J Lab Med 2010;34(6):311–318 © 2010 by Walter de Gruyter • Berlin • New York. DOI 10.1515/JLM.2010.054

Molekulargenetische und zytogenetische Diagnostik

Redaktion: H.-G. Klein

Translation of next-generation sequencing (NGS) into molecular diagnostics

Umsetzung von Next Generation Sequencing in der molekularen Diagnostik

Stefan Kotschote[1], Carola Wagner[1], Christoph Marschall[2], Karin Mayer[2], Kaimo Hirv[2], Martin Kerick[3], Bernd Timmermann[3] and Hanns-Georg Klein[1,2,*]

[1] IMGM Laboratories GmbH, Martinsried, Germany
[2] Center for Human Genetics and Laboratory Medicine Dr. Klein and Dr. Rost, Martinsried, Germany
[3] Max-Planck-Institute for Molecular Genetics, Berlin, Germany

Abstract

In the past 5 years, next-generation sequencing (NGS) has been established as a valuable tool for several research applications. Commercially available platforms from Helicos, Illumina, Life Technologies, Pacific Biosciencies, and Roche allow for massively parallel sequencing and analysis in the fields of genomics, transcriptomics, and epigenomics. As in most projects, data throughput of the sequencers is not the limiting factor; genomic DNA samples are directly prepared for sequencing without prior conditioning. However, there are some applications such as targeted resequencing that do not require sequencing of whole genomes. Therefore, a technology called target enrichment was established more than 2 years ago. Different PCR- or hybridization-based approaches were further commercially developed and refined. The combination of this method with NGS can improve analysis of disease-related gene sets in molecular diagnostics by reducing time and costs. By taking advantage of the enormous data output, several genes and patients can be analyzed in parallel in one single instrument run.

Keywords: immunogenetics; molecular diagnostics; molecular genetics; next-generation sequencing (NGS); target enrichment.

*Correspondence: Hanns-Georg Klein, MD, Center for Human Genetics and Laboratory Medicine Dr. Klein and Dr. Rost, Lochhamer Str. 29, 82152 Martinsried, Germany
Tel.: +49-89-895578-0
Fax: +49-89-895578-780
E-Mail: hanns-georg.klein@medizinische-genetik.de

Zusammenfassung

In den letzten 5 Jahren hat sich Next Generation Sequencing zu einer wertvollen Methode für verschiedene Forschungsanwendungen entwickelt. Die kommerziell von Helicos, Illumina, Life Technologies, Pacific Biosciencies und Roche erhältlichen Plattformen ermöglichen eine massive parallele Sequenzierung und Analysen in den Bereichen Genomik, Transkriptomik und Epigenomik. Da in den meisten Projekten der von den Sequenziergeräten generierte Datendurchsatz nicht limitierend ist, werden genomische DNA Proben ohne weitere Vorbehandlung direkt für die Sequenzierung vorbereitet. In Anwendungen wie der gerichteten Resequenzierung wird jedoch keine Sequenzierung von Gesamtgenomen benötigt. Deshalb wurde vor über 2 Jahren eine Technologie namens Zielregion-Anreicherung (Target Enrichment) entwickelt. Verschiedene PCR- oder Hybridisierungs-basierte Ansätze wurden danach erarbeitet und weiterentwickelt. Die Kombination dieser Methode mit Next Generation Sequencing kann die Analyse von krankheitsrelevanten Gensets in der Molekulardiagnostik durch Reduktion von Zeit und Kosten verbessern. Unter Ausnutzung der enormen Datenausgabe können mehrere Gene und Patienten gemeinsam in einem Gerätelauf analysiert werden.

Schlüsselwörter: Immungenetik; Molekulare Diagnostik; Molekulargenetik; Next Generation Sequencing (NGS); Zielregion-Anreicherung.

Introduction of next-generation sequencing into molecular diagnostics

In 2005, Roche 454 Life Sciences was the first company that commercialized a next-generation sequencing (NGS) platform [1]. In the following years, NGS has dramatically changed basic genomics research [2, 3]. It is now possible to perform experiments such as whole-genome or transcriptome analysis, which were previously technically neither feasible nor affordable [4]. The potential of NGS applications only seems to be limited by one's imagination [5]. There are three main study areas of interest where NGS is used: the genome, the transcriptome, and the epigenome.

Table 1 Data output of the currently available NGS platforms and the theoretically possible number of samples that can be analyzed together in one sequencing run.

Platform	Manufacturer	Sequence output per run (MB)	400 bp – Calculated multiplexing with 50× coverage	1 MB – Calculated multiplexing with 50× coverage
GS Junior	454 Sequencing	40	2,000	1
GS FLX	454 Sequencing	500	25,000	10
GA IIe	Illumina	10,000	500,000	200
GA IIx	Illumina	25,000	1,250,000	500
HiSeq2000	Illumina	100,000	5,000,000	2,000
SOLiD4	Life Technologies	100,000	5,000,000	2,000
AB3730xl	Life Technologies	0.08	N/A	N/A

Regarding translation of established genome sequencing approaches, there are several possible applications for molecular diagnostics. In oncology, it could be beneficial to sequence and analyze whole cancer genomes to apply individual treatments and to predict the outcome, whereas individual whole-exome sequencing could help identify genotypes that are causative for given phenotypes. Further approaches are targeted resequencing, e.g., to analyze distinct disease-related groups or gene sets, the analysis of chromosomal rearrangements for molecular cytogenetics (e.g., in prenatal diagnostics [6–8]) and the quantitative characterization of infectious agents for molecular microbiology and virology.

With regard to transcriptome sequencing approaches such as expression profiling (whole or targeted transcriptome) or small RNA profiling (miRNA, piRNA, non-coding RNA, etc.), there is actually less demand for translation into molecular diagnostics. These applications are currently very informative to identify potential biomarkers that will later enter molecular diagnostics; a situation comparable to the application of microarrays in earlier days.

Further applications such as chromatin immune-precipitation sequencing (ChIP-Seq) can still serve well in basic research, whereas the analysis of different DNA methylation patterns as second epigenomic application seems to be of more interest for molecular diagnostics. For example, four identified genomic loci seem to be powerful epigenetic biomarkers of breast cancer in circulating DNA, as tumor samples displayed more variation in methylation level than normal samples [9]. Furthermore, a field of application might be the analysis of response to drug treatment supposed to depend on methylation status in patient DNA.

Progress in the different NGS technologies has been made, especially in robustness, accuracy, and technical procedures. Molecular diagnostics will benefit from continued improvements in the whole process including automation, simplified and standardized workflows, chemistry enhancements, cost reductions, and advanced data handling. Although costs of sequencing reagents per sequencing run are still substantial, the cost per base compared to Sanger sequencing is reduced enormously and is associated with an increase in data output. Further cost reduction per sample can be made if the full capacity of the NGS platform is not needed for analysis of individual samples. Multiple samples can be analyzed in parallel in separate compartments by ligation of unique identifier sequences (''barcodes'') to individual samples before pooling and joint sequencing. During data analysis, indexed sequence reads are reallocated to the individual samples [10]. The degree of multiplexing always depends on (i) the sequence output per run of the specific platform, (ii) the length of the target sequence, and (iii) the expected average coverage rate. Examples for 50× coverage and two target regions of 400 bp and 1 MB, respectively, are shown in Table 1.

Validation of target enrichment methods

The combination of NGS with target-enrichment techniques provides a promising approach for the translation of technology into molecular diagnostics by facilitating the identification and characterization of genetic variants at specific loci associated with complex diseases or phenotypes. Currently, two approaches are predominantly being utilized for enrichment of target sequences from whole genomes. The first method is PCR-based and commercially available for different set-ups from RainDance Technologies (Lexington, MA, USA) (RainStorm™ [11]) or Fluidigm (South San Francisco, CA, USA) (AccessArray™ [12]). Although target enrichment by common single or multiplex PCR is highly specific and sensitive, scaling of the method is difficult. The second method is based on hybridization of the target sequences to oligonucleotide probes, either on-array [13–15] or in-solution [16] and is also commercially available from Agilent Technologies (Santa Clara, CA, USA) or Roche NimbleGen (Madison, WI, USA). It was reported that target enrichment by hybridization in biotinylated-cRNA probe solution was highly efficient, uniform, and reproducible [17]. This method could be well suited for population studies of loci in the mega-base pair scale using current sequencing technologies.

The interest in translating NGS into molecular diagnostics is primarily driven by the need for the analysis of multiple disease-related genes (gene sets) to make diagnostic sequencing more efficient. Recently, we initiated a study to assess different target-enrichment methods for their potential use in targeted resequencing. This validation study includes a set of 67 selected genes that are known or hypothesized to be linked to cardiomyopathies or connective tissue disorders.

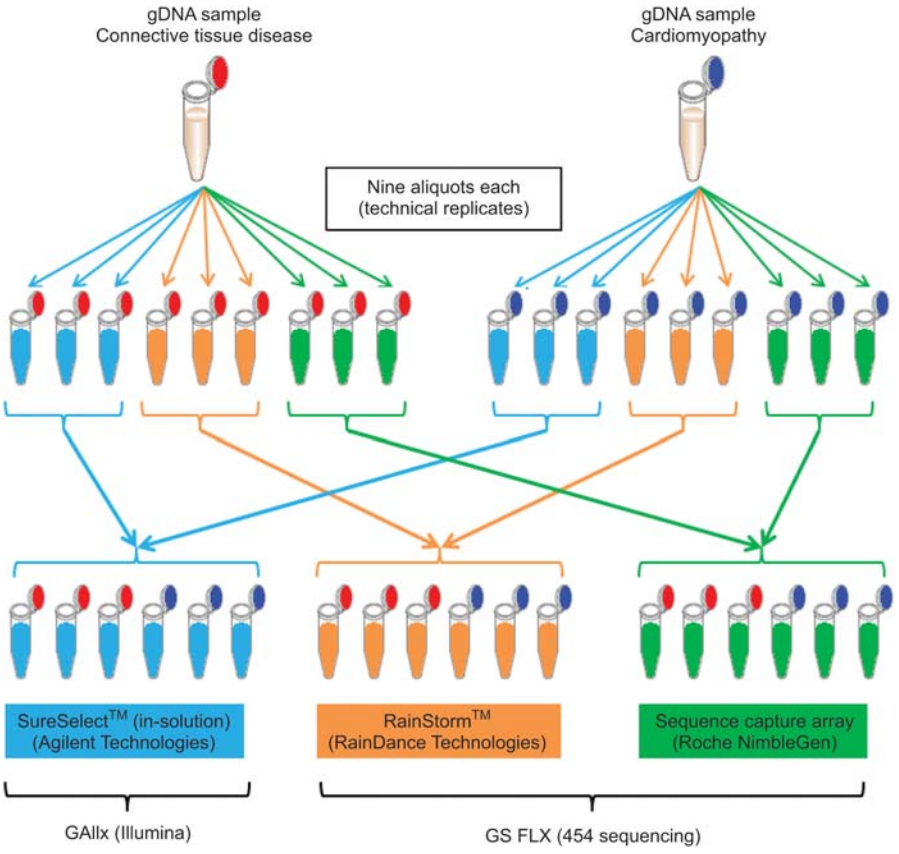

Figure 1 Schematic outline of the design of the target enrichment validation study.
DNA from two different samples with known mutations were selected for target enrichment and separated into nine aliquots each. Three replicates per sample were subjected to individual target enrichment and library preparation on three different platforms. Enriched samples from RainDance Technologies and Roche NimbleGen were sequenced with 454 Genome Sequencer FLX, those from Agilent Technologies with the Illumina Genome Analyzer II.

The target region of the 67 genes comprised 2,224 exons with a total length of 544,031 bp. Complete sequence information was received from the UCSC genome browser (hg18) based on exon coordinates and RefSeq entries. Two genomic DNA samples with known phenotypes and previously identified genomic mutations were analyzed in technical triplicates and three enrichment methods: micro-droplet PCR-based (RainDance Technologies), in-solution hybridization (Agilent Technologies), and on-array hybridization (Roche NimbleGen; Figure 1).

Target enrichment – basic background information

The following designs were implemented according to the manufacturer's recommendations: for the Agilent Technologies in-solution enrichment, selected 120-mer probes with 2× tiling frequency of all exon sequences were chosen with +20 bp of the adjacent intron sequences resulting in 632,991 bp of targeted sequence. This strategy was superior compared to longer probes (170-mer) and an end-to-end approach [17]. The Roche NimbleGen on-array enrichment design is based on the SSAHA (Sequence Search and Alignment by Hashing Algorithm) where at most two mismatches were allowed using multiple tiling on all exons, extending small exons to 200 bp. The primer design strategy of the PCR-based RainDance Technologies approach is based on single nucleotide polymorphism (SNP) masking, a minimum tiling of potential amplicons and at most one off-target match of the primer pair. The final primer design was performed using the Primer3 software (http://fokker.wi.mit.edu/primer3/input.htm). Oligonucleotide primers for the likewise PCR-based Fluidigm enrichment approach were designed to specifically and uniquely match the target sequence without any off-targets and to avoid known SNP positions.

For comparison and interpretation of the performance of different enrichment strategies, diverse handling of repetitive sequences, the size of targeted intron-exon boundaries, multiple tiling approaches and accepted mismatches have to be considered during the design process. In addition, inert technical specifications of sequence hybridization and PCR for target enrichment could influence performance and outcome. Hybridization based methods have to deal with diverse binding affinities due to different melting temperatures (Tm) of each probe and a known hybridization preference of GC-rich regions. This can lead to an imbalance of captured sequences and consequently of matched reads to the targeted regions

regarding the GC-content. Furthermore, secondary structures as well as sequence repeats of the target regions can also influence the binding affinity and finally the capturing efficiency. Apart from similar difficulties relative to Tm, GC-preference and secondary structures and their effect on the performance of primer binding and hence PCR-based capturing efficiency, the detection of reliable SNPs in the target region could be additionally influenced due to the error rate of polymerase activity in the PCR. Ultimately, all enrichment strategies aim at the increase of reads and coverage of the genomic regions of interest. The most efficient enrichment would minimize off-target reads in combination with high on-target coverage.

According to one study [18], at least an average coverage of 40× to 50× should be aspired to finally reach a coverage of 20× with a likelihood of 95%. Based on a theoretical model, the hypothesis of reaching an equal and sufficient coverage holds true, but the final coverage is highly dependent on the sequence composition of the target region to be enriched [19, 20]. Regarding the generated reads in our experiment, an average coverage of approximately 2,400× was expected for the Agilent Technologies in-solution enrichment and short-read sequencing by using the Genome Analyzer II (GAII, Illumina, San Diego, CA, USA). Using the NimbleGen on-array enrichment (Roche NimbleGen) in combination with the long-read sequencing by the Roche 454 (Branford, CT, USA) Genome Sequencer FLX (GS FLX), an average coverage of 136× was expected (Table 2).

After mapping against the whole human genome reference sequence (hg18) and filtering on uniquely mapped reads, an average coverage of approximately 1,000× (max.>4,000×) was reached with regard to Agilent Technologies in-solution/GAII, whereas a mean coverage of 20× to 30× was determined regarding NimbleGen/GS FLX (Table 3). Finally, many reads mapped off-target; however, unspecific mapping could be excluded with >80% (Agilent Technologies in-solution/GAII) and >90% (Roche NimbleGen/GS FLX) uniquely mapped reads were identified. The major reasons for the lower coverage seem to be (i) unspecific capturing due to sequence homologies (gene families), (ii) capturing of pseudogenes, and (iii) loosing sequencing capacity on the flanking regions of the target sequences. Therefore, if using hybridization-based enrichment technology, on-target coverage highly depends on the theoretical coverage that is based on the targeted size of all genes as well as parallel analyzed samples by indexing. Apart from the theoretically possible coverage, the following parameters of the target region strongly influence the final coverage: GC-content, repetitive elements, homopolymer stretches, possible pseudogenes, and genes targeted by sequence homology within members of the same gene family.

PCR-based enrichment strategies can avoid the problem of unspecific enrichment to some extent and therefore increase the on-target coverage, because gene specific primers could be designed for only one member of a gene family or without detecting pseudogenes. By contrast, a study [21] revealed that the coverage variability highly depends on the library size. Moreover, amplicon ends are overrepresented if not applying a PCR primer with a 5′ modification.

Diagnostic applications of NGS in molecular genetics

Research studies using NGS applications confirmed the importance of obtaining more human genome sequence information and suggested that this information will have a noteworthy impact on molecular medicine. So far, NGS approaches have been primarily used in research for either rapid whole-genome sequencing or region specific resequencing supporting gene mapping and population genetics studies [22, 23]. Currently, there is an increasing demand for whole gene sequencing, but diagnostic applications for whole-genome sequencing are rather unclear. The number of genes characterized in association with genetically heterogeneous hereditary diseases has continuously increased, but the clinical utility of this knowledge is still limited [24, 25].

The implementation of such powerful techniques such as NGS for diagnostic applications using resequencing of targeted disease genes also requires specific training and extensive experience to achieve a safe handling of the technology, particularly in terms of the sequencing methods, target enrichment, as well as the analysis and the management of the massive amount of data. The low-scale amplicon analysis currently dominating the genetic diagnostics will most likely be replaced by large-scale resequencing of entire disease gene pathways and networks, particularly for the so-called complex disorders. Although whole-genome sequencing can become standard for some diagnostic applications, resequencing of defined genomic regions currently remains the most

Table 2 Run descriptive values of technical triplicates after Roche NimbleGen on-array enrichment and GS FLX sequencing (KN1-3), as well as Agilent Technologies in-solution enrichment and GAII sequencing (KA1-3).

	Total no. of generated reads	No. of mappable reads[a]	Mappable reads, %	No. of uniquely mapped reads	Uniquely mapped reads, %
KN1	233,465	231,533	99	217,919	93
KN2	250,988	249,157	99	234,765	94
KN3	239,275	237,367	99	225,229	94
KA1	20,953,595	18,530,115	88	17,445,575	83
KA2	21,970,535	19,662,556	89	18,528,931	84
KA3	24,562,250	20,933,116	85	19,327,748	79

[a]Mappable reads = reads which mapped against the human reference sequence hg18.

Table 3 Coverage and reads on-target after NimbleGen enrichment (N) in combination with GS FLX (Roche 454) sequencing, as well as Agilent in-solution enrichment (A) combined with GAIIx sequencing (Illumina).

	Avg. coverage on target region[a]	% of Bases on target region not covered[b]	No. of reads on target region[c]	% of Reads on target region	No. of reads ±200 bp of target region	% of Reads ±200 bp of target region
KN1	39.57	0.36	73,842	33.89	131,629	60.40
KN2	26.97	0.58	53,762	22.90	94,043	40.06
KN3	41.11	0.33	77,964	34.62	139,120	61.77
KA1	962.40	0.65	8,613,192	49.37	11,614,612	66.58
KA2	1,033.40	0.59	9,249,278	49.92	12,312,629	66.45
KA3	1,094.25	0.80	9,819,708	50.81	12,921,777	66.86

[a]Average coverage is calculated on the single base level, i.e., number of reads per base of target region. [b]Calculation is based on 665,484 bases corresponding to the enriched target region. [c]Each read which mapped with at least one base on-target.

reasonable approach in terms of both cost and clinical benefit. From our perspective, the most likely diagnostic NGS application in molecular genetics is targeted resequencing of multiple disease-related genes leading to an increase in diagnostic sensitivity and allowing an improved identification of causative genetic alterations in families with previously uncharacterized disorders. NGS should be particularly useful in heterogeneous disorders with currently low mutation detection rates (e.g., Brugada syndrome). Targeted resequencing requires substantially less throughput per sample compared to whole-genome sequencing, but owing to the high capacity of next generation sequencers, unprecedented requirements are placed on the upfront methods of sample preparation. However, combining enrichment technologies with NGS is a powerful sequencing tool that definitely has the potential to be adapted for diagnostic applications.

Aside from ethical and legal questions related to the German Gene Diagnostics Law, the technical challenges need to be mastered, before the technology is safe for diagnostic purposes. Defined quality criteria and generally accepted guidelines are needed for SNP and INDEL detection in terms of library preparation, targeted enrichment, and data analysis settings (including the accuracy of reads, the quality scores for reads, and sequencing coverage). To ensure reliable results, accuracy should be optimized by automation of sample processing (library preparation, enrichment, and sequencing). Before specific QC guidelines can be elaborated for diagnostic laboratories, a large sample set needs to be studied to validate the technique. NGS needs to be compared to Sanger sequencing, the current gold standard for diagnostic sequencing. The enrichment design and the algorithm utilized for data analysis is particularly important. The existence of multiple pseudogenes (or highly homologous sequences) has to be taken into account, if capture-based technologies and highly sensitive NGS for gene sequence analysis are being used. Compared with short-read NGS platforms (i.e., SOLiD from Life Technologies and the GAII systems from Illumina), the long sequence reads provided by the GS FLX system not only allows more efficient assembly for the detection of large insertions (particularly repetitive sequences) but also facilitates identification of genomic signatures from highly homologous sequences [26].

Although improvements will be necessary in accuracy, speed, data handling, and cost, the identification of novel mutations in more than 1,000 exons representing 100 candidate genes for ocular birth defects has been published, demonstrating that NGS on the Roche GS FLX is a valuable tool for mutation detection [27]. The detection, however, of deletions or duplications in long homopolymer stretches of one single nucleotide can be cumbersome; thus, different software packages should be used for analyses [23, 24]. High throughput diagnostic sequencing of many genes at a time, or even sequencing of the entire genome, is expected to reveal thousands, perhaps millions, of novel genetic variants in the patient analyzed. To minimize the risk of causing uncertainties and anxieties among patients, if no clear clinical consequences can be offered, the differentiation of benign polymorphisms and potentially pathogenic mutations have to be carefully examined. Diagnostic application of NGS can be reasonable in diseases caused by mutations in numerous and large genes, particularly if the differential diagnosis of the clinical phenotype is difficult. Examples for diseases with strong phenotypic overlaps include cardiac arrhythmias, cardiomyopathies, connective tissue disorders, and idiopathic epilepsies. Instead of a step-by-step analysis of several genes according to the frequency of described mutations, NGS provides the opportunity of massive parallel sequencing of multiple disease genes in a reasonable time period.

Some disease panels including the number of individual genes and exons to be analyzed are depicted in Table 4. In syndromes, associated with sudden cardiac death, 30–40 genes with 500 exons could be analyzed simultaneously. These genes include those being characterized as causative for long-QT syndrome, Brugada syndrome, arrhythmogenic right ventricular dysplasia, catecholaminergic polymorphic ventricular tachycardia, and hypertrophic cardiomyopathy. If the clinical phenotype is more specific, the analyses could be restricted to 5–15 genes. Owing to phenotypic pleiotropy, the classification of connective tissue disorders is often difficult. Particularly, various types of Ehlers-Danlos syndrome, Marfan syndrome, and related phenotypes, as well as different diseases characterized by aortic aneurysms and dissections require classification and risk stratification [28, 29]. The analysis of specific gene panels by NGS could become a powerful approach to identify the underlying genetic variants and allow genetic counseling of relatives at risk for aortic aneurysm. Genome research on epilepsies has led to

Table 4 Examples of gene panels for the analysis of genetically heterogeneous disorders.

	Number of genes	Number of exons
Connective tissue disorders		
Marfan syndrome and related phenotypes	4	148
Ehlers-Danlos syndromes	14	512
Aortic aneurysm syndromes	8	226
Idiopathic non-syndromic epilepsies	16	266
Benign neonatal/infantile seizures	3	58
Febrile seizures beginning in infancy	4	67
Childhood epilepsies	5	53
Adolescence-adult epilepsies	4	59
Arrhythmogenic cardiac disorders/ cardiomyopathies		
Long QT syndrome	11	124
Brugada syndrome	5	103
ARVD	5	85
CPVT	2	116
HCM	13	160
Sudden cardiac death	32	493

ARVD, arrhythmogenic right ventricular dysplasia; CPVT, catecholaminergic polymorphic ventricular tachycardia; HCM, hypertrophic cardiomyopathy.

the identification of more than 20 genes with a major effect on susceptibility to idiopathic epilepsies. Although seizures and epilepsies are still classified by seizure type and age of onset, new concepts focus on the underlying type of cause and differentiate into genetic, structural-metabolic, and unknown. In a revised terminology, the International League Against Epilepsy commission on classification proposed the term electroclinical syndromes to define a specific diagnosis on the basis of age onset, EEG characteristics, and seizure types [30, 31]. If these electroclinical syndromes are arranged by age of onset, NGS provides a helpful tool to analyze several known genes in an appropriate panel. If a causative mutation can be identified in epilepsy, the index patient remarkably benefits from the analysis, because further diagnostic testing can be omitted, clinical management decisions revised, and antiepileptic drug therapy optimized.

For all the above-mentioned examples, exon amplification strategies are preferred to hybridization based target enrichment technologies to adapt gene panels to individual clinical questions.

Use of NGS for histocompatibility testing and immunogenetics

The selection of solid organ or hematopoietic cell transplant (HCT) donors is based on the compatibility of the HLA region. In solid organ transplantation, HLA-A, -B (low resolution typing) and -DRB1 (high resolution typing) matching is considered for the donor selection. In a HCT setting, matching for HLA-A, -B, -C, -DRB1, -DQB1 (all with high resolution testing) alleles is associated with superior survival. Recent studies have shown that every HLA mismatch contributes to a 9%–10% decrease in patient survival in unrelated transplant setting [32]. *HLA* genes contain 6–8 exons but only exons 2 and 3 of *HLA* class I genes (*HLA-A*, *-B* and *-C*), and exon 2 of *HLA* class II genes (*HLA-DRB*1, *-DQB1*) are regarded as clinically/biologically relevant and consequentially analyzed [33].

Even if a HLA-matched unrelated donor can be found, which can be achieved for up to 70% of patients of Caucasian origin [34], graft-versus-host disease (GvHD), an immune mediated reaction initiated by donor T cells in response to host alloantigen, remains a significant complication after HCT [35]. Clinical GvHD occurs when genetic differences between donor and recipient are sufficient to induce T cell activation. Next to the strongest histocompatibility antigens, *HLA* class I and class II genes mentioned above, there are many other genes, located throughout the genome and in particular in the major histocompatibility complex (MHC), which can initiate an antihost alloimmune reaction (Figure 2). More than 400 genes within the MHC region have immune-related functions and serve as potential candidates for developing a GvHD. By matching for extended HLA haplotypes, the incidence of GvHD can be lowered and transplant outcomes can be improved [36].

Histocompatibility antigen mismatches can lead to the opposite effects in the recipient. Development of GvHD is accompanied by decreased risk of relapse at the same time. This is due to the beneficial graft-versus-leukemia (GvL) effect of the transplant. Because both GvHD and GvL are caused by histocompatibility antigen mismatches, separating these opposite effects is a key clinical challenge to transplantation medicine [37]. This goal can be reached only by collecting more detailed information about variations inside the immune-related gene regions, preferably by complete sequencing of regions of interest.

NGS technology could become the method of choice in histocompatibility testing. Apart from the capability to analyze new target genes, *HLA* typing by itself can benefit from the new technology. The complexity and high variability of *HLA* genes makes *HLA* typing an ambitious effort for most laboratories. With increasing numbers of alleles, ambiguous typing results are more and more difficult to resolve. Furthermore, expansion of the sequenced region is inevitable in cases where alleles with reduced or no expression, arising from the mutations outside of exons 2 and 3, must be excluded. Currently, Sanger sequencing is still is the gold standard in HLA laboratories. However, it has many limitations that do not support the feasibility of quick and cost-effective typing of *HLA* genes. Clonal sequencing of single molecules by NGS allows a better if not maximum resolution of ambiguities, depending on the read lengths and extent of the sequenced region. Owing to extremely high variability of HLA alleles, read lengths of at least 250 nucleotides are needed for the resolution of ambiguities. Currently, only the Roche 454 GS FLX genome sequencer generates sequence read lengths greater than 250 nucleotides. First studies are published, which demonstrate that a rapid and accurate determination of *HLA* alleles is feasible by NGS. It allows for simultaneous typing of multiple HLA loci of many individuals in a single run [38].

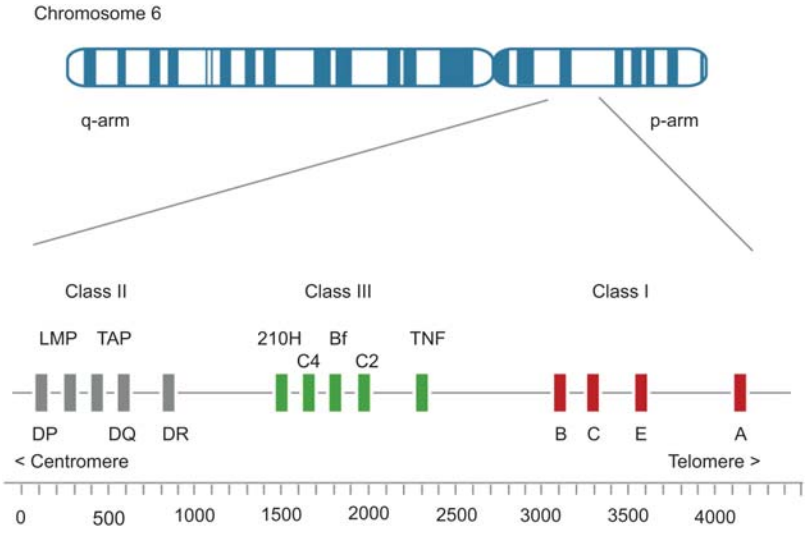

Figure 2 Ideograph of chromosome 6.
Schematic illustration of the genes of MHC in the region 6p21.1–6p21.3. Coding regions are represented by small boxes (gray, green, and red).

NGS could therefore become a valuable tool in research and diagnostics of autoimmune diseases. The HLA region has been associated with hundreds of human diseases, including many autoimmune diseases [39]. Unfortunately, for most of these diseases, underlying molecular mechanisms are still unknown. Some of the reasons include: (i) existence of extended linkage disequilibrium within the HLA region; (ii) HLA-associated diseases can be the result of a combination of different HLA molecules; and (iii) nearly all HLA-associated diseases are multifactorial polygenic diseases in which particular HLA allele(s), in combination with other genetic variants and environmental factors, is involved in disease susceptibility [40]. A cost-effective and rapid sequencing of gene regions of interest in the size of hundreds of kb and hundreds of patients and controls is definitely more promising in recognition of genetic background of autoimmune diseases compared to the typing for individual SNPs or sequencing for single genes. One of the first targets for extended sequencing could be the MHC as a whole, or specific regions inside the MHC with putative disease association.

References

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376–80. Erratum in: Nature 2006;441:120.
2. Stangier KA. Neue Sequenziertechnologien: ein kurzer Vergleich/Next-generation sequencing: a short comparison. J Lab Med 2009;33:267–70.
3. Cullen P, Hoffmann G, Klein H-G, Funke H. Next-generation sequencing und hochparallele Genexpressionsanalyse in der klinischen Diagnostik. J Lab Med 2010;34:349–356.
4. Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009;55:641–58.
5. Metzker ML. Sequencing technologies – the next generation. Nat Rev Genet 2010;11:31–46.
6. Fan HC, Quake SR. Detection of aneuploidy with digital polymerase chain reaction. Anal Chem 2007;79:7576–9.
7. Lo YM, Lun FM, Chan KC, Tsui NB, Chong KC, Lau TK, et al. Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc Natl Acad Sci USA 2007;104:13116–21.
8. Dennis Lo YM, Chiu RW. Prenatal diagnosis: progress through plasma nucleic acids. Nat Rev Genet 2007;8:71–7.
9. Korshunova Y, Maloney RK, Lakey N, Citek RW, Bacher B, Budiman A, et al. Massively parallel bisulphite pyrosequencing reveals the molecular complexity of breast cancer-associated cytosine-methylation patterns obtained from tissue and serum DNA. Genome Res 2008;18:19–29.
10. Meyer M, Stenzel U, Myles S, Prüfer K, Hofreiter M. Targeted high-throughput sequencing of tagged nucleic acid samples. Nucleic Acids Res 2007;35:e97.
11. Tewhey R, Warner JB, Nakano M, Libby B, Medkova M, David PH, et al. Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 2009;27:1025–31.
12. Introduction to the Fluidigm Access Array™ system. http://www.fluidigm.com/applications/access.html.
13. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, et al. Direct selection of human genomic loci by microarray hybridization. Nat Methods 2007;4:903–5.
14. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;4:907–9.
15. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522–7.
16. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182–9.

17. Tewhey R, Nakano M, Wang X, Pabón-Peña C, Novak B, Giuffre A, et al. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol 2009; 10:R116.
18. Craig DW, Pearson JV, Szelinger S, Sekar A, Redman M, Corneveaux JJ, et al. Identification of genetic variants using barcoded multiplexed sequencing. Nat Methods 2008;5:887–93.
19. Garber K. Fixing the front end. Nat Biotechnol 2008;26: 1101–4.
20. Lee H, O'Connor BD, Merriman B, Funari VA, Homer N, Chen Z, et al. Improving the efficiency of genomic loci capture using oligonucleotide arrays for high throughput resequencing. BMC Genomics 2009;10:646.
21. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 2009;10:R32.
22. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. N Engl J Med 2010;362: 1181–91.
23. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010;42:30–5.
24. Goossens D, Moens LN, Nelis E, Lenaerts AS, Glassee W, Kalbe A, et al. Simultaneous mutation and copy number variation (CNV) detection by multiplex PCR-based GS-FLX sequencing. Hum Mutat 2009;30:472–6.
25. Hoischen A, Gilissen C, Arts P, Wieskamp N, van der Vliet W, Vermeer S, et al. Massively parallel sequencing of ataxia genes after array-based enrichment. Hum Mutat 2010;31:494–9.
26. Chou LS, Liu CS, Boese B, Zhang X, Mao R. DNA sequence capture and enrichment by microarray followed by next-generation sequencing for targeted resequencing: neurofibromatosis type 1 gene as a model. Clin Chem 2010;56:62–72.
27. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA. Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 2010;100:184–92.
28. Callewaert B, Malfait F, Loeys B, De Paepe A. Ehlers-Danlos syndromes and Marfan syndrome. Best Pract Res Clin Rheumatol 2008;22:165–89.
29. von Kodolitsch Y, Rybczynski M, Bernhardt A, Mir TS, Treede H, Dodge-Khatami, et al. Marfan syndrome and the evolving spectrum of heritable thoracic aortic disease: do we need genetics for clinical decisions? Vasa 2010;39:17–32.
30. Ottman R, Hirose S, Jain S, Lerche H, Lopes-Cendes I, Noebels JL, et al. Genetic testing in the epilepsies – report of the ILAE Genetics Commission. Epilepsia 2010;51:655–70.
31. Berg AT, Berkovic SF, Brodie MJ, Buchhalter J, Cross JH, van Emde Boas W, et al. Revised terminology and concepts for organization of seizures and epilepsies: report of the ILAE Commission on Classification and Terminology, 2005–2009. Epilepsia 2010;51:676–85.
32. Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 2007;110:4576–83.
33. European Federation for Immunogenetics. Standards for Histocompatibility Testing. http://www.efiweb.eu/fileadmin/user_upload/pdf/Accreditation/version_5_6_1_Revision_Nomenclature.pdf. Accessed October 18, 2010.
34. Hirv K, Bloch K, Fischer M, Einsiedler B, Schrezenmeier H, Mytilineos J. Prediction of duration and success rate of unrelated hematopoietic stem cell donor searches based on the patient's HLA-DRB1 allele and DRB1-DQB1 haplotype frequencies. Bone Marrow Transplant 2009;44:433–40.
35. Hansen JA. Genomic and proteomic analysis of allogeneic hematopoietic cell transplant outcome. Seeking greater understanding the pathogenesis of GVHD and mortality. Biol Blood Marrow Transplant 2009;15(Suppl 1):e1–7.
36. Petersdorf EW, Malkki M, Gooley TA, Martin PJ, Guo Z. MHC haplotype matching for unrelated hematopoietic cell transplantation. PLoS Med 2007;4:e8.
37. Kawase T, Matsuo K, Kashiwase K, Inoko H, Saji H, Ogawa S, et al. HLA mismatch combinations associated with decreased risk of relapse: implications for the molecular mechanism. Blood 2009;113:2851–8.
38. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, Trachtenberg EA, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 2009;74:393–403.
39. Shiina T, Inoko H, Kulski JK. An update of the HLA genomic region, locus information and disease associations: 2004. Tissue Antigens 2004;64:631–49.
40. Caillat-Zucman S. Molecular mechanisms of HLA association with autoimmune diseases. Tissue Antigens 2009;73:1–8.