 cells grown in the presence or absence of doxycycline (1μg/ml) for 12 hours and HEK 293T cells grown with or without presence of Wnt3a for 12 hours as described previously [16]. In brief, approximately $30x10^6$ cells per IP were crosslinked with 1% formaldehyde for 20 minutes at room temperature. For β-catenin ChIP-seq, cells were crosslinked for 40 minutes using ethylene glycol-bis(succinimidylsuccinate) (Thermo scientific) at 12.5mM final concentration, with addition of formaldehyde (1% final concentration) after 20 minutes of incubation. The reaction was quenched with glycine and the cells were washed with phosphate buffered saline, buffer B (0.25% Triton-X 100, 10 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and buffer C (0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]). The cells were then resuspended in ChIP incubation buffer (0.3% sodium dodecyl sulfate [SDS], 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) and sheared using Covaris S2 (Covaris) for 8 minutes with the following settings: duty cycle: max, intensity: max, cycles/burst: max, mode: Power Tracking. The sonicated chromatin was diluted to 0.15 SDS, incubated for 12 h at 4°C with appropriate antibody with 100 μl of protein G beads (Upstate). The beads were successively washed 2 times with buffer 1 (0.1% SDS, 0.1% deoxycholate, 1% Triton-X 100, 0.15 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 2 (0.1% SDS, 0.1% sodium deoxycholate, 1% Triton-X 100, 0.5 M NaCl, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), one time with buffer 3 (0.25 M LiCl, 0.5% sodium deoxycholate, 0.5% NP-40, 1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]), and two times with buffer 4 (1 mM EDTA, 0.5 mM EGTA, 20 mM HEPES [pH 7.6]) for 5 min each at 4°C. Chromatin was eluted by incubation of the beads with elution buffer (1% SDS, 0.1 M $NaHCO_3$), the eluted fraction was reconstituted to 0.15% SDS with ChIP incubation buffer and the immunoprecipitation repeated for second time with half the amount of antibody. After washing and elution, the immunoprecipitated chromatin was de-cross-linked by incubation at 65°C for 5 h in the presence of 200 mM NaCl, extracted with phenol-chloroform, and ethanol precipitated. Immunoprecipitated chromatin was additionally sheared, end-repaired, sequencing adaptors were ligated and the library was amplified by ligation mediated PCR (LMPCR). After LMPCR, the library was purified and checked for the proper size range and for the absence of adaptor dimers on a 2.2% agarose gel and sequenced on SOLiD/AB sequencer to produce 50 basepair long reads. Sequencing reads were mapped against the reference genome (hg18, NCBI built 36 for HEK 293T and hg19, NCBI build 37 for Ls174T-L8) using the MAQ or BWA package [19, 20]. Reads mapped to multiple positions were excluded from further analysis. Cisgenome software package [21] was used for the identification of binding peaks from the ChIP-seq data and identification of most over-represented sequence motifs in binding peaks.

Data files from ENCODE project

ChIP-seq data of 9 histone marks, CTCF and coordinates of different chromatin types in HEK 293T cells [17] were downloaded from http://genome.ucsc.edu/ENCODE/downloads.html website [22].

Identification of TIBR sites and meta-gene analysis

Regions from -1000 to +1000 base pairs from centers of β-catenin peaks were aligned around peak center and divided into 50bp bins. Read density profiles per bin were normalized for sequencing depth and antibody and were clustered using k-means clustering with R kmeans() command (centers=8 settings), according density profiles from TCF4 ChIP-seq. Cluster of β-catenin peaks with lowest density of TCF4 tags was defined as TIBR sites.

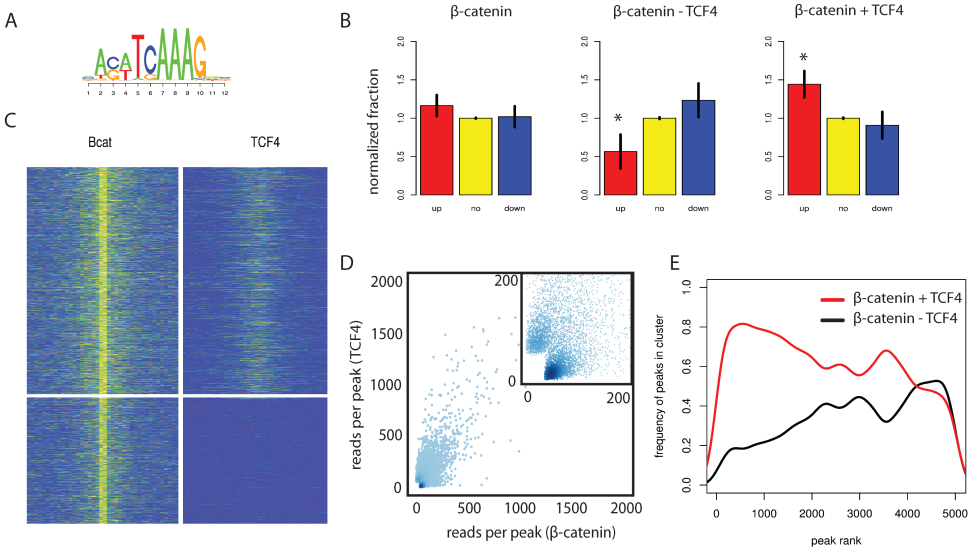
Reporter assays

Reporter assays for TIBR sites were performed as described previously [15]. In brief, typically 1kb long fragments containing candidate TIBR site was cloned in front of firefly luciferase gene in pGL4.10 in both directions. The reporters were

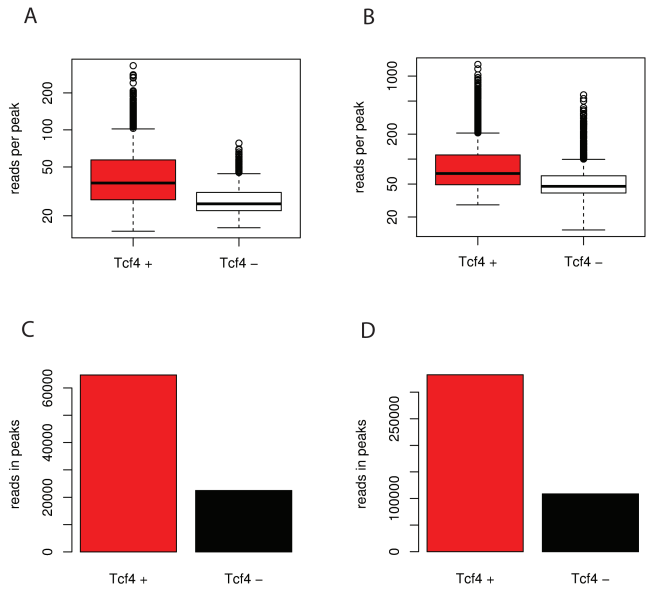
transfected into Ls174T-L8 cells with *Renilla* luciferase as transfection control and their activity was measured using a dual-luciferase reporter assay system (Promega).

REFERENCES

1. Clevers, H., *Wnt/beta-catenin signaling in development and disease.* Cell, 2006. **127**(3): p. 469-80.
2. Le, N.H., P. Franken, and R. Fodde, *Tumour-stroma interactions in colorectal cancer: converging on beta-catenin activation and cancer stemness.* Br J Cancer, 2008. **98**(12): p. 1886-93.
3. Korinek, V., et al., *Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma.* Science, 1997. **275**(5307): p. 1784-7.
4. Molenaar, M., et al., *XTcf-3 transcription factor mediates beta-catenin-induced axis formation in Xenopus embryos.* Cell, 1996. **86**(3): p. 391-9.
5. Behrens, J., et al., *Functional interaction of beta-catenin with the transcription factor LEF-1.* Nature, 1996. **382**(6592): p. 638-42.
6. Botrugno, O.A., et al., *Synergy between LRH-1 and beta-catenin induces G1 cyclin-mediated cell proliferation.* Mol Cell, 2004. **15**(4): p. 499-509.
7. Mosimann, C., G. Hausmann, and K. Basler, *Beta-catenin hits chromatin: regulation of Wnt target gene activation.* Nat Rev Mol Cell Biol, 2009. **10**(4): p. 276-86.
8. Mulholland, D.J., et al., *Interaction of nuclear receptors with the Wnt/beta-catenin/Tcf signaling axis: Wnt you like to know?* Endocr Rev, 2005. **26**(7): p. 898-915.
9. Townsley, F.M., A. Cliffe, and M. Bienz, *Pygopus and Legless target Armadillo/beta-catenin to the nucleus to enable its transcriptional co-activator function.* Nat Cell Biol, 2004. **6**(7): p. 626-33.
10. de la Roche, M., J. Worm, and M. Bienz, *The function of BCL9 in Wnt/beta-catenin signaling and colorectal cancer cells.* BMC Cancer, 2008. **8**: p. 199.
11. Billin, A.N., H. Thirlwell, and D.E. Ayer, *Beta-catenin-histone deacetylase interactions regulate the transition of LEF1 from a transcriptional repressor to an activator.* Mol Cell Biol, 2000. **20**(18): p. 6882-90.
12. Kitagawa, H., et al., *A regulatory circuit mediating convergence between Nurr1 transcriptional regulation and Wnt signaling.* Mol Cell Biol, 2007. **27**(21): p. 7486-96.
13. Kelly, K.F., et al., *beta-Catenin Enhances Oct-4 Activity and Reinforces Pluripotency through a TCF-Independent Mechanism.* Cell Stem Cell, 2011. **8**(2): p. 214-27.
14. van de Wetering, M., et al., *The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells.* Cell, 2002. **111**(2): p. 241-50.
15. Hatzis, P., et al., *Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells.* Mol Cell Biol, 2008. **28**(8): p. 2732-44.
16. Mokry, M., et al., *Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles.* PLoS One, 2010. **5**(11): p. e15092.
17. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
18. Li, J. and C.Y. Wang, *TBL1-TBLR1 and beta-catenin recruit each other to Wnt target-gene promoter for transcription activation and oncogenesis.* Nat Cell Biol, 2008. **10**(2): p. 160-9.
19. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
20. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
21. Ji, H., et al., *An integrated software system for analyzing ChIP-chip and ChIP-seq data.* Nat Biotechnol, 2008. **26**(11): p. 1293-300.
22. Myers, R.M., et al., *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.



Supplementary Figure 1: Two distinct classes of beta catenin bound regions in HRK 293T cells. A) One of two most overrepresented sequence motifs found in β-catenin peaks is similar to the canonical TCF4 binding motif. B) β-catenin sites without significant presence of TCF4 (TIBR) are more enriched in down-regulated genes compared to β-catenin/TCF4 sites, which are enriched in up-regulated genes. Error bars represent standard deviation of 10,000 randomized datasets. *p<0.05 C) Heatmap presenting results of k-means clustering of β-catenin bound regions according to the presence of TCF4. D) Occupation of sites bound by either TCF4 or β-catenin as determined by ChIP-seq. E) Representation of different peak ranks among two β-catenin peak classes.



Supplementary Figure 2: Quantification of bound β-catenin. Quantitative distribution of sequencing tags forming TCF4 positive and negative β-catenin bound regions in A) Ls174T-L8 and B) HEK 293T cells. Total numbers of sequencing tags forming TCF4 positive and negative β-catenin bound regions in C) Ls174T-L8 and D) HEK 293T cells.

9

CONCLUSIONS AND PERSPECTIVES

CONCLUSIONS AND PERSPECTIVES

In this thesis, several new improvements in next generation sequencing technologies are shown. These improvements opened new horizons in targeted resequencing and allowed us to relieve biological principles behind the transcriptional regulation.

There are only few examples of scientific technologies that evolved so dramatically in such a short time period as targeted resequencing. Nowadays, together with information from linkage analysis and homozygosity mapping, targeted resequencing represents a standard tool for the identification of disease causing variants in Mendelian inherited disorders (1-3). The past two years have resulted in a rapidly growing list of genes linked to particular diseases. In addition, large-scale patient sequencing projects have provided a novel view on some complex pathological states like mental retardation (4), where examples of de novo monogenic causes were identified. Apart from research applications, targeted resequencing also starts entering the field of routine clinical diagnostics and most likely will serve as a ground-base of personalized medical care before it becomes obsolete due to cheaper and more complete whole genome sequencing. All of this has happened within four years since the first (and maybe even preliminary) proof of principle in back to back reports in 2007 (5-7). To illustrate the revolutionary possibilities of targeted NGS resequencing in clinical diagnostics, one can compare with the classical Sanger-sequencing based tests in which single genes are subsequently screened for genomic variants at the cost of several hundreds of Euros per gene. With NGS technology hundreds or even thousands of genes can be screened with the same sensitivity and price (8,9) as demonstrated in **chapter 2 and 3** of this thesis. This opens great possibilities for screening in inherited disease diagnostics in newborns or more precise cancer diagnostics with possibility for more targeted treatment. In addition, while implementing methods like multiplexed enrichment as introduced in **chapter 3 and 4** of this thesis (9), together with an expected increase in throughput of next generation sequencers the same price may be sufficient to sequence all genes in the genome in the upcoming year. However in the same way as knowing an alphabet and a few words or phrases does not mean that we understand the whole book, most of our efforts in upcoming decade(s) will have to be put into understanding the meaning and outcomes of millions of genomic variations that are present in every individual human genome. Only then true personalized medicine will become a reality. Comprehensive analysis and interpretation of all genomic variation will be required to predict disease susceptibility and treatment success. .Apart from medical research and clinical applications, targeted resequencing has also great potential to improve on basic research possibilities or efficiencies. Introduction of NGS-based techniques for the identification of genomic variant in reverse genetics screens in model organisms like zebrafish [unpublished data] or rat (9) as demonstrated in **chapter 3** or techniques for faster and less laborious identification of causative genes in forward screens (10) as shown in **chapter 5,** allow researchers to spend more time on studying the biology aspect of a particular model system rather than performing laborious mapping experiments.

During the four years after its introduction (11), another method, chromatin immunoprecipitation and sequencing (ChIP-seq), dramatically changed our way of studying biology, i.e. mechanisms behind the transcriptional regulation. With possibilities to produce genome wide maps of various histone marks and transcription factors in virtually any model system, the

major challenges in upcoming years would be in proper understanding or decoding of their combinatorial meaning. This might be a difficult and complex task due to the many known and maybe still several unknown histone modifications. Furthermore, the number of combinations and outcomes is very high with hundreds of different transcription factors that are simultaneously active in a single cell and may share or compete for the same regulatory regions. In addition, when moving from transcriptional regulation to regulation of gene expression in general other layers of control take place including the nuclear organization, splicing, small RNAs, RNA trafficking or regulation of RNA stability, making regulatory networks even more complex and more difficult to model and understand. Within this complex framework, single signaling event or activation of a single transcription factor may directly work on several regulatory layers that independently target different sets of genes (12,13)RNA polymerase II binding and steady state RNA levels identifies differentially regulated functional gene classes while triggering numerous other indirect events. As an example, transcription factors may serve as a repressor by directly binding to promoters of target genes and at the same time increases RNA levels of other set of genes by independent mechanisms (12).

To allow for more proper studying of mechanisms and principles behind transcriptional regulations more reliable and versatile tools to measure transcription rate need to be developed. Nowadays RNA levels and their changes measured by microarrays or RNA sequencing are often translated as a measure for transcription rate. However steady-state RNA levels often do not mirror rates how these transcripts are produced due to large differences in RNA stability. Other methods like GRO-seq and nuclear run-on require laborious experimental setups and do not allow easily for studying transcription in intact systems. To fill this gap, we introduced RNA Polymerase II (Pol II) ChIP-seq in **chapter 9** as a method for more directly studying transcriptional events as compared to steady state RNA sequencing or micro-arrays. In the future, this method may allow us to more precisely study direct transcriptional changes upon signal transduction changes. In addition, together with methods assaying RNA levels it will allow us to independently investigate separate mechanisms involved in maintaining RNA levels. For example, a combination of Pol II ChIP-seq and steady RNA level measurements from the same system could allow us to separately study changes in production of RNA and mechanisms responsible for RNA stability.

To further increase the repertoire of possible tools, we modified the way of preparing ChIP-seq libraries to achieve highly reproducible protocol especially for samples where only limited amounts of immunoprecipitated material is available, as described in **chapter 6**. This allowed us to produce genome wide binding maps of transcription factors which interact with DNA only indirectly like β-catenin and allowed us to uncover two different classed of β-catenin bound regions as described in **chapter 8**.

It is very likely that NGS techniques and applications will develop further in the near future. Systematic and comprehensive measurements of biological systems will thereby increasingly become reality providing important opportunities for systems biological approaches. However, one should often realize that every independent cell is a system by its own, indicating that single cell and single molecule measurements will become increasingly important to properly understand individual molecular processes and their joint contribution to the characteristics of a cell, a tissue and the individual.

REFERENCES

1. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461, 272-276.
2. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A. et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet, 42, 30-35.
3. Walsh, T., Shahin, H., Elkan-Miller, T., Lee, M.K., Thornton, A.M., Roeb, W., Abu Rayyan, A., Loulus, S., Avraham, K.B., King, M.C. et al. (2010) Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am J Hum Genet, 87, 90-94.
4. Vissers, L.E., de Ligt, J., Gilissen, C., Janssen, I., Steehouwer, M., de Vries, P., van Lier, B., Arts, P., Wieskamp, N., del Rosario, M. et al. (2010) A de novo paradigm for mental retardation. Nat Genet, 42, 1109-1112.
5. Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J. et al. (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods, 4, 903-905.
6. Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J. et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet, 39, 1522-1527.
7. Okou, D.T., Steinberg, K.M., Middle, C., Cutler, D.J., Albert, T.J. and Zwick, M.E. (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods, 4, 907-909.
8. Mokry, M., Feitsma, H., Nijman, I.J., de Bruijn, E., van der Zaag, P.J., Guryev, V. and Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res, 38, e116.
9. Nijman, I.J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E. and Cuppen, E. (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nat Methods, 7, 913-915.
10. Mokry, M., Nijman, I.J., van Dijken, A., Benjamins, R., Heidstra, R., Scheres, B. and Cuppen, E. (2011) Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach. BMC Genomics, 12, 256.
11. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science, 316, 1497-1502.
12. Westendorp. B, Mokry, M., Koerkamp, M.G., Holstege, F.C., Cuppen, E. and de Bruin, A. (submitted) E2F7 represses a transcriptional network of oscillating cell cycle genes to control S-phase progression.
13. Mokry, M., Hatzis, P., Schuijers, J., Lansu, N., Ruzius, F.P., Clevers, H. and Cuppen, E. (2011) Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady state RNA levels identifies differentially regulated functional gene classes. Nucleic Acids Research, (in press).

&

SAMENVATTING
ACKNOWLEDGEMENTS
CURRICULUM VITAE
LIST OF PUBLICATIONS

SAMENVATTING

In dit proefschrift worden meerdere verbeteringen in verschillende toepassingen van de volgende generatie sequensen beschreven. Deze verbeteringen maakt het mogelijk om gericht te hersequencen en maakte het mogelijk om de biologische principes achter trancriptionale regulatie verder te begrijpen.

In hoofdstuk 2 hebben we laten zien dat door het gebruik maken van kleinere insertie sequens banken en langere sequenses de verrijking van specifieke genomische delen kon worden vergroot. De belangrijkste verbetering van deze aanpak was niet alleen dat we, voor dat moment, de beste verrijking hadden, maar de meer homogene distributie van sequenses over de verrijkte regio's en de positieve en negatieve DNA strengen. Door deze gelijke distributie van sequenses over de regio's van interesse waren wij in staat om varianties waar te nemen met een zeer lage vals positieve en een lage vals negatieve ratio.

Hoofdstuk drie beschrijft een nieuwe methode om van meerdere monsters specifieke regios van het genoom te hersequensen. Door het hersequensen van 770 genen van 30 N-ethyl-N-nitrosourea- gemuteerde ratten, lieten we zien dat we meerdere monsters samen konden voegen in een experiment voor de genomische verrijking. Voor deze vinding was het nodig om elk monster apart te verrijken voor het sequensen. Door het gebruiken van deze methode, konden wij de prijs van gericht hersequensen erg drukken, voornamelijk in studies waar veel monsters gebruikt worden en de kandidaat genen bekend zijn.

In hoofdstuk 4 presenteren we het reeds gepubliceerde protocol om van meerdere monsters tegelijkertijd genomisch te verrijken en te sequensen. Door middel van dit protocol, laten wij de toepasbaarheid van onze methode op een brede reeks toepassing zien, beginnend bij de genomische verrijking en hersequensen van 96 patiënten in een enkel experiment voor een beperkte hoeveelheid genen tot het volledige humane exome.

In hoofdstuk 5 introduceren we een "fast forward genetics" aanpak voor het opsporen van potentiele variaties in forward genetic experimenten. Deze techniek combineert de traditionele bulk segregant techniek met genomische verrijking en de volgende generatie sequensen .Wij hebben deze twee stappen methode succesvol toegepast op twee Arabidopsis mutanten en identificeerde op deze manier een nieuwe factor die nodig is voor stam cel activiteit.

In hoofdstuk 6 laten we zien dat het verder doen afknappen van ge-mmunoprecipiteerd DNA een voordeel oplevert bij Chip-Seq. Door het toepassen van deze methode waren wij in staat sequence banken te generen van hoeveelheden die normaal gesproken niet toereikend waren door het gebruik maken van DNA fragmenten die normaal gesproken te lang waren voor sequencing.

In hoofdstuk 7 introduceren we RNA polymerase II (polII) Chip-Seq als een betrouwbare en veelzijdige manier op gen activiteit te bepalen. Door het integreren van Pol II chip-seq data, micro-array en RNA-seq gegevens, stelde ons in staat om functionele gen groepen te definiëren die op verschillende manieren worden gereguleerd.

In hoofdstuk 8, gebruik makend van de in hoofdstuk 6 beschreven techniek, waren we in staat om reproduceerbaar chip-seq van beta-catenin en TCF/LEF familie leden te doen en identificeerde we twee klassen van beta-catenin gebonden genomische elementen.

ACKNOWLEDGEMENTS

I can't believe that my thesis is finished!!! Indeed this was only possible thanks to the contribution and help of many people.

Edwin, thank you for everything. You strongly influenced the way how I perform research and my perception of science, both practically and philosophically. During those four years you not only helped me to develop scientifically but also personally. Thanks for the chance you gave me, when you answered to my email and allowed me to join your lab for a short internship, which led to a PhD position in your lab. I am looking forward to our future collaborations!

Joram, you have been the first person I talked to in Netherlands and my first daily supervisor. I still remember that evening when you and guys from lab took me to Winkel van Sinkel to watch football and even though you did not score against Argentina, you definitely scored against Slovakia in beer drinking. Thanks Marco..., you know what for ;).

Harma, Ruben, Sam and Jos, my old PhD mates. Thanks for lot of fun and discussions during those years. The youngsters - Sebas and Roel thanks for the enthusiasm you brought into our PhD-room.

Ewart, I learned a lot from you. I'll always remember the fun and talks we had during our trips to San Francisco and Boston.

Nico, together with Ewart you managed to tame the wild blue beast in sequencing room, making the life of us who analyze the deep sequencing data much easier. Thanks for that.

Of course I can't forget the inhabitants of the Nerds Lair. Victor and Ies, thanks for introducing me into the world of Linux and Perl. Elzo, probably I would not jump into R without your example. Marieke, thanks for discussions and advices on how to do the data analysis. Long Frans, Sander, Slavik, thanks for all the help. I still time to time (mis)use the scripts you wrote.

Henk, Pim, Mark, Jose and Esther and the whole Cuppen group, thanks for help and support when it was needed.

I supervised only two students Veronika and Jarka. But quality goes above quantity! Veronika, good luck with your second masters and I hope to see you around again.

Pantelis and Jurian, thank you for introducing me into completely new field, which at the end forms half of my thesis and allowed me to collaborate with many fantastic people. Pantelis, good luck with your own lab. Jurian, good luck with your thesis.

Lucas, thanks for all the discussions and reminding me to use more stairs instead of elevator. And not to forget... did you know that there is Nature Climate Change journal and PloS Neglected Tropical Diseases journal? I think it seriously interferes with our plans.

Harma and Pieter, collaboration with you resulted in nice publication and served as a cornerstone for some of the others. Thanks for that.

Anja, Rene, Renze and Ben, I would never expect that a plant will play central role in one chapter of my thesis. Thank you for nice collaboration.

Bart, Astrid, Sylvia, Joost, Frank and Sjoerd. Thank you for giving me the opportunity to be part of your projects. I know that sometimes it took me longer than I promised to analyze the data or I mixed some adaptors ;), but I hope that I helped a bit.

Rodičia a starí rodičia, ďakujem Vám za Vašu lásku a podporu, ktorú stále pociťujem, aj keď len na diaľku.

I would have never imagined that one of my chapters would be written together with my wife. Magdi, nie každý má šťastie mať pri sebe niekoho, kto ho chápe a podporuje tak ako ty mňa. Ďakujem za všetko a teším sa na chvíle, ktoré ešte len prídu.

I did not realize how wonderful experience it is be to write the acknowledgment part. Many nice moments from those 4 years just suddenly revoked in my mind - making me feel like I experience it again. Thank you all again for those great four years.

Michal

CURRICULUM VITAE

Michal Mokry was born on 7th December 1982 in Košice, Slovakia. He finished secondary education at the Gymnasium Srobarova 1 with an emphasis on the natural sciences in 2001. In the same year, he entered P.J. Safarik University in Kosice, Faculty of Medicine, where he voluntarily participated in numerous research projects. For his research work he was awarded "Slovak Student Personality of the Year Award" in medical sciences and pharmacy. He received his medical degree in 2007 with the Dean's Prize. In summer 2007 he joined the Hubrecht Institute in Utrecht to begin his PhD research at the laboratory of prof. Edwin Cuppen. The results of this research are described in this thesis. Since July 2011 he works in the laboratory of prof. Edward Nieuwenhius at the ~~University Medical Center~~ Utrecht.

PUBLICATION LIST

Publications related to PhD study

Westendorp B, **Mokry M**, Koerkamp MG ,Holstege F, Cuppen E, de Bruin A (2011) E2F7 represses a transcriptional network of oscillating cell cycle genes to control S-phase progression (submitted)

Boj SF, vEs JH, Li V, Hatzis P, **Mokry M**, Huch M, Haegebarth A, vdBorn M, Voshol P, Dor Y, Chambon P, Cuppen E, Clevers H (2011) The Wnt effector TCF4 encoded by diabetes risk gene Tcf7l2 controls liver metabolism during birth and fasting (submitted)

Koval V, **Mokry M**, Cuppen E, Guryev V (2011) DWAC-seq — Dynamic Window Approach for CNV detection using next-generation sequencing tag density (in preparation)

Vermaat JS, Nijman IJ, Koudijs M, Gerritse FL, Scherer SJ, **Mokry M**, Roessingh W, Lansu N, de Bruijn E, van Hillegersberg R, van Diest PJ, Cuppen E, Voest EE (2011) Primary colorectal cancers and their metastases are genetically different: implications for selection of patients for targeted treatment (submitted)

*Harakalova M, ***Mokry M**, Hrdlickova B, de Bruijn E, Nijman IJ, Renkens I, Duran K, van Roekel H, Kloosterman W, Cuppen E (2011) Multiplexed genomic enrichment for flexible targeted next-generation sequencing. Nature Protocols,~~(in press)~~
*equal contribution

***Mokry M**, *Hatzis P, *Schuijers J, Lansu N, Ruzius FP, Clevers H and Cuppen E (2011) Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady state RNA levels identifies differentially regulated functional gene classes. Nucleic Acids Research 2011, doi: 10.1093/nar/gkr720.
*equal contribution

***Mokry M**, *Nijman IJ, van Dijken A, Benjamins R, Heidstra R, Scheres B, Cuppen E (2011) Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach BMC Genomics 12, 256
*equal contribution

Harakalova M, Nijman IJ, Medic J, **Mokry M**, Renkens I, Blankensteijn JD, Kloosterman W, Baas AF, Cuppen E (2011) Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm (AAA) regions of interest - challenges and limitations. J of Cardiovasc Trans Res 4, 271–280

Mokry M, Hatzis P, de Bruijn E, Koster J, Versteeg R, Schuijers J, van de Wetering M, Guryev V, Clevers H, Cuppen E. (2010) Efficient double fragmentation ChIP-seq provides nucleotide resolution protein-DNA binding profiles. PLoS One. 5(11), e15092.

*Nijman IJ, ***Mokry M**, *van Boxtel R, Toonen P, de Bruijn E, Cuppen E (2010) Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. Nature Methods. 7(11),913-915.
*equal contribution

Johannes F, Wardenaar R, Colome-Tatche M, Mousson F, de Graaf P, **Mokry M**, Guryev V, Timmers MHTh, Cuppen E, Jansen RC (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq, Bioinformatics, 26(8):1000-1006

Mokry M, Feitsma H, Nijman I.J, de Bruijn E, van der Zaag PJ, Guryev V, Cuppen, E. (2010) Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res. 36(10), e116

Mokry M, Cuppen E (2008) The Atp1a1 gene from inbred Dahl salt sensitive rats does not contain the A1079T missense transversion. Hypertension, 51, 922-927.

Other publications

Gal P, Kravcukova P, **Mokry M**, and Kluchova D, (2009) Chemokines as possible targets in modulation of the secondary damage after acute spinal cord injury: a review. Cell Mol Neurobiol, 29, 1025-1035

Gal P, **Mokry M**, Vidinsky B, Kilik R, Depta F, Harakalova M, Longauer F, Mozes S, Sabo J (2009) Effect of equal daily doses achieved by different power densities of low-level laser therapy at 635 nm on open skin wound healing in normal and corticosteroid-treated rats. Lasers Med Sci, 24, 539-547.

Mokry M, Joppa P, Slaba E, Zidzik J, Habalova V, Kluchova Z, Micietova L, Rozborilova E, Salagovic J, Tkacova R (2008) Beta2-adrenergic receptor haplotype and bronchodilator response to salbutamol in patients with acute exacerbations of COPD. Med Sci Monit, 14, CR392-398.

Gal P, Toporcer T, Vidinsky B, **Mokry M**, Grendel T, Novotny M, Sokolsky J, Bobrov N, Toporcerova S, Sabo J (2008) Postsurgical administration of estradiol benzoate decreases tensile strength of healing skin wounds in ovariectomized rats. J Surg Res, 147, 117-122.

Kilik R, Bober J, Gal P, Vidinsky B, **Mokry M**, Longauer F, Sabo J (2007) [The influence of laser irradiation with different power densities on incisional wound healing in healthy and diabetic rats]. Rozhl Chir, 86, 384-387 [article in Slovak]

Mokry M, Gal P, Harakalova M, Hutnanova Z, Kusnir J, Mozes S, Sabo J (2007) Experimental study on predicting skin flap necrosis by fluorescence in the FAD and NADH bands during surgery. Photochem Photobiol, 83, 1193-1196.

Gal P, Toporcer T, Vidinsky B, **Mokry M**, Novotny M, Kilik R, Smetana K Jr., Gal T, Sabo J (2006) Early changes in the tensile strength and morphology of primary sutured skin wounds in rats. Folia Biol (Praha), 52, 109-115.

Mokrý M (2006), [Current possibilities of intraoperative malign glioma demarcation using fluorescence methods]. Cesk Slov Neurol N, 102, 426-430 [article in Slovak]

Mokry M, Gal P, Vidinsky B, Kusnir J, Dubayova K, Mozes S, Sabo J (2006) In vivo monitoring the changes of interstitial pH and FAD/NADH ratio by fluorescence spectroscopy in healing skin wounds. Photochem Photobiol, 82, 793-797

Gal P, Vidinsky B, Toporcer T, **Mokry M**, Mozes S, Longauer F, Sabo J (2006) Histological assessment of the effect of laser irradiation on skin wound healing in rats. Photomed Laser Surg, 24, 480-488

Mokrý M, Gál P, Novotný M, Kušnír J, Dubayová K, Sabo J (2006) [Fluorescence spectra as a possible marker of the process of wound healing.] Cesk Slov Derm, 81, 277-281 [article in Slovak]

Gál P, Toporcer T, Vidinský B, Grendel T, **Mokrý M**, Kilík R, Bober J, Sabo J (2005) Comparison of Wound Tensile Strength Using Two Different Types of Skin Sutures. A Mech Slov, 2-A, 9, 57–62

Mokrý M, Kušnír J, Gál P, Dubayová K, Synek M, Sabo J (2005) Modification of the Optical Probe of the Perkin-Elmer LS 55 Luminescent Spectrometer for Measurement of Fluorescence Spectra of Skin Surface. Chem Listy, 2005, 99, 653-656 [article in Slovak]