Abstract

DNA sequencing has revolutionized biological and medical research, and is poised to have a similar impact in medicine. This tool is just one of a number of developments in our capability to identify, quantitate and functionally characterize the components of the biological networks keeping us healthy or making us sick, but in many respects it has played the leading role in this process. The new technologies do, however, also provide a bridge between genotype and phenotype, both in man and model (as well as all other) organisms, revolutionize the identification of elements involved in a multitude of human diseases or other phenotypes, and generate a wealth of medically relevant information on every single person, as the basis of a truly personalized medicine of the future.

The human genome

Starting from having little knowledge of any of the information in the human genome a few decades ago, the combination of cloning [1] and sequencing [2,3] gave us our first access to (initially very small) parts of the human/mouse genome [4]. Through the development of automated sequencing machines [5], this first phase of technology development culminated in the first sequence(s) of the human genome as the result of the human genome project [6,7] followed up by a number of single genomes, all but the first [8] sequenced on different next generation sequencing platforms [9-15].

The variation between the different individuals and their haplotypes was first addressed systematically in the HapMap project [16-19], resulting in the identification of 3.1 million human single nucleotide polymorphisms (SNPs) typed in 270 individuals from 4 major populations, still based on a combination of Sanger sequencing with chip-based genotyping approaches.

With the availability of next generation sequencing platforms [10,14,15,20-24] (summarized in [25]), much larger scale analyses of genomes and genome variations became possible. The 1000genomes project [26,27] aims to provide information on rarer single nucleotide and structural variations in the human genome, by combining medium deep (typically defined as 4× coverage) genome and complete exome coverage of about 2500 individuals from 27 populations, combined with deep sequencing (>30× coverage) of a limited number of individuals/trios. In parallel, Grand Opportunity Exome sequencing project (GO-ESP), a project to sequence the exomes of 6700 individuals funded by the National Heart, Lung and Blood Institute (NHLBI) has focused specifically on the variations within the coding regions in specific patient groups with over 80 heart, lung and blood-related traits and other diseases of major importance [28].

In particular, the combination of genome, exome and transcriptome analysis is playing a key role in our understanding of mechanisms underlying cancer development, addressed particularly by the International Cancer Genome Consortium (ICGC, www.icgc.org) [29,30] and The Cancer Genome Atlas (TCGA, cancergenome.nih.gov), an analogous US-only project [31], generating comprehensive
catalogues of the somatic changes in different tumor entities by genome/exome sequencing (plus often additional information) of both tumor and germ line cells.

In contrast to these projects, which typically only collect very limited phenotypic information on the individuals analysed, the Personal Genome Project (PGP, www.personalgenomes.org) aims to sequence the genomes of up to 100,000 volunteers and to combine this information with a wide range of their phenotypic/disease history information [32].

The functional consequences of base pair changes or small deletions or insertions in the coding regions, which change the amino acid sequence of the protein, are typically easier to predict than for non-coding (e.g. regulatory) sequences. Different procedures [33-37] have therefore been used to enrich and sequence the exome (or other relevant regions of the genome), either alone, or in combination with more limited coverage of the entire genome. As an interesting variant [38], sequencing can also be targeted directly to specific sequences, by modifying the oligonucleotides on the sequencing chip.

Different genomes (and particularly cancer genomes) do not just differ in their genomic sequence, but are often also characterized by translocations, copy number variation and loss of heterozygosity. Specific translocations, for example, have been identified early as characteristic for specific types of tumors [39]. Next generation sequencing of the genome by so-called mate-pairs (sequencing the ends of large, circularized DNA fragments on a single fragment) has proven an effective technique to identify such rearrangements [40]. Similarly, translocations resulting in the fusion of transcripts observed, for example, in the case of fusion proteins, can be identified by RNAseq analysis [41,42].

The identification of larger scale copy number changes, first identified by comparative genome hybridization on chromosomes (CGH) [43,44], and then at higher resolution by array analysis (array-CGH) [45], has essentially been overtaken by sequence analyses [46-49], providing much better resolution, but also information on the copy number changes of the two haplotypes separately [50,51].

Given the same overall sequence composition, the biological function of the genome depends critically on the haplotypes contained within it, illustrated by the original definition of a gene in bacteria and phages as cistron on the basis of a cis-trans test (two mutations were considered to be in the same gene, if the phenotype of the cell or phage was different, if the mutations were carried on the same DNA [in cis] or on two different fragments [in trans]). Although this definition is not possible in eucaryotes, due to alternative splicing, mixed diploid sequencing will usually still not be able to determine, if, for example, two heterozygous loss-of-function mutations are in cis, with two mutations in one copy of the gene, but leaving the other copy intact, or in trans, inactivating both copies of the gene. However, this information can be gained by statistical analyses [52] or experimentally by an increasing number of experimental strategy approaches, for example based on the sequencing of pools of fosmid clones, sorted chromosomes or longer DNA fragments [53-57].

**Difficult materials**

Beyond the different platforms used in these analyses (Illumina, Solid, 454, PacBio, Complete Genomics, Ion Torrent) with more still under development (e.g. Oxford Nanopore), there are a number of technical variations focusing on specific aspects of the sequencing process. An important burden in the analysis is, for example, the still relatively high error rate of many of these sequencing platforms, as well as, in some cases, the effect of partially damaged DNA. A major step to address this problem has been taken by Schmitt et al. who labeled every DNA fragment with a random tag during library construction. In the case of errors in the sequencing process or of damage in the original DNA fragment, the sequences of the two strands of the original DNA will differ, flagging these variants as being caused either by damage to the original DNA or errors in one of the amplification steps in the sequencing process [58]. Special technologies also had to be developed for the analysis of badly preserved DNA, e.g. due to the age of the material [59,60] or to the formaldehyde action in formalin fixed paraffin embedded (FFPE) material [61].

Another difficult challenge with current technologies lies in the sequencing of very small amounts of DNA, e.g. from free DNA in serum, or from individual cells (e.g. circulating tumor cells) [56,62,63], a problem that could be simplified in the future by techniques able to analyse un-modified, un-amplified DNA samples.

**Functional analysis of the genome**

Ultimately, the sequence of the DNA has to be understood in terms of its many functions. With the recent publications on the Encyclopedia of DNA Elements (ENCODE, http://www.genome.gov/10005107) project, a large amount of information has been created. This work illustrates many of the techniques available to functionally characterize the genome [64-72] etc.
The epigenome/methylome
Methylation and other modifications of DNA play a major role in affecting its transcription [73]. A number of different approaches have been used to detect such DNA modifications [74-82]. The “gold standard”, based on treating the DNA with bisulfite, a reagent converting cytosin into uracil, while leaving 5-methylcytosine unchanged, followed by whole-genome sequencing, gives by far the largest amount of information, albeit at correspondingly high cost. At lower cost (and generating correspondingly less information), restriction digests can be used to enrich particularly informative segments of the genome for bisulfite sequencing [77], or to focus on specific positions using either next generation sequencing [78] or chip-based sequencing techniques. The Infinium 450K Methylation array allows the determination of the bases remaining after bisulfite treatment of more than 480,000 cytosines in the genome, selected as particularly informative for methylation analyses [83]. Alternative procedures (e.g. MeDip) rely on the selective isolation of DNA fragments carrying the appropriate modified base by antibodies or proteins binding selectively to a particular type of modified DNA, followed by next generation sequencing, or sequencing of short fragments generated by restriction enzymes targeting regions with many GpCs, the dinucleotide sequence carrying most methylation in mammalian DNA. However, methyl C is not the only modified base in mammalian genomes: hydroxymethyl C, an alternatively modified base, cannot be distinguished from methyl C through bisulfate-based analysis methods [84] but can be selectively identified by antibodies recognizing this DNA modification [85]. New sequencing platforms promise to allow direct detection of modified bases, very significantly simplifying this type of analysis [86,87]. An additional ‘epigenetic code’ beyond the direct modification of the DNA is, however, provided by histone modification [88], which can be analysed by chromatin immunoprecipitation (ChIP)-seq techniques [89,90].

Regulatory sequences, genome structure
While it is relatively straightforward to identify protein-coding sequences, the identification of regulatory sequences or other sequence elements still represents a significant challenge. Interesting, new sequence-based approaches to identify such elements include, for example, the identification of sequences containing DNase-sensitive sites [91]; the mapping of open chromatin by crosslinking chromatin with formaldehyde; extracting the DNA cross-linked to protein by phenol extraction, and sequencing the remaining free DNA (FAIRE sequencing [92]); ChIP-seq with antibodies or other specific binding molecules (aptamers, affibodies, etc. [93-96]) against transcription factors [89]; the identification of binding sites by fusing the protein of interest (in this case a transcription factor) to a Dam methylase gene and then sequencing the DNA protected from digestion by the methylation (DamID, [97]); and the generation of quantitative enhancer maps by cloning the genome in fragments downstream of a minimal promoter, causing the enhancers to be transcribed at a level proportional to their enhance strength (STARR-sequencing, [98]).

A novel variant in the use of next-generation sequencing equipment to identify genomic sequences binding to specific proteins, high throughput sequencing-fluorescent ligand profiling (HiTS-FLIP) [99], takes advantage of the optics and fluidics of an Illumina sequencer to score the binding of a fluorescence-labeled protein or other ligand, as an interesting alternative to previously described protocols [100].

A wide range of new, sequence-based techniques has also been developed to analyze the proximity of different elements of the genome to each other either directly (C3, C4, C5, HiC) or after selection for the presence of specific proteins (ChIP-loop, ChLA-PET), or their proximity to specific structural elements (e.g. nuclear membrane) by DamID or other protocols [101-104]. Similarly, next-generation sequencing has allowed a detailed analysis of the pattern of replication of the genome [105].

Transcriptome analysis
The analysis of transcripts by next-generation sequencing techniques (RNAseq) in its many different forms addresses a key step in the flow of information from the genome (and epigenome) to the phenotype of the organism. It provides information on many different types of transcripts (protein coding, short and long non-coding RNAs, micro RNAs), and has revolutionized the analysis of expression patterns, alternative splicing and allele-specific expression, providing unbiased digital information far beyond the results provided by different hybridization-based platforms [106,107].

In general, RNAseq has been carried out by first converting the RNA into cDNA, which is then sequenced as described above. Direct RNA sequencing has, however, also been described [108]. Another interesting variant is provided by on-flowcell reverse transcription sequencing (FRT-seq), the direct reverse transcription of mRNA on the flowcell, eliminating steps that could lead to specific artifacts [109].

A variety of protocols have been developed to handle more samples in parallel [110]. Earlier protocols did not preserve the information on the strandedness of the transcripts,
neglecting essential information [111,112]. This has been addressed in more recent protocols [113,114].

A wide range of protocols has been developed to focus on particular types of transcripts: coding, long and short non-coding and micro RNA etc, addressed by different protocols (poly-A plus, ribo-minus, etc.) [115]. An interesting variant to the ribo-minus strategy, relying on the removal of ribosomal RNA (and, in newer versions, also mitochondrial RNA) from the library involves “not so random” hexamer primers, selected not to contain sequences able to prime on ribosomal RNA, but to prime cDNA/double strand production [116].

To selectively analyse the 5' end of RNAs from very small amounts of RNA, two new protocols, nanoSCAN and CAGEscan, have been described [117] as modifications of the original CAGE protocol [118,119]. Paired-end (PET) sequencing [120] can be used to advantage to compensate, to some extent, for the typically short read length of most current next-generation sequencing platforms. The analysis of alternative splicing patterns, however, still remains a difficult problem. Longer reads (PacBio, 454) could contribute significantly to the identification of the exact structure of transcripts for every gene expressed in a sample [121-126]. The mapping of branch points also adds relevant information [127]. Analysis of allele-specific expression patterns [128] and changes by RNA editing also provides essential information as to the function of a gene [129], as ultimately the genome acts through the RNA.

Many techniques have been developed to identify and quantitate microRNAs, to identify their targets, and to analyze their functions [130-134]. In addition, more and more long non-coding RNAs have become increasingly associated with many different regulatory processes [135].

There are obvious technical difficulties in cases where only small amounts of RNA, e.g. from single cells, have to be sequenced. Significant progress has been made, but there are obvious limitations due to the inherent technical and biological noise in this type of data [136-140].

Transcript abundance can change due to differences in synthesis or degradation rate. To be able to distinguish between these parameters, different procedures allowing the selection and subsequent sequencing of newly synthesized RNAs have been developed [141-145].

In complex tissues, alternative techniques are needed to selectively analyze transcripts from specific cell types [130,146,147]. As an alternative, in-situ transcriptome sequencing could combine spatial resolution with the information content of transcriptome analysis. Direct in situ sequencing protocols, based, for example, on polony sequencing [148,149] have inherent limitations in the combined resolution versus sequencing depth, which can be overcome by the use of spatially encoded oligonucleotide primers (unpublished).

**Proteins and protein interactions**

While nucleic acids can be essentially detected and characterized down to the level of single molecules, this is typically not the case with proteins. In spite of major progress in mass spectroscopy (e.g. [150-153]) and other analysis techniques (e.g. [154-159]), we are still far from the power provided, for example, by next-generation sequencing in transcriptome analysis. A significant effort has, therefore, been directed at converting the analysis of proteins into a nucleic acid analysis problem. Sequencing the RNA protected by the ribosome does, for example, give detailed information on the protein synthesis, determines translation rates and identifies previously unknown proteins [160].

To be able to apply the sensitivity and throughput of nucleic acid analysis, a number of techniques have been developed to tag proteins, antibodies or other binding agents (or even chemicals) with nucleic acids, which can then be analyzed by deep sequencing. In proximity ligation for example, two or more binders are tagged with different oligonucleotides, which can form amplifiable sequences, if they are held in close proximity, allowing the highly sensitive detection of proteins (two binders to the same protein), protein modifications (one binder to the protein, one to the modification), protein-protein complexes (one binder each for both proteins), or larger structures (multiple binders carrying sequences), which will only give an amplifiable result if all components are present in close proximity [161,162]. Similarly, the results of different types of protein interaction assays (e.g. from a two-hybrid analysis) can be read out by selective amplification and sequence analysis [163-166].

**Metagenome analysis, cell phenotypes, and much more**

Microbial populations can play an important role in human diseases and other phenotypes. Next generation sequencing has allowed much more detailed analyses of these complex populations [167-169].

Next-generation sequencing techniques can also be used to great advantage to analyze the effect of specific conditions on a cell population marked by specific sequences with little effort. Next-generation sequencing can also help to identify causal variants for interesting phenotypes at the cell level, by either submitting populations of different
cells recognizable by a (introduced or naturally present) specific sequence, followed by analyzing the differences in sequence representation before and after the selection step, or by selection for specific phenotypes, followed by sequencing the genome to identify the causal variants [170-173].

As sequencing costs have dropped (and, in spite of the current reversal, they are likely to continue to drop over the longer term, driven by new technology platforms) and more applications are transformed into DNA sequencing (see [174] for a proposed approach to analyse neuron connectivity by DNA sequencing), we can expect to have increasingly different types of information available, not only for basic research but also for the individualized application of this knowledge for the benefit of individuals. This is likely to be the case, in spite of the currently high analysis costs often dwarfing the costs of data generation [175], as the existing, only partly automated, analysis pipelines mature, with data analysis ultimately limited by the costs of the electricity required for the computation.

**Human phenotypes and diseases**

At the beginning of positional cloning of Mendelian traits in man and other mammals, it took many years to identify some of the first human and mouse genes defined only by their phenotype [74,176]. Today, this could be carried out within weeks, after the family members have been collected or appropriate crosses have been carried out. Exome sequencing has already allowed the identification of causative mutations in a large number of analyses (e.g. [177-182]). Next-generation sequencing is, in particular, able to also identify causative mutations for diseases or phenotypes, for which no or too little family material is available (e.g. new mutations) [183-185].

The analysis of multifactorial traits by genotyping in general is limited to common alleles (the ‘common disease-common allele’ hypothesis) [186]. It has, however, become increasingly obvious that many phenotypes are caused by many rare alleles, or copy number variants, leaving exome or genome sequencing as the most obvious analysis route [187,188]. Interestingly, it has been shown that the combination of low coverage sequencing and imputation can be a cost-effective alternative to standard chip-based genotyping techniques [189]. Genome/exome sequencing, therefore, increasingly complements or replaces genotyping-based analyses, and is particularly important for providing clinically relevant information on the individual [190,191].

Next-generation sequencing has also proven itself as a relevant and powerful tool to detect disease-causing mutations in the genome of the embryo in preimplantation diagnosis [192], or to analyse fetal nucleic acids in maternal plasma (e.g. for diagnosis of trisomy 21) [193,194].

Sequencing and other -omics techniques have proven particularly important for the analysis of tumors, since cancer, in a sense, is a “genomic” disease [29]. Analysis of tumors or tumor-derived cell lines [195] by deep genome/exome and transcriptome sequencing, combined with sequencing of the genome/exome of the patient, plays an increasingly important role in guiding the therapy choice [196-198], either through the identification of “actionable” variants, or, in future, increasingly through computer models [199,200] able to incorporate many different types of information to generate “virtual patient” models, on which the treatment of the individual patient can be optimized [201].

**The future**

We come from a situation a few decades ago when we hardly knew anything about our genome and its components. Development of cloning, sequencing and polymerase chain reaction (PCR) techniques has allowed us to identify specific genes and analyze their function, or, increasingly, identify the gene responsible for a specific (organismal) phenotype or disease. Completing the sequence of the human genome of 3 billion bases has revolutionized our understanding of biology and medicine, and has made many tasks, which were either impossible or very, very difficult, (relatively) easy. We now know the sequence of thousands of genomes, and are likely to know, sooner or later, the sequence of everybody’s genome, complemented by a wide range of different analysis techniques (limited much more by the availability of samples, than by the complexity or cost of the analyses). We have moved from hybridisation-based array/chip analyses [202,203], generating the first “big data” types, more and more to sequence-based analyses, from gene sets, to exomes and whole genomes. Many of these analyses have become fairly straightforward; others, and in particular the analyses of indels/copy-number variations (CNVs), are still difficult, even combining a wide range of different analysis techniques [51,204].

In going from essentially no knowledge to knowing billions of bases of one human genome (as well as a lot of additional information on transcripts, proteins and metabolites, etc.) to many billions of bases of the genome and other -omics data for billions of humans puts us roughly half way (in a log scale) on a long road aiming to use abundant information (and computing power) to optimize treatment, prevention and well-being for everybody.
But what will we do with this abundance of information? A total of 32 (binary) biomarkers could give 7 billion different combinations, one for every person alive. 270 such biomarkers with an appropriate distribution, a very small number compared to the millions of differences between different human genomes, would be sufficient to identify every single atom in the observable universe. Computer models like those we use for weather forecasting give increasingly better predictions the more (and the more different types of) information they can be based on, which is typically not the case for many statistical procedures. Every patient is different. Every tumor (and in fact almost every cell of every tumor) could be considered a different “orphan” disease. We will therefore probably need all the information we can get to address this individuality, integrated by “virtual patient/virtual individual” models of every patient (including every functionally distinct subset of tumor cells in every tumor and, at least for prevention, every individual), which we had proposed in our FET-flagship project IT future of medicine (ITFoM, see www.itfom.eu) as the only reasonable way to integrate the huge medical datasets generated by the wide range of technologies available now, and likely to become available in the future.

Abbreviations
CGH, comparative genome hybridization; ChIP, chromatin immunoprecipitation; CNVs, copy-number variations; ENCODE, Encyclopedia of DNA elements; FFPE, formalin fixed paraffin embedded; FRT-seq, on-flowcell reverse transcription sequencing; GO-ESP, Grand Opportunity Exome sequencing project; HiTS-FLIP, high throughput sequencing-fluorescent ligand profiling; ICGC, International Cancer Genome Consortium; NHLBI, National Heart, Lung and Blood Institute; PCR, polymerase chain reaction; PGP, Personal Genome Project; SNPs, single nucleotide polymorphism; TCGA, The Cancer Genome Atlas.

Disclosures
I am founder of and advisor to Alacris Theranostics, which aims to develop ‘virtual patient’ models for use in therapy choice and drug development.

Acknowledgements
I am very grateful to many colleagues at the Max Planck Institute for Molecular Genetics, the Dahlem Centre for Genome Research and Medical Systems Biology and Alacris Theranostics, who have contributed to this review, in particular Sylvia Krobitsch and Wilfried Nietfeld, who have helped me fight with the many references, and Sylvia Krobitsch, Marie-Laure Yaspo, Michal Schweiger, Marc Sultan, Ralf Sudbrak, Margret Hoehe, Albert Pousta, Tatjana Borodina, Jörn Glöckler and Bodo Lange for discussions and feedback. I am also very grateful to many colleagues outside, especially Ulf Landegren, Stefan Beck and George Church, for free exchange of ideas and to two anonymous referees for helpful suggestions.

References


175. Mardis ER: The $1,000 genome, the $100,000 analysis? Genome Med 2010, 2:84.


