Application of high-throughput sequencing for studying genomic variations in congenital heart disease

Cornelia Dorn, Marcel Grunert* and Silke R. Sperling*

Advance Access publication date 3 October 2013

Abstract

Congenital heart diseases (CHD) represent the most common birth defect in human. The majority of cases are caused by a combination of complex genetic alterations and environmental influences. In the past, many disease-causing mutations have been identified; however, there is still a large proportion of cardiac malformations with unknown precise origin. High-throughput sequencing technologies established during the last years offer novel opportunities to further study the genetic background underlying the disease. In this review, we provide a roadmap for designing and analyzing high-throughput sequencing studies focused on CHD, but also with general applicability to other complex diseases. The three main next-generation sequencing (NGS) platforms including their particular advantages and disadvantages are presented. To identify potentially disease-related genomic variations and genes, different filtering steps and gene prioritization strategies are discussed. In addition, available control datasets based on NGS are summarized. Finally, we provide an overview of current studies already using NGS technologies and showing that these techniques will help to further unravel the complex genetics underlying CHD.

Keywords: *next-generation sequencing; congenital heart disease; sequence variations; variation filtering; whole-exome datasets; genomics*

INTRODUCTION

Over the last years, the application of automated Sanger sequencing and microarrays for genomic and genetic analyses has been increasingly replaced by next-generation sequencing (NGS) technologies. These high-throughput technologies are able to generate far more sequence data, in less time and with lower costs. This adds a particular advantage to many non-Mendelian diseases with a clear genetic component, where it has been a great challenge to identify the contributions made by single or even multiple genes. Doing so might permit the establishment of a profile for the disease that could be used for diagnostic purposes as well as predicting the likely outcome of particular therapeutic interventions. Using NGS, previously inaccessible insights into cognitive and neurological disorders, schizophrenia, cancer and cardiovascular diseases have been gained [1–5] and its application in clinical settings is increasingly being explored [6–9]. These technologies also open

Corresponding author. Silke R. Sperling, Department of Cardiovascular Genetics, Experimental and Clinical Research Center (ECRC), Charité–University Medicine Berlin and Max Delbrück Center (MDC) for Molecular Medicine, Lindenberger Weg 80, 13125 Berlin, Germany. Department of Biochemistry, Free University Berlin, Berlin, Germany. Tel.: +49-(0)30-450540123; Fax: +49-(0)30-84131699; E-mail: silke.sperling@charite.de

*These authors contributed equally to this work.

Cornelia Dorn holds a diploma in biology from the Humboldt University of Berlin and is a PhD student in the Department of Cardiovascular Genetics at the Charité and the MDC for Molecular Medicine. Her research focuses on omics of congenital heart disease.

Marcel Grunert received his PhD in bioinformatics from the Free University Berlin and is a post-doctoral researcher in the Department of Cardiovascular Genetics at the Charité and the MDC for Molecular Medicine. His primary research interest is the computational analysis of next-generation sequencing data.

Silke R. Sperling is a professor in cardiovascular genetics and biochemistry. She holds a doctoral degree in cardiac physiology and a habilitation in molecular biology and bioinformatics. She is head of the Cardiovascular Genetics at the ECRC jointly established between the Charité and the MDC for Molecular Medicine.

new opportunities for the study of cardiovascular development and complex human disorders like congenital heart disease (CHD). Here, we give an overview about the latest NGS technologies and provide a roadmap for study design and analysis of genomic CHD data. Furthermore, available control datasets and studies using NGS to investigate the genetics of congenital heart malformations are summarized.

CHD are the most common birth defect in human with an incidence of ~1% in all live births [10, 11]. For the United States it is estimated that ~760 000 individuals with CHD born in 1990 or later will be alive by the year 2020 [12]. In Germany, a prevalence of ~280 000 individuals with CHD in 2020 is expected [13]. CHD comprise a heterogeneous group of cardiac malformations that arise during heart development and the long-term clinical outcome after corrective surgery or intervention varies depending on the malformation as well as associated non-cardiac abnormalities [14]. Already decades ago, a multifactorial background of CHD with genetic–environmental interactions has been assumed [15]. A number of environmental influences during pregnancy are well-known to increase the risk of CHD, such as alcohol, teratogens and infectious agents [16–18] as well as common diseases like obesity and diabetes [19, 20]. Approximately 30% of cardiac malformations are part of syndromic disorders like Down syndrome, 22q11.2 deletion syndrome and Holt–Oram syndrome [21–23]; however, the majority of CHD occurs sporadically and does not follow Mendelian inheritance [15].

In the last decades studying familial cases using classical linkage analyses or performing candidate gene approaches based on knowledge gained in model organisms such as knockout mice have helped to gather major insights into the genetic background of CHD. Examples are families with atrial septal defect and conduction delay harboring mutations in the homeobox transcription factor *NKX2-5* [24], or a large family suffering from isolated septal defects related to a missense mutation in the transcription factor *GATA4* [25]. Based on knockout mouse data, the gene *CITED2* was analyzed using denaturing high-performance liquid chromatography (DHPLC) and direct sequencing, resulting in the identification of mutations in patients with different types of cardiac malformations [26]. Chromosomal aberrations including copy number variations (CNVs) can be identified by cytogenetic analysis including fluorescent *in situ* hybridization (FISH). For example, the majority of DiGeorge syndrome cases are caused by a chromosomal microdeletion (22q11) [27, 28]. However, some of the patients lack this deletion but harbor mutations in the T-box gene *TBX1* located in 22q11, displaying its important role for CHD [29]. Array comparative genomic hybridization (array CGH) offers a higher resolution for screening of submicroscopic chromosomal imbalances. The first studies using array CGH showed that ~17% of CHD patients harbor potentially disease-causing rare chromosomal aberrations [30, 31]. More recently, genome-wide SNP arrays [32] were used to identify copy number changes in sporadic CHD [33, 34]. To find single nucleotide polymorphisms (SNPs) associated with complex disorders, genome-wide association studies (GWAS) have been performed in large cohorts comprising hundreds to thousands of individuals [35]. The first studies on CHD identified loci associated with the risk of Tetralogy of Fallot (TOF) and septation defects [36–38].

Taken together, these studies have provided valuable insights into the genetics of CHD, as reviewed elsewhere [39–41]. However, there is still a large proportion of cardiac malformations for which no underlying cause could be identified. NGS techniques now provide a powerful novel approach to further elucidate the genetic background of CHD. They allow the simultaneous analysis of thousands of genes or even the whole genome in large patient cohorts. In contrast to microarrays, they are not dependent on DNA hybridization to preselected probes, which facilitates the identification of novel variations at a single-base resolution without *a priori* sequence information. Thus, they will enable the discovery of novel disease genes and networks.

Nevertheless, the identification of true disease-related genes is complicated by the huge amount of data that is generated. Large-scale population studies showed that a high number of potentially pathogenic variations can be observed in any healthy individual [42–44], with >95% being rare [42]. Using NGS, some variations identified to be disease causing in the past based on their exclusive occurrence in patients are now also found, albeit at very low frequencies, in healthy individuals and seem to be tolerated in the individual context. Thus, the application of novel sequencing approaches to the analysis of complex disorders like CHD remains challenging.

STUDY DESIGN

The current high-throughput sequencing technologies offer a variety of different study designs, which have to be considered carefully with regard to the scientific question being asked. The number of individuals selected for sequencing, the pooling of samples (multiplexing), the number of selected target bases, the choice of the sequencing platform as well as the desired read depth and length determine the costs and the major bottlenecks for research projects.

One important aspect of study design is the selection of individuals for sequencing. Depending on the research question, the availability of samples and costs, one might focus on families (e.g. trios), unrelated individuals or cases with extreme phenotypes. For small cohorts, the selection of well-defined, homogenous (sub)phenotypes can increase the significance of the study, as has been shown for apical hypertrophic cardiomyopathy [45]. However, large consortia now enable the analysis of hundreds of patients, which represent diverse CHD phenotypes. Using barcodes for multiplexing allows the simultaneous sequencing of a larger number of samples and thus reduces the time and costs for data generation when analyzing large cohorts.

Another crucial step is the choice between whole-genome sequencing (WGS), whole-exome sequencing and targeted resequencing. They all have their individual strengths and limitations and are suitable for different scientific questions. WGS allows gaining a broad understanding of the full range of genomic variations including e.g. enhancers and promoters. However, due to high costs and time needed to achieve an adequate read depth, WGS is not feasible for many studies. Thus, whole-exome and targeted resequencing approaches have been established as an alternative. Whole-exome sequencing enables the sequencing of almost all protein-coding regions, often combined with a high coverage. For rare inherited disorders, it has been shown that focusing on the exome is reasonable because the majority of mutations responsible for Mendelian diseases affect protein-coding sequences [46]. If knowledge about possible candidate genes and disease pathways is already available, the targeted resequencing of selected regions is a promising option. To select genomic regions for targeted resequencing, data from previous projects like sequencing analyses, GWAS studies, animal models as well as publicly available databases and other web resources can be used. In addition, gene prioritization tools can be employed to narrow down the list of genes of interest [47, 48], which enables their analysis in a much larger cohort of patients and controls. For CHD, the CHDWiki offers a repository of current knowledge on the genetic basis of the disease [49]. However, due to the constantly decreasing sequencing costs, whole-exome sequencing should be applied if possible, because it is not limited to the selection of genes. Thus, the more comprehensive data allow the discovery of novel disease pathways and can be used for subsequent projects. Both exome and targeted resequencing require sequence enrichment technologies like array-based sequence capturing. Care should be taken when comparing different datasets, because the use of different enrichment techniques can lead to differences in the captured regions ranging from whole genes down to single bases.

In addition to identifying the genomic positions and nucleotide changes of a wide range of alterations, recent advances in sequencing technologies also enable the independent determination of both haplotype sequences of individual genomes. This phase information, i.e. the separation of maternally and paternally derived sequences, are important for understanding gene function and disease [50]. The successful application of this approach has been demonstrated in several studies [51–53]. It allows the discrimination of *cis* and *trans* configurations of mutations and can provide valuable insights into disease mechanisms like compound heterozygosity.

Next-generation sequencing platforms

Once a decision about the samples and genomic regions to be sequenced has been made, the next step is to select a sequencing platform. NGS technologies are evolving rapidly and during the last years several platforms were released. Although they differ in their biochemistry, all follow the principle of cyclic-array sequencing, where an array of DNA features is iteratively enzymatically sequenced combined with imaging-based data detection [54].

Recently, the HiSeq 2000/2500 instrument (Illumina), GS FLX+ system (Roche/454) and SOLiD 5500/5500xl Wildfire system (Life Technologies) have set the standard for high-throughput sequencing. There are differences between these platforms resulting in specific advantages and disadvantages (Figure 1), which also have to be taken into account when comparing different datasets. In addition to the standard high-throughput sequencing

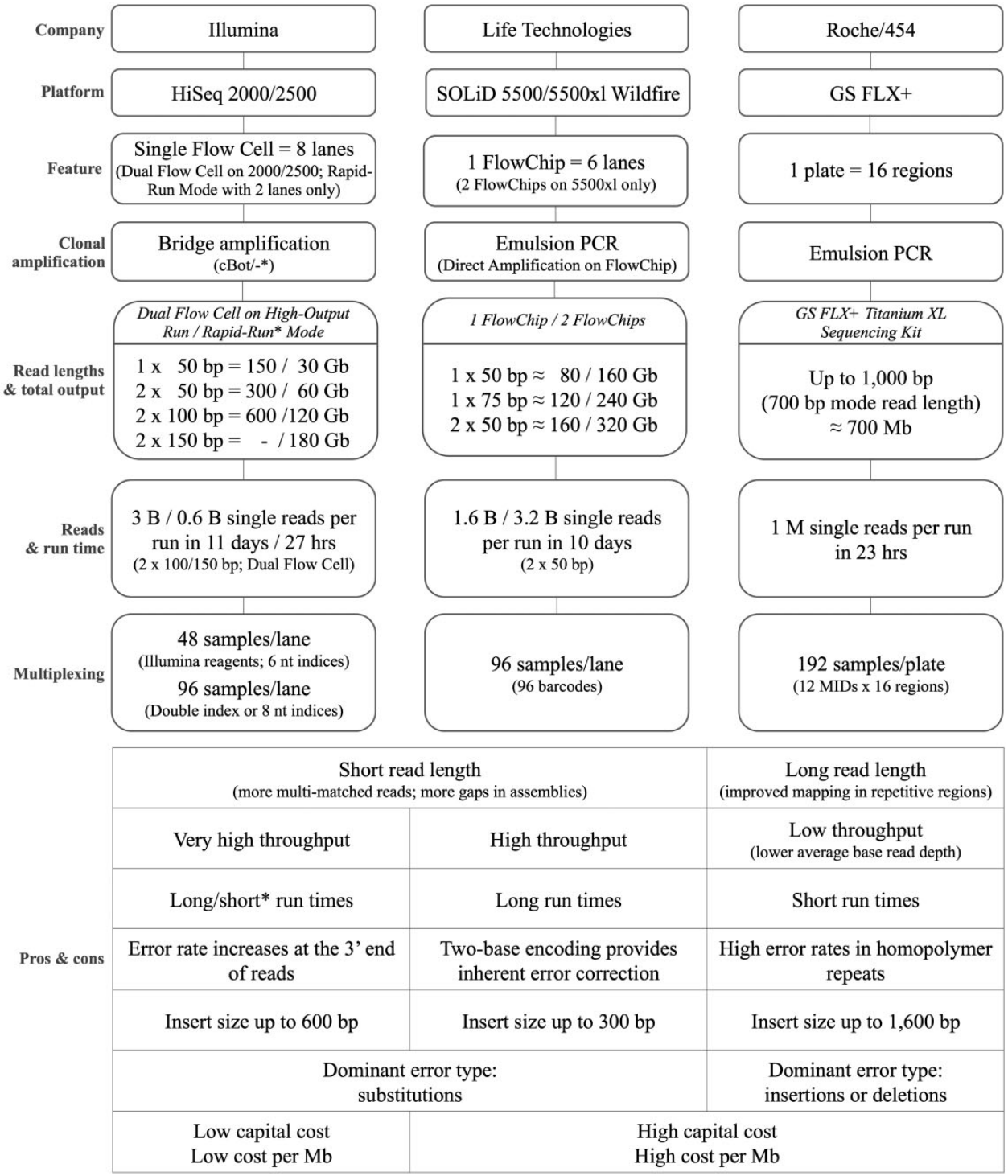

Figure 1: NGS platforms. Overview of the three most common high-throughput sequencing platforms currently available. The information provided are based on company sources alone. B, Billion; bp, base pairs; hrs, hours; Gb, Giga bases; Mb, Mega bases; nt, nucleotides; PCR, polymerase chain reaction. *Illumina HiSeq 2500 only.

platforms, three benchtop platforms have been released, envisaged for smaller laboratories and the clinical diagnostic market [55]. The MiSeq (Illumina), 454 GS Junior (Roche/454) and Ion Torrent PGM/Ion Proton (Life Technologies) are lower throughput fast-turnaround instruments, which need much less instrument space and offer less set-up [55, 56].

In general, a higher number of sequence reads from NGS results in greater sequencing depth and thus in higher sequence confidence. For example, the different phases within the 1000 Genomes Projects range from low coverage (2–6×) for whole-genome sequence data to high coverage (50–100×) for exome sequence data [57]. Moreover, the overall accuracy and specific error distribution (e.g. tendency for systematic errors) of the different technologies have to be considered [54].

IDENTIFICATION OF GENOMIC VARIATIONS

Single nucleotide variations (SNVs) represent the most abundant type of genomic variation, followed by short (<50 bases) insertions and deletions (InDels), summarized as local variations. Common SNVs (SNPs) occur in >1% of a population. InDels are both less frequent and subjected to a stronger purifying selection compared with SNVs because they create larger changes in coding regions such as frameshifts and insertions/deletions of amino acids. In contrast, SNVs often produce synonymous changes with less or no impact on gene function [58].

After sequencing, the first preprocessing step is the quality assessment of the raw sequence reads. Several tools are available for this purpose, including FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html) [59], which all can also be used for the handling of NGS data in general.

The method of choice for the identification of local variations is the mapping of sequence reads to a known reference genome (alignment-consensus approach). Many algorithms have been developed specifically for this purpose (e.g. Bowtie 2, BWA, RazerS 3, SOAP3 [60–63]), which has been reviewed elsewhere [64].

After taking sequencing and alignment problems into account (e.g. using the GATK realignment [65]), several SNV and InDel calling tools can be used, such as mpileup (samtools), GATK, VarScan2, SOAPsnp, SOAPindel and Pindel [65–71], as reviewed elsewhere [72, 73]. From the computational perspective, algorithms for the detection of SNVs are much more advanced than for the detection of InDels, partially due to the difficulties of detecting InDels in relatively short sequence reads. However, having long reads from NGS does not necessarily help to find true InDels. The platform from Roche/454 produces reads with a length of up to 1000 bp (Figure 1) but tends to identify many false InDels because of its problem to correctly assess the length of homopolymer repeats, resulting in over- and undercalls [74].

CNVs are much larger genetic alterations (up to millions of DNA bases). There are four main computational methods for detecting copy numbers from NGS data, namely read-depth, read-pair, split-read and assembly-based methods. Assembly-based approaches perform best for smaller genomes and are less widely used for the human genome because the assembly in repeat regions is difficult with short read lengths [75]. Split-read methods (e.g. Pindel [71]) can detect deletions and small insertions at single base pair resolution, thus defining the exact breakpoint [32]. They were first applied to longer reads from Sanger sequencing [76] and they are currently used to identify rearrangement points in the long sequence reads from Roche/454 (Roche GS Reference Mapper). Read-pair approaches (e.g. PEMer, BreakDancer, VariationHunter [77–79]) consider the span and orientation between two pairs of reads (paired-end) [32] but they are limited by the insert size when detecting insertions (see 'Pros and Cons' in Figure 1) [80]. Read-depth methods (e.g. mrCaNaVar and CNVnator [81, 82]) assume that the mapped reads are randomly distributed across the reference genome or targeted regions. They investigate differences from the expected read distribution to detect duplications (higher read depth) and deletions (reduced read depth) [32].

IDENTIFICATION OF CANDIDATE GENES

Genes affected by raw local variations must be reduced to potential disease-causing genes, which are candidates for further downstream analyses. This includes the filtering of all local variations for functional relevance and frequency in control datasets as well as the gene prioritization process and the validation of related variations.

Filtering of local variations

Variation calling often results in false positives and negatives resulting from technical bias (duplicate reads, strand and GC bias), sequencing errors

(e.g. increased error probability at the 3′-end of Illumina reads and at homopolymer repeats in Roche/454 sequencing) and alignment artifacts in low mappability regions [32]. The variation calling methods already try to minimize the number of false positives. However, to further reduce these errors and to identify functionally relevant variations additional filtering steps have to be applied.

First, variations should be filtered by the sequencing quality including the read depth (coverage), number of supporting reads, average base quality (e.g. Phred score ≥ 20), supporting strands (i.e. forward and/or reverse) and variation allele frequency (step 1 in Figure 3). Variations with an allele frequency <0.2 should be discarded, while frequencies between 0.2 and 0.8 are called heterozygous and those >0.8 are considered as homozygous [83]. Moreover, variations with an excessively high coverage are usually caused by structural variations like CNVs or other alignment artifacts.

The next filtering step is the annotation and functional characterization of the variations (step 2 in Figure 3). Several tools are available to annotate variations from NGS data such as SeattleSeq (http://snp.gs.washington.edu/SeattleSeqAnnotation/), ANOVA, VAT, F-SNP and snpEff [84–87]. To identify putatively deleterious SNVs, functional prediction methods (e.g. PolyPhen-2, SIFT, MutationTaster and a likelihood ratio test [88–91]) and conservation-based methods (e.g. PhastCons, GERP++, PhyloP, SCONE [92–95]) can be applied. Using multiple methods can help to obtain more reliable functional predictions and thus, to focus on the most likely relevant variations [96]. Over all, variations not predicted to be damaging, nonsense, frame-shifting or inserting/deleting amino acids as well as variations not affecting splice sites, non-coding RNAs (e.g. seeds of microRNAs) or other regulatory regions (e.g. promoter or enhancer) might be discarded. An overview of the different types of genomic annotations and functional characterizations of local variations is given in Figure 2.

The retained variations can subsequently be reduced to novel variations or rare variations with a minor allele frequency (MAF) of ≤ 0.01 in the dbSNP database [97] or other public datasets, which will be presented in the next section. Rare and *de novo* mutations are the main genetic cause for CHD and thus, filtering for rare variations or variations not present in dbSNP can reduce the search space by 2- to 10-fold while retaining the most promising candidates. However, known disease-associated variations present for example in the OMIM, KEGG disease, HGMD and ClinVar database [98–102] or the CHDWiki [49] might be retained (step 3 in Figure 3).

AVAILABLE CONTROL DATASETS

Several datasets and databases are available that can be used to filter for rare variations (step 3 in Figure 3). To catalog short genomic variations dbSNP was established [97]. The database summarizes data from various projects using different genotyping methods. Examples for contributing large-scale projects are the HapMap Project, the 1000 Genomes Project and the ClinSeq Study (CSAgilent) [57, 103–105]. Of course, dbSNP is not always complete and accurate, as it contains false positives and might be contaminated by single nucleotide differences arising from paralogous sequences in the genome [106, 107].

As an alternative or in addition to dbSNP, separate datasets can also be used as controls. Currently, several NGS datasets of large cohorts are being generated and are already partly available. The 1000 Genomes Project aims to sequence the genomes of 2500 individuals of European, East Asian, West African, American and South Asian ancestry using a combination of low-coverage WGS (2–6×), targeted deep exome sequencing (50–100×) and dense SNP genotyping [57, 104]. The Exome Sequencing Project of the National Heart, Lung and Blood Institute (NHLBI) provides whole-exome sequencing data of 6503 individuals from multiple cohorts to study heart, lung and blood disorders [42, 96]. The ClinSeq Study focuses on cardiovascular health and aims to recruit >1500 participants. A total of 662 participants of European descent have already undergone whole-exome sequencing and the data are publicly available [9, 105].

Li and colleagues analyzed 200 healthy Danish individuals by whole-exome sequencing at low coverage. This dataset also contains individual genotypes in addition to the accumulated frequency for each variation [44]. The project ‘Genome of the Netherlands’ (GoNL) aims to capture the genetic variation present in the Dutch population and has performed WGS of 769 individuals belonging to 250 families with two parents and one or two children [108].

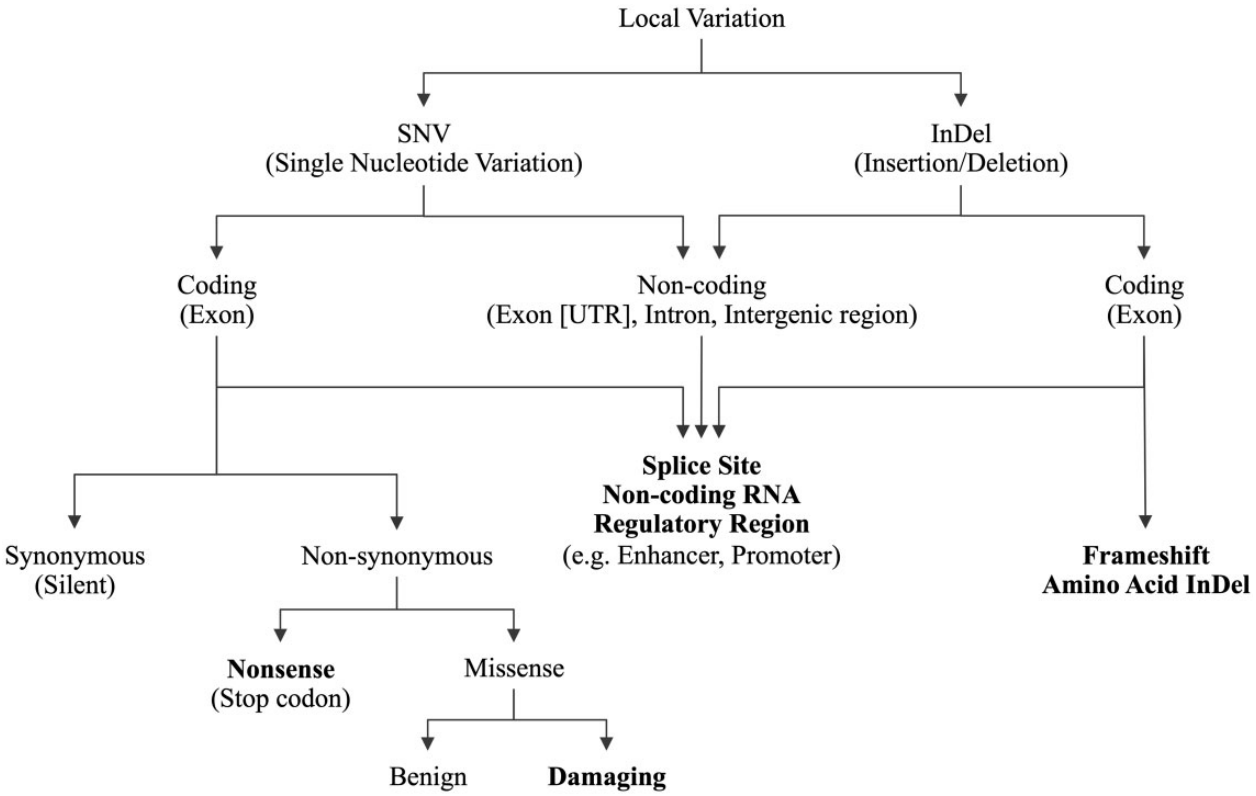

Figure 2: Genomic annotations and functional characterizations of local variations. The main functionally relevant types of variations are marked in bold. UTR, untranslated region.

The UK10K project is currently performing WGS for 4000 individuals (including twin pairs) and whole-exome sequencing for 6000 individuals showing disease phenotypes in the three large groups obesity, neurodevelopmental disorders and rare diseases including 125 CHD cases. The sequencing data are already partly available through the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) at the EBI. In general, NGS datasets can also be obtained from the database of Genotypes and Phenotypes (dbGaP; http://www.ncbi.nlm.nih.gov/gap) [109] at the NCBI. A summary of the described datasets is given in Table 1. When analyzing trio samples or larger families, a further opportunity is to use healthy family members as additional controls and to filter for *de novo* mutations, which will be discussed in the following section.

In general, the selection of a suitable control dataset mainly depends on the technical comparability and the individuals selected for sequencing. The control and the study cohort should preferably be enriched with the same technique and sequenced with the same platform. Moreover, the control individuals should belong to the same ethnical group as the studied cases and phenotypic information of the controls should also be considered. For example, the individuals sequenced within the NHLBI project belong to multiple disease groups like atherosclerosis, asthma and cystic fibrosis.

Gene prioritization

After filtering for rare deleterious variations using prediction tools and control datasets, the resulting affected genes can be prioritized to identify the most likely disease-related genes and to further reduce the number of candidate genes for downstream studies (step 4 in Figure 3).

Automated gene prioritization approaches require prior knowledge about the disease and associated genes and pathways. They integrate diverse data including protein–protein interactions, animal models, coexpression, gene ontologies (e.g. GO [110]), sequence homologies and literature co-occurrences and include tools like GeneSeeker and Endeavour [111, 112]. They can link the candidate genes to known disease-associated genes and generate a ranked list, with the most promising candidates at the top [47, 48].

Pathway databases and analysis tools, which can also be used for the gene prioritization process by

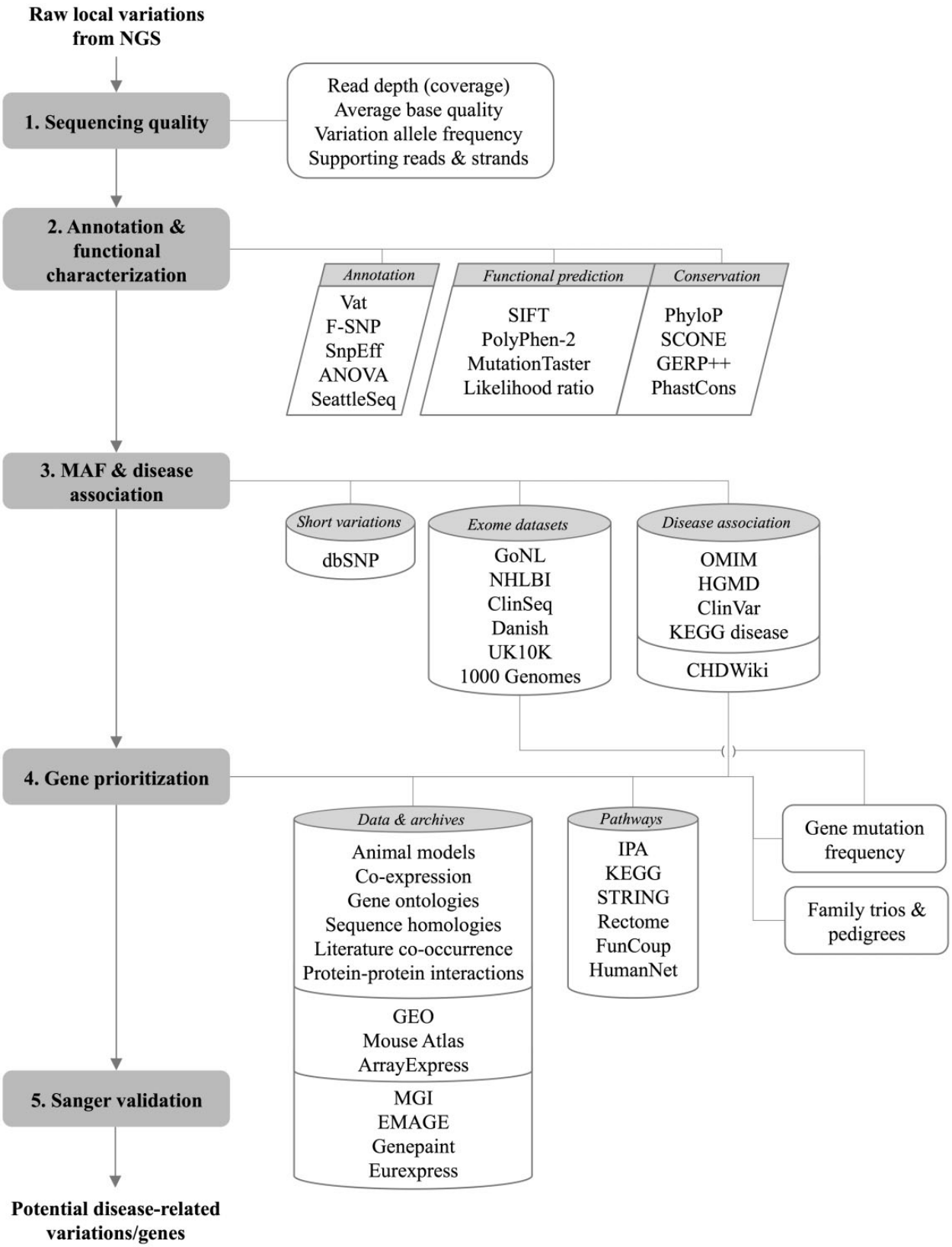

Figure 3: Identification of candidate genes. The individual filtering steps from raw local variations to potential disease-related variations and genes are shown including different data sources, tools and approaches that can be used.

Table 1: Available NGS control datasets

Dataset	Cohort Size	Cohort ancestry	Data	Exome capture	NGS Platform	URL	References
1000 Genomes Pilot	697	7 populations (European, Africian and East Asian)	SNVs; InDels	Exons of 906 genes using mainly NimbleGen 385 K capture array	Genome Analyzer IIx (Illumina); 454 GS FLX/ Titanium (Roche)	www.1000genomes.org (Pilot 3 study)	[104]
1000 Genomes	1092	14 populations (European, East Asian, sub-Saharan African and American)	SNVs; InDels; SVs	Whole exome using NimbleGen SeqCap EZ Human Exome Library VI & V2 (Roche) and SureSelect All Exon V2 Target Enrichment (Agilent)	Genome Analyzer IIx (Illumina); SOLiD (Life Technologies)	www.1000genomes.org (Phase I)	[57]
	2500*	25 populations (European, African, American, East and South Asian)	SNVs; InDels; SVs	Whole genome and whole exome	–	www.1000genomes.org (Phase 3)	–
NHLBI	6503	4300 European American and 2203 African American individuals	SNV; InDels	Whole exome using NimbleGen capture (Roche)	Genome Analyzer IIx; HiSeq 2000 (Illumina)	http://evs.gs.washington.edu/EVS/	[42, 96]
ClinSeq	662	European American individuals	SNVs	Whole exome using SureSelect Human All Exon 38 or 50 Mb kit (Agilent Technologies)	Genome Analyzer IIx (Illumina)	http://www.genome.gov/20519355 (Data available from dbSNP; id=CSAgilent)	[9]
	1500*	European American and Africian American individuals	SNV; InDels	Whole exome	–		
Danish	200	Danish individuals	SNVs	Whole exome using NimbleGen 2.1 M HD array (Roche)	Genome Analyzer IIx (Illumina)	http://soap.genomics.org.cn/soapsnp.html (SOAP website)	[44]
GoNL	769	Dutch individuals	SNVs; InDels; SVs	Whole exome	HiSeq 2000 (Illumina)	http://www.nlgenome.nl (SNV data currently available for 498 parents)	–
UKI0K	10 000	British (4000 healthy and 6000 disease-affected individuals)	SNVs; InDel; SVs	Whole genome and whole exome using SureSelect Human All Exon 50-Mb kit (Agilent Technologies), respectively	Genome Analyzer IIx; Hiseq 2000 (Illumina)	http://www.uki0k.org (data available at EGA)	–

EGA, European Genome-phenome Archive; InDels, insertions/deletions; *denotes the overall size of the study including all sub-cohorts.

placing candidate genes into the context of known molecular pathways, include KEGG, STRING, HumanNet, FunCoup, Reactome [99, 100, 113–116] or the commercial Ingenuity Pathway Analysis (IPA; www.ingenuity.com/). For Mendelian diseases the human gene connectome (HGC) approach was recently introduced [117].

Since CHD is a developmental disorder, the genes causing it must be functional during heart development. Moreover, the gene expression in adult heart might be important regarding the long-term clinical outcome. Thus, cardiac expression might be used as a further criterion for prioritizing candidate genes. Expression can be measured by e.g. quantitative PCR, serial analysis of gene expression (SAGE), expression array or RNA-seq, which was recently used in a high-throughput screen of CHD candidate genes [118]. Already published data can be retrieved from e.g. Mouse Atlas, ArrayExpress, Bgee and GEO [119–122]. For individual genes, literature research for expression data can be guided by e.g. MGI (Mouse Genome Informatics; http://www.informatics.jax.org/). Moreover, spatial mouse expression data based on *in situ* hybridization can be obtained from the EMAGE, Genepaint and Eurexpress databases [123–125].

When analyzing family trios, one approach is to further filter for genes with *de novo* mutations, assuming that a unique mutation is causing CHD in the offspring [118]. In addition, large CHD pedigrees offer the opportunity to analyze Mendelian inheritance of disease-causing mutations. Another approach is to consider the combination of variations in different genes and the mutation frequencies of the genes, assuming an oligo- or multigenic background with disease-related genes more often affected by deleterious mutations in cases compared to controls.

Validation of NGS results

Today's NGS platforms generate highly accurate data, as shown by validation studies using Sanger sequencing. When using a high coverage threshold for variation calling (≥30×), ∼100% of variations can be confirmed [126, 127]. However, when selecting for rare variations overrepresented in a cohort, one also runs the risk of enriching false-positive variations that result from characteristic sequence features or mapping problems at a specific position and thus are likely to occur in several individuals.

In a study on TOF patients, 35 variations in 20 genes with a significantly higher mutation frequency in cases compared with controls were initially identified. Four of these variations were found in multiple patients. Using Sanger sequencing, seven variations (20%) could not be validated, including the four variations that were detected in more than one patient. Nevertheless, comparison to related RNA-seq data showed that 94% of variations covered at least 10× could be confirmed and thus demonstrated a high sequencing quality (unpublished data). Notably, true-positive variations could still be missed by RNA-seq due to allelic expression, which is a widespread phenomenon that can be mediated through mechanisms like alternative mRNA processing or differential transcription factor binding [128–130]. Taken together, validation of variations detected by NGS sequencing should additionally be performed as one of the last filtering steps (step 5 in Figure 3).

HIGH-THROUGHPUT SEQUENCING STUDIES ON CHD

To date, only few CHD studies based on NGS have been published. One combined approach of whole-exome sequencing, high-resolution melting analysis and direct DNA sequencing of selected genes identified possible disease-causing mutations in a family with heterogeneous CHD [131]. Whole-exome sequencing of one heterotaxy patient with CHD could identify a recessive missense mutation in *SHROOM3*. Subsequent screening of 96 heterotaxy patients using Sanger sequencing identified four additional cases with rare variants in the gene, suggesting a role of *SHROOM3* in left–right patterning [132]. Another application of NGS identified a dominant missense mutation causing the rare Cantú syndrome, which includes cardiac manifestations. Here, whole-exome sequencing of the index patient and his unaffected parents was performed, identifying a single *de novo* missense mutation in the potassium channel gene *ABCC9*. Subsequently, missense mutations in the *ABCC9* gene could be detected in 13 of 15 additional cases and functional studies showed that the mutations lead to dominant channel opening [133].

Studying large cohorts of CHD patients using high-throughput sequencing will hopefully lead to a better understanding of the complex genetics underlying the disease. One example is the Congenital Heart Disease Genetic Network Study established by the Pediatric Cardiac Genomics Consortium, which enrolled >3700 patients representing a diverse

range of congenital heart defects. The study aims to investigate the relationships between genetic factors, clinical features and outcomes in CHD patients. Medical data and biospecimen were collected and ongoing studies include the identification of CNVs, the resequencing of candidate genes as well as the search for somatic mutations and skewed allelic expression using whole-exome sequencing and RNA sequencing, respectively, from cardiac tissue samples [134]. For a subset of 362 patients and their parents whole-exome sequencing from venous blood DNA has already been completed. Most interestingly, this study revealed an accumulation of *de novo* mutations in histone-modifying genes in CHD cases, underlining the important role of epigenetic regulation in heart development [118].

Another large-scale project is the Deciphering Developmental Disorders (DDD) study (http://www.ddduk.org/) headed by the Wellcome Trust Sanger Institute, which aims to collect clinical data and DNA samples from 12 000 undiagnosed children with developmental disorders and their parents. The study also includes CHD patients and uses high-resolution array CGH, SNP genotyping, and whole-exome sequencing to identify the genetic causes underlying the diverse disorders [135]. Finally, the UK10K project (http://www.uk10k.org/) is performing whole-exome sequencing for 125 CHD patients enclosed in its rare disease sample set. Both studies are still ongoing and so far, no results on CHD have been published.

CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The heart is the first organ that functions during embryonic development, and congenital cardiac malformation are the most common birth defect in human. CHDs represent a heterogeneous group of disorders with a complex genetic background. Although many disease-causing genetic alterations have been identified, there is still a large proportion of CHD with unknown precise origin. During the last years, high-throughput sequencing technologies were established, which are still rapidly developing. These NGS techniques offer novel opportunities to further study the genetics underlying congenital heart malformations. Furthermore, high-throughput sequencing as well as many of the tools and databases described here can also be applied to a wide range of other complex diseases.

Besides the analysis of genomic variations, NGS can be used for studying genetic and epigenetic alterations such as RNA and small RNA expression, alternative splicing, DNA methylation and protein–DNA interactions. Individual NGS datasets can already provide a wealth of information. However, the combination of genomic, genetic and epigenetic as well as proteomic and metabolic data in a systems biology approach enables a more comprehensive understanding of disease processes [136–138]. Just recently, exome sequencing data of CHD patients were linked to gene expression data from mouse to filter for potentially disease-causing mutations [118].

Moreover, considering sequence variations in regulatory regions like enhancers or promoters can lead to valuable insights into regulatory changes underlying a disease. These variations might disrupt the assembly of the transcription machinery or change transcription factor binding affinities [139]. Combining such findings with corresponding expression data can show the functional consequences of observed variations. To date, several computational and experimental tools are available to assess the pathogenicity of variations in regulatory elements and numerous examples for disease associations have been identified [140]. In a patient suffering from ventricular septal defect, a homozygous variation in the *TBX5* enhancer could be shown to abrogate the gene's expression in the heart [141].

To analyze the effect of individual mutations in combination with the complex genetic background, the differentiation of patient-specific induced pluripotent stem cells might be a valuable approach. This strategy was used to model several cardiac phenotypes [142–144] and can prospectively be used for drug discovery and development [145]. Hopefully, a better understanding of the causes underlying cardiac malformation will enable the development of novel therapeutic and preventive strategies in the future.

Key points

- NGS technologies offer novel opportunities to study complex genetic disorders like congenital heart disease.
- The huge amount of data generated with high-throughput sequencing requires a sophisticated study design and analysis to identify potentially disease-related variations and genes.
- There are several large-scale projects providing sequence information for control cohorts comprising hundreds to thousands of individuals.
- Current studies already using NGS technologies were able to gain new insights into the genetic background of CHD.

FUNDING

This work was supported by the European Community's Seventh Framework Programme contracts ('CardioGeNet') 2009-223463 and ('CardioNet') People-2011-ITN-289600 (all to S.R.S), a PhD scholarship to C.D. by the Studienstiftung des Deutschen Volkes, and the German Research Foundation (Heisenberg professorship and grant 574157 to S.R.S.).

Acknowledgements

We thank Andreas Perrot and Kerstin Schulz for discussion and review of the manuscript. We apologize to those who have made contributions to this field of research and are not cited in this review.

References

1. Najmabadi H, Hu H, Garshasbi M, *et al*. Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature* 2011;**478**(7367):57–63.
2. Lupski JR, Reid JG, Gonzaga-Jauregui C, *et al*. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 2010;**362**(13):1181–91.
3. Kong A, Frigge ML, Masson G, *et al*. Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012;**488**(7412):471–5.
4. Puente XS, Pinyol M, Quesada V, *et al*. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 2011;**475**(7354):101–5.
5. Herman DS, Lam L, Taylor MRG, *et al*. Truncations of titin causing dilated cardiomyopathy. *N Engl J Med* 2012; **366**(7):619–28.
6. Berg JS, Evans JP, Leigh MW, *et al*. Next generation massively parallel sequencing of targeted exomes to identify genetic mutations in primary ciliary dyskinesia: implications for application to clinical testing. *Genet Med* 2011;**13**(3):218–29.
7. Welch JS, Westervelt P, Ding L, *et al*. Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. *JAMA* 2011;**305**(15):1577–84.
8. Ashley EA, Butte AJ, Wheeler MT, *et al*. Clinical assessment incorporating a personal genome. *Lancet* 2010; **375**(9725):1525–35.
9. Biesecker LG. Opportunities and challenges for the integration of massively parallel genomic sequencing into clinical practice: lessons from the ClinSeq project. *Genet Med* 2012; **14**(4):393–8.
10. Hoffman JIE, Kaplan S. The incidence of congenital heart disease. *J Am Coll Cardiol* 2002;**39**(12):1890–900.
11. Reller MD, Strickland MJ, Riehle-Colarusso T, *et al*. Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *J Pediatr* 2008;**153**(6):807–13.
12. Webb CL, Jenkins KJ, Karpawich PP, *et al*. Collaborative care for adults with congenital heart disease. *Circulation* 2002;**105**(19):2318–23.
13. National Register for Congenital Heart Defects. http://www.kompetenznetz-ahf.de/en/research/register-biobank/ (20 May 2013, date last accessed).
14. Michielon G, Marino B, Formigari R, *et al*. Genetic syndromes and outcome after surgical correction of tetralogy of Fallot. *Ann Thorac Surg* 2006;**81**(3):968–75.
15. Nora JJ. Multifactorial inheritance hypothesis for the etiology of congenital heart diseases. The genetic-environmental interaction. *Circulation* 1968;**38**(3):604–17.
16. Zhu H, Kartiko S, Finnell RH. Importance of gene-environment interactions in the etiology of selected birth defects. *Clin Genet* 2009;**75**(5):409–23.
17. Kopf PG, Walker MK. Overview of developmental heart defects by dioxins, PCBs, and pesticides. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 2009; **27**(4):276–85.
18. Dewan P, Gupta P. Burden of Congenital Rubella Syndrome (CRS) in India: a systematic review. *Indian Pediatr* 2012;**49**(5):377–99.
19. Watkins ML, Rasmussen SA, Honein MA, *et al*. Maternal obesity and risk for birth defects. *Pediatrics* 2003;**111**(5 Pt 2):1152–8.
20. Loffredo CA, Wilson PD, Ferencz C. Maternal diabetes: an independent risk factor for major cardiovascular malformations with increased mortality of affected infants. *Teratology* 2001;**64**(2):98–106.
21. Antonarakis SE, Lyle R, Dermitzakis ET, *et al*. Chromosome 21 and down syndrome: from genomics to pathophysiology. *Nat Rev Genet* 2004;**5**(10):725–38.
22. Momma K. Cardiovascular anomalies associated with chromosome 22q11.2 deletion syndrome. *Am J Cardiol* 2010;**105**(11):1617–24.
23. Basson CT, Bachinsky DR, Lin RC, *et al*. Mutations in human TBX5 [corrected] cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet* 1997;**15**(1):30–5.
24. Schott JJ, Benson DW, Basson CT, *et al*. Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science* 1998;**281**(5373):108–11.
25. Garg V, Kathiriya IS, Barnes R, *et al*. GATA4 mutations cause human congenital heart defects and reveal an interaction with TBX5. *Nature* 2003;**424**(6947):443–7.
26. Sperling S, Grimm CH, Dunkel I, *et al*. Identification and functional analysis of CITED2 mutations in patients with congenital heart defects. *Hum Mutat* 2005;**26**(6):575–82.
27. Driscoll DA, Salvin J, Sellinger B, *et al*. Prevalence of 22q11 microdeletions in DiGeorge and velocardiofacial syndromes: implications for genetic counselling and prenatal diagnosis. *J Med Genet* 1993;**30**(10):813–7.
28. Yamagishi H, Srivastava D. Unraveling the genetic and developmental mysteries of 22q11 deletion syndrome. *Trends Mol Med* 2003;**9**(9):383–9.
29. Yagi H, Furutani Y, Hamada H, *et al*. Role of TBX1 in human del22q11.2 syndrome. *Lancet* 2003;**362**(9393): 1366–73.
30. Thienpont B, Mertens L, de Ravel T, *et al*. Submicroscopic chromosomal imbalances detected by array-CGH are a frequent cause of congenital heart defects in selected patients. *Eur Heart J* 2007;**28**(22):2778–84.
31. Erdogan F, Larsen LA, Zhang L, *et al*. High frequency of submicroscopic genomic aberrations detected by tiling path array comparative genome hybridisation in patients with

isolated congenital heart disease. *J Med Genet* 2008;**45**(11):704–9.

32. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011;**12**(5):363–76.
33. Greenway SC, Pereira AC, Lin JC, *et al*. De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot. *Nat Genet* 2009;**41**(8):931–5.
34. Soemedi R, Wilson IJ, Bentham J, *et al*. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am J Hum Genet* 2012;**91**(3):489–501.
35. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;**363**(2):166–76.
36. Cordell HJ, Töpf A, Mamasoula C, *et al*. Genome-wide association study identifies loci on 12q24 and 13q32 associated with Tetralogy of Fallot. *Hum Mol Genet* 2013;**22**(7):1473–81.
37. Cordell HJ, Bentham J, Töpf A, *et al*. Genome-wide association study of multiple congenital heart disease phenotypes identifies a susceptibility locus for atrial septal defect at chromosome 4p16. *Nat Genet* 2013;**45**(7):822–4.
38. Hu Z, Shi Y, Mo X, *et al*. A genome-wide association study identifies two risk loci for congenital heart malformations in Han Chinese populations. *Nat Genet* 2013;**45**(7):818–21.
39. Bruneau BG. The developmental genetics of congenital heart disease. *Nature* 2008;**451**(7181):943–8.
40. Blue GM, Kirk EP, Sholler GF, *et al*. Congenital heart disease: current knowledge about causes and inheritance. *Med J Aust* 2012;**197**(3):155–9.
41. Fahed AC, Gelb BD, Seidman JG, Seidman CE. Genetics of congenital heart disease: the glass half empty. *Circ Res* 2013;**112**(4):707–20.
42. Tennessen JA, Bigham AW, O'Connor TD, *et al*. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 2012;**337**(6090):64–9.
43. Marth GT, Yu F, Indap AR, *et al*. The functional spectrum of low-frequency coding variation. *Genome Biol* 2011;**12**(9):R84.
44. Li Y, Vinckenbosch N, Tian G, *et al*. Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 2010;**42**(11):969–72.
45. Arad M, Penas-Lado M, Monserrat L, *et al*. Gene mutations in apical hypertrophic cardiomyopathy. *Circulation* 2005;**112**(18):2805–11.
46. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. *Cell* 2012;**148**(6):1242–57.
47. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* 2012;**13**(8):523–36.
48. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 2011;**12**(9):628–40.
49. Barriot R, Breckpot J, Thienpont B, *et al*. Collaboratively charting the gene-to-phenotype network of human congenital heart defects. *Genome Med* 2010;**2**(3):16.
50. Tewhey R, Bansal V, Torkamani A, *et al*. The importance of phase information for human genomics. *Nat Rev Genet* 2011;**12**(3):215–23.
51. Wang J, Wang W, Li R, *et al*. The diploid genome sequence of an Asian individual. *Nature* 2008;**456**(7218):60–5.
52. Suk E-K, McEwen GK, Duitama J, *et al*. A comprehensively molecular haplotype-resolved genome of a European individual. *Genome Res* 2011;**21**(10):1672–85.
53. Kitzman JO, Mackenzie AP, Adey A, *et al*. Haplotype-resolved genome sequencing of a Gujarati Indian individual. *Nat Biotechnol* 2011;**29**(1):59–63.
54. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;**26**(10):1135–45.
55. Quail MA, Smith M, Coupland P, *et al*. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;**13**:341.
56. Loman NJ, Misra RV, Dallman TJ, *et al*. Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 2012;**30**(5):434–9.
57. Abecasis GR, Auton A, *et al*. 1000 Genomes Project Consortium An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**(7422):56–65.
58. Mills RE, Pittard WS, Mullaney JM, *et al*. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* 2011;**21**(6):830–9.
59. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One* 2012;**7**(2):e30619.
60. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
61. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;**25**(14):1754–60.
62. Weese D, Holtgrewe M, Reinert K. RazerS 3: faster, fully sensitive read mapping. *Bioinformatics* 2012;**28**(20):2592–9.
63. Liu C-M, Wong T, Wu E, *et al*. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* 2012;**28**(6):878–9.
64. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012;**28**(24):3169–77.
65. McKenna A, Hanna M, Banks E, *et al*. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;**20**(9):1297–303.
66. Li H, Handsaker B, Wysoker A, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**(16):2078–9.
67. Koboldt DC, Zhang Q, Larson DE, *et al*. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 2012;**22**(3):568–76.
68. Li R, Li Y, Fang X, *et al*. SNP detection for massively parallel whole-genome resequencing. *Genome Res* 2009;**19**(6):1124–32.
69. Li R, Yu C, Li Y, *et al*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 2009;**25**(15):1966–7.
70. Li S, Li R, Li H, *et al*. SOAPindel: efficient identification of indels from short paired reads. *Genome Res* 2013;**23**(1):195–200.
71. Ye K, Schulz MH, Long Q, *et al*. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 2009;**25**(21):2865–71.

72. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**(6):443–51.
73. Neuman JA, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform* 2013;**14**(1):46–55.
74. Salmela L. Correction of sequencing errors in a mixed set of reads. *Bioinformatics* 2010;**26**(10):1284–90.
75. Teo SM, Pawitan Y, Ku CS, *et al*. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 2012;**28**(21):2711–8.
76. Mills RE, Luttig CT, Larkins CE, *et al*. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* 2006;**16**(9):1182–90.
77. Korbel JO, Abyzov A, Mu XJ, *et al*. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;**10**(2):R23.
78. Chen K, Wallis JW, McLellan MD, *et al*. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 2009;**6**(9):677–81.
79. Hormozdiari F, Hajirasouliha I, Dao P, *et al*. Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 2010;**26**(12): i350–7.
80. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods* 2009;**6**(11 Suppl):S13–20.
81. Alkan C, Kidd JM, Marques-Bonet T, *et al*. Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* 2009;**41**(10):1061–7.
82. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res* 2011;**21**(6):974–84.
83. Harismendy O, Ng PC, Strausberg RL, *et al*. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol* 2009;**10**(3):R32.
84. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**(16):e164.
85. Habegger L, Balasubramanian S, Chen DZ, *et al*. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics* 2012;**28**(17):2267–9.
86. Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res* 2007;**36**(Database):D820–4.
87. Cingolani P, Platts A, Wang LL, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 2012;**6**(2):80–92.
88. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**(4):248–9.
89. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protocol* 2009;**4**(8):1073–81.
90. Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**(8):575–6.
91. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res* 2009; **19**(9):1553–61.
92. Siepel A. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005; **15**(8):1034–50.
93. Davydov EV, Goode DL, Sirota M, *et al*. Identifying a high fraction of the human genome to be under selective constraint using GERP. *PLoS Comput Biol* 2010; **6**(12):e1001025.
94. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 2010;**20**(1):110–21.
95. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput Biol* 2007;**3**(12):e254.
96. Fu W, O'Connor TD, Jun G, *et al*. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 2013;**493**(7431):216–20.
97. Sherry ST, Ward MH, Kholodov M, *et al*. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**(1):308–11.
98. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online Mendelian inheritance in man (OMIM(R)). *Nucleic Acids Res* 2009;**37**(Database):D793–6.
99. Kanehisa M. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**(1):27–30.
100. Kanehisa M. Molecular network analysis of diseases and drugs in KEGG. *Methods Mol Biol* 2013;**939**:263–75.
101. Stenson PD, Mort M, Ball EV, *et al*. The Human Gene Mutation Database: 2008 update. *Genome Med* 2009; **1**(1):13.
102. Baker M. One-stop shop for disease genes. *Nature* 2012; **491**(7423):171.
103. International HapMap Consortium. The International HapMap Project. *Nature* 2003;**426**(6968):789–96.
104. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**(7319):1061–73.
105. Biesecker LG, Mullikin JC, Facio FM, *et al*. The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 2009; **19**(9):1665–74.
106. Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 2004;**20**(7):1022–32.
107. Musumeci L, Arthur JW, Cheung FSG, *et al*. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum Mutat* 2010;**31**(1):67–73.
108. Boomsma DI, Wijmenga C, Slagboom EP, *et al*. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* 2013 doi:10.1038/ejhg.2013.118.
109. Mailman MD, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007; **39**(10):1181–6.
110. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–29.

111. van Driel MA, Cuelenaere K, Kemmeren PPCW, *et al*. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 2005;**33**(Web Server issue):W758–61.
112. Aerts S, Lambrechts D, Maity S, *et al*. Gene prioritization through genomic data fusion. *Nat Biotechnol* 2006; **24**(5):537–44.
113. Franceschini A, Szklarczyk D, Frankild S, *et al*. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 2012; **41**(D1):D808–15.
114. Lee I, Blom UM, Wang PI, *et al*. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 2011;**21**(7):1109–21.
115. Alexeyenko A, Schmitt T, Tjärnberg A, *et al*. Comparative interactomics with Funcoup 2.0. *Nucleic Acids Res* 2011; **40**(D1):D821–8.
116. Matthews L, Gopinath G, Gillespie M, *et al*. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res* 2009;**37**(Database):D619–22.
117. Itan Y, Zhang S-Y, Vogt G, *et al*. The human gene connectome as a map of short cuts for morbid allele discovery. *Proc Natl Acad Sci USA* 2013;**110**(14):5558–63.
118. Zaidi S, Choi M, Wakimoto H, *et al*. De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 2013;**498**:220–23.
119. Siddiqui AS, Khattra J, Delaney AD, *et al*. A mouse atlas of gene expression: large-scale digital gene-expression profiles from precisely defined developing C57BL/6J mouse tissues and cells. *Proc Natl Acad Sci USA* 2005; **102**(51):18485–90.
120. Rustici G, Kolesnikov N, Brandizi M, *et al*. ArrayExpress update–trends in database growth and links to data analysis tools. *Nucleic Acids Res* 2013;**41**(Database issue):D987–90.
121. Bastian F, Parmentier G, Roux J, *et al*. Bgee: integrating and comparing heterogeneous transcriptome data among species. *Data Integr Life Sci* 2008(5109):124–31.
122. Barrett T, Wilhite SE, Ledoux P, *et al*. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2012;**41**(D1):D991–5.
123. Richardson L, Venkataraman S, Stevenson P, *et al*. EMAGE mouse embryo spatial gene expression database: 2010 update. *Nucleic Acids Res* 2010;**38**(Database issue):D703–9.
124. Visel A, Thaller C, Eichele G. GenePaint.org: an atlas of gene expression patterns in the mouse embryo. *Nucleic Acids Res* 2004;**32**(Database issue):D552–6.
125. Diez-Roux G, Banfi S, Sultan M, *et al*. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol* 2011;**9**(1):e1000582.
126. Sikkema-Raddatz B, Johansson LF, de Boer EN, *et al*. Targeted next-generation sequencing can replace sanger sequencing in clinical diagnostics. *Hum Mutat* 2013;**34**: 1035–42.
127. Sivakumaran TA, Husami A, Kissell D, *et al*. Performance evaluation of the next-generation sequencing approach for molecular diagnosis of hereditary hearing loss. *Otolaryngol Head Neck Surg* 2013;**148**:1007–16.
128. Serre D, Gurd S, Ge B, *et al*. Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet* 2008;**4**(2):e1000006.
129. Reddy TE, Gertz J, Pauli F, *et al*. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res* 2012;**22**(5):860–9.
130. Li G, Bahn JH, Lee J-H, *et al*. Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res* 2012;**40**(13):e104.
131. Arrington CB, Bleyl SB, Matsunami N, *et al*. Exome analysis of a family with pleiotropic congenital heart disease. *Circ Cardiovasc Genet* 2012;**5**(2):175–82.
132. Tariq M, Belmont JW, Lalani S, *et al*. SHROOM3 is a novel candidate for heterotaxy identified by whole exome sequencing. *Genome Biol* 2011;**12**(9):R91.
133. Harakalova M, van Harssel JJT, Terhal PA, *et al*. Dominant missense mutations in ABCC9 cause Cantú syndrome. *Nat Genet* 2012;**44**(7):793–6.
134. Pediatric Cardiac Genomics Consortium The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ Res* 2013;**112**(4):698–706.
135. Firth HV, Wright CF.DDD Study. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* 2011;**53**(8):702–3.
136. Sperling SR. Systems biology approaches to heart development and congenital heart disease. *Cardiovascular Res* 2011;**91**(2):269–78.
137. MacLellan WR, Wang Y, Lusis AJ. Systems-based approaches to cardiovascular disease. *Nat Rev Cardiol* 2012;**9**(3):172–84.
138. Kohl P, Crampin EJ, Quinn TA, Noble D. Systems biology: an approach. *Clin Pharmacol Ther* 2010;**88**(1): 25–33.
139. Haraksingh RR, Snyder MP. Impacts of variation in the human genome on gene regulation. *J Mol Biol* 2013;**425**: 3970–77.
140. Jarinova O, Ekker M. Regulatory variations in the era of next-generation sequencing: implications for clinical molecular diagnostics. *Hum Mutat* 2012;**33**(7):1021–30.
141. Smemo S, Campos LC, Moskowitz IP, *et al*. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet* 2012;**21**(14):3255–63.
142. Kim C, Wong J, Wen J, *et al*. Studying arrhythmogenic right ventricular dysplasia with patient-specific iPSCs. *Nature* 2013;**494**(7435):105–10.
143. Moretti A, Bellin M, Welling A, *et al*. Patient-Specific Induced Pluripotent Stem-Cell Models for Long-QT Syndrome. *N Engl J Med* 2010;**363**(15):1397–409.
144. Dambrot C, Passier R, Atsma D, Mummery CL. Cardiomyocyte differentiation of pluripotent stem cells and their use as cardiac disease models. *Biochem J* 2011; **434**(1):25–35.
145. Davis RP, van den Berg CW, Casini S, *et al*. Pluripotent stem cell models of cardiac disease and their implication for drug discovery and development. *Trends Mol Med* 2011; **17**(9):475–84.