Next generation sequencing in functional genomics

Thomas Werner

Submitted: 8th March 2010; Received (in revised form): 5th May 2010

Abstract
Genome-wide sequencing has enabled modern biomedical research to relate more and more events in healthy as well as disease-affected cells and tissues to the genomic sequence. Now next generation sequencing (NGS) extends that reach into multiple almost complete genomes of the same species, revealing more and more details about how individual genomes as well as individual aspects of their regulation differ from each other. The inclusion of NGS-based transcriptome sequencing, chromatin-immunoprecipitation (ChIP) of transcription factor binding and epigenetic analyses (usually based on DNA methylation or histone modification ChIP) completes the picture with unprecedented resolution enabling the detection of even subtle differences such as alternative splicing of individual exons. Functional genomics aims at the elucidation of the molecular basis of biological functions and requires analyses that go far beyond the primary analysis of the reads such as mapping to a genome template sequence. The various and complex interactions between the genome, gene products and metabolites define biological function, which necessitates inclusion of results other than sequence tags in quite elaborative approaches. However, the extra efforts pay off in revealing mechanisms as well as providing the foundation for new strategies in systems biology and personalized medicine. This review emphasizes the particular contribution NGS-based technologies make to functional genomics research with a special focus on gene regulation by transcription factor binding sites.

Keywords: pathways; networks; integrated analysis; ChIP-Seq; RNA-Seq

INTRODUCTION
Sanger introduced his revolutionary method of polymerase-based sequencing in the early 1970s, which due to various technical improvements finally led to the complete sequencing of the human and many other genomes. Two seminal papers originating from the publicly funded Human Genome Sequencing project led by Francis Collins [1] and the parallel (if not dependent) commercial approach of Celera [2] marked the change from sporadic to systematic genome-wide sequencing. The first genome sequence drafts were called so correctly as they of still consisted of thousands of sequence bits and pieces—not contiguous sequence (termed a contig) of chromosomes. Consequently, a better draft, a more complete draft, a real complete draft and a final draft (only to be followed by the next refined draft) of the human and many other genomes followed during the last decade of the last century (or as a matter of taste: millennium). Two important new concepts emerged from this development: The first is shifting the focus from isolated parts to analyzing genes, RNAs and proteins in context. cDNA projects had only provided isolated part lists, the cDNAs and the proteins derived from them. The genomic sequence, taking finally the shape of more or less complete chromosomal contigs, now revealed a much more detailed blueprint of whole genomes and thus the physical context of genes and transcripts. The second important principle, reinforced rather than newly introduced, was the observation that the same physical items, such as genes, transcripts and proteins, can vary depending on functional conditions (e.g. tissue- or condition-specific transcription or protein-modifications). However, the blueprint provided by the genome sequence is not sufficient
to reveal how genes actually functions, requiring additional analyses and data to make the transition from physical context to functional context.

For example, the ever dwindling number of human genes (from first estimates of about 100,000 we are now down to about 23,000, http://www.ensembl.org/Homo_sapiens/Info/StatsTable) coming dangerously close to the numbers found for worms (20,000) and flies (14,000) raised a lot of questions, how the obvious differences between those organisms and humans could be accommodated by genomes of such similar complexity.

It is common knowledge now that the difference is not directly correlated to the number of genes but rather to the complexity of regulation of the expression of these genes. The number of different variants (alternative transcripts and proteins including post-translational modifications) derived from the genes of an organism is what really correlates with the functional complexity. In other words, the most important part is not the blueprint of the genome, but the way its components interact, which is the central theme of functional genomics.

**WHAT IS FUNCTIONAL GENOMICS?**

Functional genomics aims at turning the huge amount of data obtained by observation and experiments into knowledge about life and life functions, albeit with a focus on genomic sequences as an anchor point. This also explains why the genomic sequence is a natural coordinate system for almost all functional genomics approaches and why the completion of so many genome sequencing projects fuels more and more projects in the area of functional genomics.

However, even for humans genomes are not identical and there are probably as many distinct human genomes as humans [3]. Genomic differences such as single nucleotide polymorphisms (SNPs) can result in functional differences by changing particular entities (e.g. protein coding parts) or affecting gene regulation (regulatory SNPs). Part of this question is now being addressed in the 1000 genome project, which basically was only made possible by the next generation sequencing (NGS) technology (http://www.1000genomes.org/page.php). Before coming to this in more detail, there are a few other points that need to be considered in order to realize the advantages and shortcomings of the new technology in the right context.

As already mentioned above, genomes and the differences between them only become important where such differences result in tangible effects. For example, as long as a woman does not have breast cancer, it is not clear whether she will get a Neu+ or Neu− genotype in case she develops breast cancer. However, in case a herceptin treatment for breast cancer is considered this makes all the difference between responders and non-responders [4]. In short, genotypic differences need to be seen in a functional context, and do not provide conclusive information on their own or in isolation. If we look a bit closer at cancer and genomics there are almost invariably genotypic changes in tumors versus the healthy tissues of the same patient that are crucial for the tumor phenotype [5]. Thus, even the individual genome sequence of the patient would not be enough but comparison of the healthy and the tumor-derived genomes is required. All in all, we need to look at many genomes from one species. If the aim is to elucidate general features of an organism’s biology, a few genomes may be sufficient, in terms of medical application this may come to several genomes from even the same patient.

Many diseases, especially all cancer tumors are invariably correlated with changes in the regulation of the cell cycle, which to a large extent is transcriptional [6]. However, the genomic sequence alone does usually not reveal what turned the cell malign. Therefore, additional detailed information about the particular transcriptomes, transcription factor binding, proteomes and/or metabolomes will be required. Taken together with temporal and quantitative changes upon internal or external signals the emerging picture will help to explain the functional differences caused by the observed changes. NGS is well suited to provide all the primary data required (i.e. the sequences) for genomics, epigenetics, transcription factor binding and transcriptomics in virtually unlimited detail. However, this data is not answering the salient functional questions directly. A thorough down-stream analysis of functional connections and consequences is mandatory to close the gap between mere data (sequence tags) and functional changes observed on a higher level paralleling the requirements already known from microarray analyses [7–9].

The very first step on this path from data to knowledge is to put the raw sequence reads into the coordinate system of a reference genome via genome annotation establishing the physical context.
GENOME ANNOTATION—CHANGING REQUIREMENTS

Genome annotation is available from various genome browser systems (UCSC Genome Browser, NIH Entrez, EBI’s ENSEMBL, as well as many others) and denotes primary sequence features such as genes, transcripts, exons, introns, promoters, genomic repeats (at least some), SNPs, CpG islands and other features that can be directly linked to the nucleotide sequence. This will be referred to as ‘static’ annotation, as these are the features thought to be fixed, which—in functional terms—they are not.

The transcriptional (and subsequent translation) of the human CDKN2A gene illustrates this point quite well: the gene has two distinct promoters giving rise to two distinct, alternative transcripts resulting in two totally different proteins [10]. Almost none of the annotated features (promoters, transcripts, exons, introns and protein sequences) are fixed functionally (Figure 1). Depending on the circumstances the active promoter is in a different physical location, the transcripts share exons but also contain different exons, and due to a frameshift between the two transcripts the proteins sequences are entirely different. Therefore, the annotation of this gene has multiple elements annotated for the exact same nucleotide sequence. For example whether a particular stretch of DNA is an intron or an exon depends on the active promoter, i.e. this becomes a dynamic feature, which depends entirely on the functional context. The DNA sequence of the gene cannot determine which promoter is active in a given cell as the genomic sequence is basically invariable (exceptions are germ and immune cells). Only functional downstream analysis, e.g. combining the genomic sequence analysis with a ChIP- and/or transcriptome, and sometimes, even proteome analyses is required to identify functional connections [11–14]. Regulatory pathways and networks responsible for the selective promoter activation need to be elucidated in order to understand why a particular transcript is expressed at a given time point. NGS data can provide important clues to this end. However, the functional interpretation of epigenetic data (ChIP-Seq [15]), transcriptome and TF-binding data (TF-ChIP-Seq) as well as elucidation and understanding of the regulatory networks and pathways is required, which reaches beyond what NGS directly delivers [16]. A similar line of analysis is also required to differentiate SNPs of functional consequence from the myriad of mere sequence variations that have no apparent effect on ‘normal’ functions, which needs to be established in the right context [17]. All of this is the domain of functional downstream-analysis without which the multitude of NGS results amount to little more than data overflow and white noise.

Figure 1: Biological flexibility of genomic annotation. Genomics sequence, transcripts and encoded proteins for the CDKN2A gene. The two promoters are shown as yellow transparent boxes, exon in gray/green (gray = non-coding part, green = coding part) and introns are indicated as the central gray rod. The two proteins encoded are entirely different due to a frameshift introduced by the alternative splicing starting from the second promoter (indicated by different pictograms, not representing any particular structure).

FUNCTIONAL GENOMICS OF TRANSCRIPTIONAL REGULATION

The regulation of transcription is one of the most important parts of functional genomics and will be the focus of this article. Within transcriptional regulation the binding and interactions of transcription factors (TFs) to their respective transcription factor binding sites (TFBSs) is of central importance to the transcription of genomic information into RNAs. Most of the downstream analysis methods and results of functional TFBS analysis so far have been established using microarray data. However, after the mapping has been completed there is no more principal difference between microarray and NGS based data, except the unbiased larger volume and more accurate quantitative nature of NGS derived sequence data. Therefore, it is not relevant for this review if microarrays or NGS data were used.
The first step in the analysis of NGS or microarray data is to establish a reference to the genomic sequence, which in case of NGS data is done by mapping of the reads to a reference genome. The next step to gain first insights into connections is usually looking at pre-defined functional collections such as ontologies [18] or pathways [19–21]. Subsequently, one can work from that towards the networks linking genes and pathways involved in the functional subset under investigation, which inevitably includes regulatory networks governing gene transcription. The final leaves of regulatory networks are TFs on the functional side and TFBSs on the physical side. Such analyses include wholesome approaches based on kinetic modeling [22] as well as mechanistically motivated approaches aiming at networks based on common molecular organization of regulatory regions involved [9, 12]. For a recent review about reverse network engineering see [23].

Therefore, functional genomics cannot be limited to genomic sequences, but has to take several –omics data into account in order to elucidate the functional consequences of any sequence variation (SNP, InDels, chromosomal rearrangements) observed for a genomic sequence (Figure 2). Here one of the major obstacles is to harmonize data that have no clear genomic sequence link (such as proteins and literature) with sequence-anchored data such as SNPs, TFBSs, transcripts, etc. This has been accomplished already in several of the studies mentioned above and needs to be intensified in the future. Integrated downstream analysis necessitated the combination of various tools and data sources. The natural turnover of PhD students and post-docs in academic laboratories favors focused problem-oriented solutions over large-scale integration and maintenance of integrated downstream analysis systems. Therefore, such integrated systems are currently maintained and provided as commercial solutions.

In addition to the need for relatively broad integration, functional genomics in general as well as TFBS-based is also faced with the intrinsic flexibility of biological systems. While these are the features that enabled evolution and are crucial to the survival of individuals as well as whole species, they can complicate functional genomics research considerably. For example, the databases of pathways and interactions of biological components are currently almost entirely generic, i.e. summaries of the most common denominator from many instances. They represent the most conserved parts of such pathways, not necessarily what really happens in an individual case. Pathways can change between cell types, tissues, individuals and species in many details, which is not reflected in the one generic pathway in the database. A nice example is the small and relatively well defined network of alternative pathways around p53, p73 and p63 in the induction of apoptosis [24]. Depending on the situation (e.g. p53 mutation) alternative pathways can be activated. Another example is represented by iso-enzymes that are active in the very same metabolic pathway but are usually under different regulation. This effectively changes the behavior of the pathway upon external stimuli as nicely demonstrated in a recent study investigating the distinct transcriptional regulation of two isoforms of Acetyl-CoA carboxylase ACC1 and ACC2 both of which are controlled by more than one promoter (ACC1 by three, ACC2 by two, [25]).

Also as a system is disturbed, e.g. by an infection or a mutation, pathways and networks are not only activated differently as demonstrated in a recent study about influenza infection [26], but signaling networks may also rearrange in parts to accommodate the new situation as in the case of the RET mutations involved in the congenital abnormality Hirschsprung’s disease [27]. This is often observed with tumors, which is the reason they behave no longer as normal cells. We know of altered pathways in case of loss-of-function mutations (or experimental knockouts), which necessarily cause some rearrangements in the connected networks as well, resulting in a different wiring of pathways in tumor cells than in other cells of the same organism as demonstrated for the phosphatidylinositol-3-kinase pathway in human colon cancer cells [28]. Any analysis restricted to generic pathways will necessarily miss out on such individual specialties. The same is true for TFBS-based sequence analysis. TFs can bind to a variety of sequences, which is taken into account by using weight-matrix based descriptions rather than individual sequences, but very often TFs can actually also bind to TFBSs that primarily bind to different TFs. This kind of cross-binding abilities has been addressed by the matrix family concept established for MatInspector [29] but nowhere else.

Real biological phenomena do only show up after the functional connections are established between all the data observed. Therefore, combination of a powerful downstream analysis with NGS data generation is an ideal setup if not a pre-requisite to adapt
Primary Analysis

DNA-Seq

RNA-Seq

CNV

chromosomal InDels

SNP annotation
de novo SNP detection

alternative splicing sites

gene expression

databases

meta-analysis

Multiple correlations

network reconstruction

Primary Analysis

DNA-Seq

RNA-Seq

pair end reads

DNA-Seq

Read clustering

de novo assembly

Downstream Analysis

DNA-Seq

ChIP-Seq

RNA-Seq

GWAS

Haplotypes

Tumor types

TFBSs

databases

pathways

regulatory networks

Primary Analysis

DNA-Seq

RNA-Seq

pair end reads

Read clustering

de novo assembly

Downstream Analysis

DNA-Seq

ChIP-Seq

RNA-Seq

GWAS

Haplotypes

Tumor types

TFBSs

databases

pathways

regulatory networks

Meta Analysis

network reconstruction

multiple correlations

multiple experiments

Figure 2: Overview of NGS-based analysis strategies. Primary analysis: This part describes analyses steps that are based directly on the reads and are physically derived from sequence comparisons: CNV, chromosomal InDels are Insertions or Deletions (including translocations), SNP annotation regards SNPs already known and described (e.g. in dbSNP) while de novo SNP detection results from alleles detected via multiple sequencing coverage of the SNP position. Alternative splice sites can be detected via mapping to a splice junction library or by direct genomic alignment of exon spanning reads (when reads or clusters are longer than 50 nucleotides), new transcripts/loci are derived by direct mapping of novel exons and splice-overlaps. Downstream analysis: This part differs for the three major application areas: genomic DNA-seq is genomic resequencing employed in genome-wide-association-studies (GWAS), the definition of Haplotypes and tumor typing usually via tumor-specific chromosomal InDels. ChIP-Seq determines genome-wide patterns of modified chromatin such as histone methylation or acetylation status as well as binding regions for DNA-binding proteins, usually DNA-dependent RNA polymerases or TFs leading to the definition of patterns such as TFBSs. RNA-Seq determines the the genome-wide expression of known as well as unknown transcripts, which can be identified by mapping of the RNA-Seq sequence tags to the genome and the transcriptome. This will also identify most splice variants. All three strategies converge into the biology-oriented downstream analysis involving identification of pathways, cis-regulatory modules and regulatory networks, which also involves the integration of prior knowledge as depicted in the flanking database areas in addition to the experimental data. Finally, meta-analysis allows merging of several lines of evidence (NGS results and other results (e.g. proteomics, metabolomics, etc)) into a more complete description of the underlying biology, via network reconstruction, multiple correlations of various lines of evidence (such as histone modifications, pol II binding and transcription rates) and the cross examination of multiple experiments such as transcriptional profiles from several patients.
and apply our generic knowledge as represented in the databases of Mendelian diseases, literature, pathways and TFBSs to the actual situation encountered in a particular case or experiment.

Moving from general features that can be observed in any number of individuals towards subtler, in extreme individual, differences poses new problems. First of all, no method is perfect, i.e. not all differences observed between individual analyses are necessarily ‘real’ or biologically true. As the sample size diminishes so does the ability to differentiate mere experimental errors (e.g. during the preparation of probes) from ‘true’, biologically manifest differences. However, meta-analysis comes to the rescue here, in the form of ‘biological consistency’. The various -omics data we can use are not independent of each other. The various -omics data we can use are not independent of each other.

Yeast-2-hybrid methods to determine protein–protein interactions (ppi) have a very high false-positive rate [30]. Using biological consistency and additional data allows to distinguish between likely true and likely false positives: If two proteins are interacting in reality they should be present in the same cell or at least tissue (proteomics can show this). They also should be present at the same time, which should be also reflected in the co-presence of their respective transcripts, which RNA-Seq NGS or microarray analysis may show. If the RNAs are coexpressed, there is a good chance that the respective promoters are also coregulated, which could be seen by common cis-regulatory modules of TFBSs (genomics analysis). If all this consistent evidence can be found, the interaction is most likely true. If no other supporting evidence than the Yeast-2-hybrid data can be found, chances are this might be an artifact. The identification of new regulatory connections between podocyte-specific proteins, initially detected by computational promoter analysis, was proven experimentally following exactly this logic of merging of genomics, transcriptomics and proteomics lines of evidence [20] (Figure 3).

The same logic can be applied to pathways, networks, epigenetics and all other results that can be found. The more different lines of evidence can be incorporated into a consistent meta-analysis the more likely the results reflect the true situation. Of course

Figure 3: Verification of prediction by several lines of evidence. The analysis started with the gene encoding for a podocyte-specific protein (kidney-glomerulus specific cell type). Comparative analysis of the promoters of the ortholog genes from human, mouse and rat identified a regulatory TFBS framework that was evolutionarily conserved. Genome-wide searches with this framework identified additional genes where the same framework was also conserved, i.e. of which were ZO-1 and Cadherin-5. Microarray based transcriptome analyses of a patient collective verified the correlated expression of all three genes in humans, and subsequently immuno-gold staining located all three proteins as part of the tight-junction complex linking podocytes in glomeruli (published in [20]).
this is far away from the initial sequencing or other experimental data but this is the point where the real value of the depth of NGS analysis materializes.

Carrying out the complete range of analysis as described based on whole genome/epigenome/TF-ChIP/transcriptome/data as obtained from NGS requires several steps of powerful stratification in order to focus on subsets of data and intermediate results relevant for a particular question or situation.

One way to focus NGS analysis onto more relevant subsets of sequences rather than going indiscriminately genome-wide is to use pre-selection based on prior knowledge. Such approaches are most valid in looking at instances of well-established and clearly delimited systems, such as diagnostic sets. However, pre-selection inevitably introduces a bias towards prior knowledge, which is the basis for the selection. In case something really new, i.e. unknown, is involved, this will be missed by pre-selection while unbiased genome-wide analysis allows to find such surprises (also known as ‘black swans’). Of course, only if the analysis is powerful enough to reveal such unknown features. This is also true for the analysis of transcriptional regulation, where the use of pre-defined TFBSs patterns introduces a strong bias.

WHAT KIND OF DATA IS PROVIDED BY NGS FOR TFBSS-BASED ANALYSIS OF TRANSCRIPTIONAL REGULATION?

Current NGS approaches can be classified into four major categories:

(i) DNA-Seq: genome-based sequencing yielding genomic deletions and rearrangements, copy-number variations (CNV) of smaller regions or elements, and SNPs.

(ii) RNA-Seq: RNA-Sequencing, yielding genome-wide and quantitative information about transcribed regions (exons, and subsequently transcripts).

(iii) ChIP-Seq 1: TF-based ChIP, yielding genome-wide information about the physical binding sites of individual TFs to within a few hundred base pairs.

(iv) ChIP-Seq 2: Epigenetic ChIP (DNA methylation and/or histone modifications), yielding information about modifications and the accessibility of genomic regions to TFs and other factors. (for a recent review of the technology see [31]).

The sensitivity of all NGS applications is mainly a matter of sequencing depth (using more flow cells/beads will increase sensitivity); it reveals things as they are—totally unbiased by any expectations or prior knowledge. Give the several-fold complete coverage of all sequence templates present, this also yields a picture of unprecedented completeness and detail down to the level of SNPs and probably also RNA editing.

NGS can detect SNPs de novo even in the absence of a reference genome. Due to the high coverage alleles stand out clearly against sequencing errors already in the clustering and mapping procedure, which can actually be directly applied in population genetics [32]. We have conducted a comparative study of two families (parents + one daughter) using de novo SNP detection (Genomatix Software, Munich, Germany) and found that almost 99% of all common SNPs showed strict adherence to the Mendelian inheritance rule (Table 1).

ChIP-Seq yields unprecedented genome-wide resolution of protein binding sequences that can be delimited quite well directly from the sequencing data analysis bringing a new dimension of power to pattern definition algorithms [33, 34]. This allows
de novo binding site pattern definition as well as differentiation between likely promoter and likely enhancer binding sites.

RNA-Seq has several advantages over microarray analysis: (i) It is independent of prior knowledge, (ii) it features a bigger dynamic range, both on end of sparsely expressed transcript all the way to a much better quantification even of highly expressed transcripts [35], (iii) it identifies transcripts directly and covers all these transcripts over the entirety length rather than interrogating with a few pre-selected probes and (iv) it allows the identification of detailed transcript structures such as alternative promoters and alternative splicing events.

Of course there is always a trade-off between better resolution and the introduction of noise as all kinds of DNA and RNA fragments and degradation products will be also picked up that might be taken as ‘new’ binding sites or transcripts. Again, downstream analysis is the only way to collect additional evidence for the real TFBSs or transcripts differentiating them from unsupported fragments and artifacts.

And finally NGS is ideally suited to help us chart the 50% of the human (and other) genome(s) still not annotated, as each new NGS experiment defines new loci and transcripts in so far ‘intergenic’ regions, enriching at least the physical map of the genome immediately.

HOW CAN NGS DATA BE USED TO REVEAL TFBS-BASED TRANSCRIPTIONAL REGULATION?

Epigenetic events mainly control which parts of the genomic sequence are accessible by the transcriptional apparatus. Given a regulatory sequence is accessible, the time, amount, and duration of transcription of a gene is under the control of various specific (TF) and general transcription factors (GTF). GTFs are usually cofactors of the RNA polymerase complexes (This review will focus on RNA polymerase II mediated transcription), while the specific TFs represent ‘classical’ transcription factors such as NFkB, SP1 or AP1. Currently more than 760 specific TFs are known for the human genome alone (MatBase 8.2, Genomatix Software, Munich). The interaction of these TFs with their respective TFBSs in regulatory regions determines the major part of direct transcription control as they form the activator complexes on promoters and enhancers that subsequently attract the pol II complex, which in turn initiates transcription [36, 37].

NGS sequence data provide information about TFBSs in several ways. The most straightforward assay is ChIP-Seq carried out with TF-specific antibodies, which allows genome-wide determination of physical binding sites for the particular TF(s). ChIP-Seq usually yields region of several hundred base pair within which the most likely binding sites can be determined either by using pre-defined weight matrices (for known factors) or by de novo pattern determination in a set of such regions. For a recent review of ChIP-Seq and RNA-Seq analysis strategies see [38] and references therein. Methods to find ChIP-Seq clusters are established, general ([39], NGS-Analyzer, Genomatix Software, Munich) as well as applications specifically for paired-end ChIP-Seq data [40]. All programs define the clusters of reads that represent the binding regions. Subsequently, these clusters can be scanned by known TFBSs descriptions for which several programs are available (only two cited here, [29, 41]; there are many more). De novo pattern detection algorithms are more in line with the genome-wide unbiased approach as they also are capable of defining novel pattern without prior expectation. There are also many pattern detection algorithms available. However, most of these are not suitable to deal with thousands of regions simultaneously and require sub-selections before the can be applied. CoreSearch is a method first published in 1996 [42], which has been updated in the meantime to allow direct analysis of unlimited numbers of sequence regions (Genomatix Software, Munich). Once, a set of TFBSs has been established (either by known weight matrix scanning or de novo pattern detection) the really crucial step is to determine which of these TFBSs are functionally relevant. Binding of a TF to its TFBSs in genomic DNA does not necessarily have any regulatory consequences as this depends crucially on the context of this binding [43–46].

There are several ways how functional aspects of TFBSs can be computationally addressed, directly or indirectly. The most straightforward approach is to correlate TFBSs with the expression of corresponding transcripts (as assessed by RNA-Seq), which works quite well if the TFBSs are located in the promoters of the transcripts. However, in case of predominantly enhancer-binding TFs (such as many nuclear factors) this is complicated by the fact that
there is no simple link between the transcripts and enhancer TFBSs. Another way to assess the functional potential of TFBSs more indirectly is by phylogenetic comparison which also works in case of conserved enhancers. If TFBSs are conserved in the same location (e.g. the same promoter) in several species they are more likely to be functional than non-conserved TFBSs. However, as already mentioned, all of these correlation studies work best for TFBSs in context of other TFBSs, constituting cis-regulatory modules. In order to take advantage of cis-regulatory module structures, it is first required to determine the potential cis-regulatory modules computationally (for a recent review see [43]). We use a fully automatic approach to build multi-TFBSs framework models suitable to scan whole genome for matches [29] for this purpose.

RNA-Seq data can be used in a similar way to identify regulated transcripts and then scan the corresponding promoters for TFBSs establishing overrepresentations of TFBSs or TFBSs frameworks with the regulated transcripts (or subset thereof). Form thereon, the same strategy as with the ChIP-Seq data can be used.

Figure 4 summarizes the strategy for both TFBSs focused approaches, which can be used separately or by merging results in the integrate step. This is very similar to the scheme shown by Pepke et al. [38] with two major differences: (i) The emphasis in the analysis downstream of TFBSs identification (either by motif definition or TFBSs matrix scanning) is put on frameworks/cis-regulatory modules, not individual TFBSs and (ii) there is an additional step ‘evaluate’ to assess the relevance of the findings for the transcriptional output.
experiment. Without the focus in the frameworks this step is impossible as usually no sufficient association of individual TFBSs with function exists.

We followed this strategy for a published ChIP-Seq data set ([47], ChIP-Seq data only) in order to demonstrate that the approach does lead to meaningful associations. All of the analysis was carried out using the Genomatix Mining station for steps 1–3 (mapping, clustering, annotation) and the Genomatix Genome Analyzer for all downstream analyses. In brief, we took the reads, clustered them, subtracted the control clusters and added the genomic annotation. STAT1 binding sites were found to be promoter associated and we carried on with TFBSs analysis and motif definition, which yielded a weight matrix highly similar to but distinct from the known STAT matrices. Using this new matrix and a selection of the most enriched regions (92 regions with more than 700 reads each) we carried out the framework analysis (four regions selected by the program) all the way to genome-wide evaluation. Figure 5 shows the summary for the best framework (of four different TFBSs) in a genome-wide analysis. Notably, the resulting (annotated) genes found are associated with a biological function (cell morphogenesis) and associated with the experiment, as a third of all matches were located in the ChIP-regions (only 5% of all promoters are in the ChIP regions but 33% of the matches, a 6.6-fold over-representation). This was not an exhaustive analysis and was only meant to demonstrate that the strategy works on real-life data.

**Figure 5:** Genome-wide framework evaluation based on a STAT1 ChIP-data set. The blue background represents all genomic promoters, the red circle the ChIP-enriched regions. The golden-brown circle inside the ChIP region represents the top-enriched regions used for automatic framework analysis, which finally selected the dark brown four regions only. The resulting frameworks is shown on the right side and genome-wide evaluation on all promoters resulted in the 419 matches shown as a green/yellow circle, 192 of which yellow part) were found in ChIP-enriched regions. The GO-term association analysis revealed a significant association with the term ‘cell morphogenesis’ for the 419 promoters, linking the molecular framework to a functional GO term.

NGS-BASED FUNCTIONAL GENOMICS AND SYSTEMS BIOLOGY

Genome-wide and high resolution brings the other buzzword of these days to mind: systems biology. What is in for this integrative discipline from the new power of NGS technology? NGS data are not obviously improving quantitative models of biological systems. However, the networks and interconnections revealed by functional genomics downstream analysis of NGS data (DNA-seq, RNA-Seq and ChIP-Seq) represent the essential topological backbone each quantitative model requires as a pre-requisite regardless whether microarray- or NGS-based [9]. Moreover, the wide range of quantitative dynamics of RNA-Seq provides essential data to model quantitative phenomena of gene expression as well as providing ample data points that can be used to verify predictions made by various models.

Thus, systems biology approaches will benefit from the enormous data density intrinsic to NGS applications, which will beyond doubt play an important role both in definition as well as verification...
of mathematical models of biological systems such as a cell or a tissue. This also extends into parasite, especially virus host interactions from a systems view [48], which was recently highlighted in HIV-related studies [49, 50].

**IS NGS-FUNCTIONAL GENOMICS THE FUTURE OF PERSONAL GENOMICS/PERSONALIZED MEDICINE?**

Elucidation and utilization of individual genetic differences for medical applications is a major hallmark of personalized medicine. NGS methods allow detailed genome-wide analyses of individuals empowering new dimensions for stratification approaches far beyond conventional population statistics. The search for biomarkers is faced with two problems: first to find the correct sub-population for which a biomarker is really useful and second to define additional biomarkers important in the same context. In most cases, no one biomarker is enough to show a sufficiently good correlation with a disease or treatment success, necessitating the definition of sets of biomarkers that may well originate from NGS-based approaches especially in cancer diagnostics [51]. Since NGS methods show all effects and changes (expected as well as unexpected) they are well suited to help avoiding the tunnel-view of prior-knowledge-based methods such as microarrays. NGS methods are also already considered for biomarker development in the diagnosis of actue leukemias [52]. Again, such data becomes only valuable and useful for personalized medicine applications after a thorough and powerful functional genomics downstream analysis has been carried out. In this case the integrative analysis has to be pushed even further: personalized medicine requires the successful merging of genomic digital data (e.g. sequence-based data) with all -omics data, biochemical data (current clinical laboratory metabolite and cell count diagnostics), qualitative data (such as histological data), quantitative data from image analyses (CT/PET Scans, microscopic images, etc), and system-level medical data (e.g. bed-side diagnosis, results from physical examinations).

**PROBLEMS REMAINING AND OUTLOOK**

NGS based technology has so many great features and has pushed many limits so far ahead that it may appear as the magic bullet to solve all genomics-related problems right now. However, there is still a list of missing features that currently limit the applicability of the method.

The first is the relatively low read length together with the intrinsic error rate that increases with read length. This poses a few problems for the mapping that either have been solved or should be solved in the near future, but are still prohibitive for true de novo assembly of mammalian genomes from NGS reads, despite some promising developments [53]. In the absence of a template genome, no really chromosome-like contigs can be assembled regardless of the technology used—so far. Another problem impairing widespread application in many fields is that NGS provides extremely deep-coverage but relatively low throughput. It is no problem to routinely scan thousands of probes using microarrays today while a similar volume in NGS runs is still extremely challenging. However, looking back to the development of Sanger-based sequencing, the solution may be similar: Multiplexing the methods may alleviate this restriction gradually. Together with technological multiplexing (see the development of Sanger sequencing from single gel-based to hundreds of capillaries in one instrument) this might allow doing large-scale parallel NGS sequencing in the not-so-distant future. At least the advances made in NGS technology just within the last five years are very impressive and justify quite some expectations for the next 5 years—or until yet another disruptive technology takes over.

**Key Points**

- Genomic complexity is in genome expression not gene lists.
- Function arises out of context and connection networks.
- NGS allows elucidation of dynamic context-dependent connections.
- Downstream analysis alone (beyond read mapping) can elucidate biological function.
- NGS empowers real personalized genomics/medicine approaches.

**References**

et al

17. Shastry BS. SNPs: impact on gene function and phenotype.


13. Werner T. Proteomics and regulomics: the yin and yang of


9. Werner T. Regulatory networks: linking microarray data


