HHS Public Access

Author manuscript

Sci China Life Sci. Author manuscript; available in PMC 2019 July 06.

Published in final edited form as:

Sci China Life Sci. 2019 July ; 62(7): 895–904. doi:10.1007/s11427-018-9479-5.

Sequencing XMET genes to promote genotype-guided risk assessment and precision medicine

Yaqiong Jin[1], **Geng Chen**[2], **Wenming Xiao**[3], **Huixiao Hong**[3], **Joshua Xu**[3], **Yongli Guo**[1], **Wenzhong Xiao**[4], **Tieliu Shi**[2], **Leming Shi**[5], **Weida Tong**[3], and **Baitang Ning**[3,*]

[1]Beijing Key Laboratory for Pediatric Diseases of Otolaryngology, Head and Neck Surgery, Beijing Pediatric Research Institute, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing 100045, China;

[2]Center for Bioinformatics and Computational Biology, and the Institute of Biomedical Sciences, Shanghai Key Laboratory of Regulatory Biology, the Institute of Biomedical Sciences and School of Life Sciences, East China Normal University, Shanghai 200241, China;

[3]National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA;

[4]Department of Surgery, Massachusetts General Hospital, Harvard Medical School, Boston, MA 02114, USA;

[5]State Key Laboratory of Genetic Engineering, School of Life Sciences and Cancer Center; Collaborative Innovation Center for Genetics and Development, Fudan University, Shanghai 200433, China

Abstract

High-throughput next generation sequencing (NGS) is a shotgun approach applied in a parallel fashion by which the genome is fragmented and sequenced through small pieces and then analyzed either by aligning to a known reference genome or by *de novo* assembly without reference genome. This technology has led researchers to conduct an explosion of sequencing related projects in multidisciplinary fields of science. However, due to the limitations of sequencing-based chemistry, length of sequencing reads and the complexity of genes, it is difficult to determine the sequences of some portions of the human genome, leaving gaps in genomic data that frustrate further analysis. Particularly, some complex genes are difficult to be accurately sequenced or mapped because they contain high GC-content and/or low complexity regions, and complicated pseudogenes, such as the genes encoding xenobiotic metabolizing enzymes and transporters (XMETs). The genetic variants in XMET genes are critical to predicate interindividual variability in drug efficacy, drug safety and susceptibility to environmental toxicity. We summarized and discussed challenges, wet-lab methods, and bioinformatics algorithms in sequencing "complex" XMET genes, which may provide insightful information in the application of NGS technology for implementation in toxicogenomics and pharmacogenomics.

*Corresponding author: (baitang.ning@fda.hhs.gov).

The author(s) declare that they have no conflict of interest. The information in these materials is not a formal dissemination of the U.S. Food and Drug Administration.

Keywords

next generation sequencing; precision medicine; xenobiotic metabolizing enzymes and transporters; toxicogenomics; pharmacogenomics

High-throughput next generation sequencing (NGS) techniques have incorporated revolutionary innovations to investigate the complexities of genomes, thus becoming the most powerful approach for the generation of genomic data (Cao et al., 2017; Escalona et al., 2016). The advances of NGS have produced a tsunami of information rapidly, and have driven genetic discovery in human disease economically and efficiently. Besides, NGS also promises to help accurate diagnosis by using clinically relevant genetic variants to stratify subgroups of patients for optimal treatments and minimization of adverse drug reactions (Bahcall, 2015; Chen and Shi, 2013; Goodwin et al., 2016). Precision medicine integrates research disciplines and clinical practice to formulate a knowledge-based and personalized treatment plan that can better guide individualized patient care (Bahcall, 2015; Wang and Zhou, 2017). One of the most important steps for the effective practice of precision medicine is to decode the disease-causative, drug-effective and drugtoxic genes and their genetic variants; therefore, a massive genetic screening approach is critical for identifying these genetic variations from individual's genome (Chen and Shi, 2013).

However, due to high percentage of GC (guanosine or cytosine) content, errors produced by sequencing-based chemistry, relatively short length of sequencing reads, and the poor quality of DNA samples, some genomic regions are difficult to sequenced, leaving un-identified gaps in genomic data (Figure 1). On the other hand, a fraction of complex genes is difficult to be sequenced accurately because these genes are either highly diversified or highly homologous (Lauschke and Ingelman-Sundberg, 2018). For example, genes encoding human leukocyte antigens (HLAs), and genes encoding T-cell receptors and B-cell receptors are highly diversified; genes encoding xenobiotic metabolizing enzymes and transporters (XMETs) are highly homologous and are associated with pseudogenes during their evolution. Herein, we refer these genes as "complex genes" since they possess challenges in the sequencing process.

Drug efficacy and safety, and the susceptibility to environmental toxicity are apparently different among individuals, which is largely caused by many genetic variants in genes encoding XMETs and drug targeted proteins (Evans, 1999; Evans and Relling, 2004; Ning et al., 2014). However, the full spectrum landscape of genetic variants in XMET genes and drug targeted genes has not been drawn. Hurdles, including highly homologous genes among the super-families or sub-families, functional variants distributed across entire genes, many pseudogenes with sequences highly similar to their corresponding functional XMET genes, make XMET genes difficult to be accurately mapped or genotyped (Lauschke and Ingelman-Sundberg, 2018). Similarly, HLA genes may also mediate adverse drug reactions (Daly et al., 2009; Guo et al., 2013; Liu et al., 2018). Genotyping of HLA genes with a high resolution by using NGS is an excessive challenge because that the large number of polymorphisms, the extensive allelic diversity of the gene loci, and the complexity of the dimerized molecules are the genetic characteristics among HLA genes (Erlich, 2012). In

addition, high GC content DNA fragments affect the DNA synthesizing efficiency and accuracy, resulting in miscalled variants from sequencing artifacts.

To improve the accuracy for sequencing these “complex” genes, challenges and strategies, including library preparation methods, gene capture approaches, and fragment mapping algorithms are summarized and discussed in this minireview.

Genes, xenobiotics, and beyond

XMETs in drug metabolisms

XMETs, often described as DMETs (drug metabolizing enzymes and transporters), include drug metabolizing enzymes (phase I/II) and transporters (phase III), playing central roles in the metabolism, elimination and detoxification of drugs (Xu et al., 2005). Comprised mostly of the Cytochrome P450 (CYP) enzymes, the phase I XMETs are of a high importance for drug metabolism (Lynch and Price, 2007). The Phase II drug conjugating or metabolizing enzymes, are usually categorized into super-families of enzymes, such as UDPglucuronosyltransferases (UGTs), sulfotransferases (SULTs), NAD(P)H:menadione reductase (NMO) or NAD(P)H:quinone oxidoreductase (NQO), N-acetyltransferases (NATs), glutathione S-transferases (GSTs) and epoxide hydrolases (EPHs) (Jancova et al., 2010; Rushmore and Kong, 2002). Phase III proteins are transporters playing crucial roles in drug absorption, distribution, and excretion (Xu et al., 2005). Genetic variants in many XMET genes have been associated with responses to specific drugs, and susceptibilities to environmental toxicity and diseases and pharmacogenetic findings have benefited patients (Lee et al., 2016). The Clinical Pharmacogenetics Implementation Consortium (CPIC) has been making efforts to identify drugvariant interaction with high importance for clinicians; based on various patient genotype, CPIC suggests guidelines for usage of drug-dosing (Lee et al., 2016). In addition, the Pharmacogenomics Knowledgebase (PharmGKB) has been widely used in functional variants annotation (Ng et al., 2017). Similarly, XMETs account for metabolizing environmental toxicants, which is highly related to health risks.

XMETs are critical factors in response to environmental toxicants

Xenobiotic metabolizing enzymes are double edged swords with hazardous or beneficial effects on human health, through their critical roles in metabolizing many drugs and xenobiotics. Each XMET responds differently to exogenous chemicals due to its substrate specificity, which has significant influence on drug metabolism, procarcinogen activation, and toxicant detoxification. The hazardous effects of XMETs result from enhanced transport or production of toxic or carcinogenic agents (Sheweita, 2000). For example, CYP1A1 is a prominent enzyme responsible for the metabolic activation of polycyclic aromatic hydrocarbons (PAHs). XMETs convert PAHs into reactive intermediates that covalently bind to genomic DNA and produce DNA adducts—an essential event for chemical carcinogenesis. Similarly, CYP1A2 can activate a battery of procarcinogens, such as aromatic amines and amides, and heterocyclic amines, which are risk factors for bladder cancer and colon cancer (Koda et al., 2017; Sheweita, 2000). Some environmental carcinogens, such as aflatoxin B1, a highly mutagenic and carcinogenic agent that may cause hepatocellular carcinoma, can be activated by several cytochrome P450 isozymes such

as CYP1A2, CYP3A4 and CYP2E1 (Manson et al., 1997). The carcinogenic potency of procarcinogens is highly correlated with the activity and substrate specificity of the specific cytochrome P450 isozyme. Inter-individual variability in XMET expression is largely influenced by genetic variations (Koturbash et al., 2015; Yang et al., 2013). In contrast, the beneficial effects of XMETs in the detoxification of toxicants are attributed to the activity the phase II XMETs, such as glutathione S-transferase (GSTs), sulfotransferases (SULTs) and UDP-glucuronyl transferase (UGTs). These phase II XMETs can inactivate chemical toxicants by conjugation of a parent compound with a charged species resulting in less toxic or inactive metabolites that are exported more efficiently (Sheweita, 2000). For example, SULT1A1 catalyzes the detoxification process of a group of structurally diverse compounds, such as small phenols, iodothyronines, environmental estrogen-like compounds and heterocyclic and aromatic amines (Ning et al., 2005).

Genetic variants in XMET genes modulate susceptibilities to toxicity and diseases

The biological activities of XMETs are essential to maintain a proper balance between detoxification and activation reactions involving xenobiotics and drugs. This balance is largely determined by many variables including genetic background, sex, age, dietary and environmental factors. Genetic factors which modify the expression and activity of XMETs exert particularly important influences on an individual's susceptibility to toxicity and disease. Therefore, XMET genetic polymorphisms may serve as molecular biomarkers providing predictive information relevant to drug efficacy, drug safety, and susceptibility to toxicity and disease, which constitute the fundamental components of precision medicine (Evans and Relling, 2004).

Well-done cooked red meat is a rich source of heterocyclic amines that can be activated metabolically to bind to DNA and create stable DNA adducts. The consumption of welldone red meat is considered an environmental risk factor for colorectal cancer for this reason (Aune et al., 2013). Phase I XMETs CYP1A2, CYP1A1, CYP1B1, and CYP2A6 catalyze *N*-oxidation of heterocyclic amines to produce hydroxylamine derivatives, and phase II XMETs, such as acetyltransferase (NAT1, NAT2) or SULT1A1, catalyze the formation of electrophilic *N*-acetyloxy or *N*-sulfonyloxy esters that react also with DNA (Nowell et al., 2002; Turesky, 2007). In contrast, phase II XMETs involved in the conjugation of these intermediate metabolites, such as GSTA1 and UGT1A1, play protective roles against carcinogenesis. Therefore, the genetic variants that provide decreased catalytic activity for these enzymes responsible for the metabolic activation of heterocyclic amines are associated with increased risks for colorectal cancer (Hein et al., 2000; Lang et al., 1994; Nowell et al., 2002). For example, genetic variants providing reduced activity for detoxification enzymes, such as *GSTA1*B* (Coles et al., 2001) and *UGT1A1*-3279 GG/TG, are associated with an increased risk for colorectal cancer (Girard et al., 2008). Exposure to benzene increases the risk for acute myeloid leukemia, acute lymphocytic leukemia, chronic lymphocytic leukemia, and other blood-related cancers. Upon exposure to humans, benzene is mainly oxidized by CYP2E1 to produce the toxic quinones that are further detoxified by GSTs. In a recent study, null alleles of *GSTT1* and/or *GSTM1* are identified as risk factors that increased hosts' susceptibility to benzeneinduced hepatotoxicity (Nourozi et al., 2017).

On the other hand, some specific mutations caused by mutagens/carcinogens can serve as diagnostic signatures. It has been demonstrated that some specific mutagens cause characteristic patterns of mutations in the DNA of tumor tissues, and these patterns are defined as mutation signatures. For example, exposure to heterocyclic amines can introduce G:C single base pair deletions, especially in the 5′-GGGA-3′ fragment, along with G:CàT:A transversions. These mutations can be identified within colorectal tumor samples, by comparison with normal adjacent tissues (Lynch et al., 1998), which are mutational signatures specifically associated with the exposure of hosts to heterocyclic amines.

Integration of environment-gene interaction into precision medicine

Genetic variations (genetics) and environmental modulators (epigenetics) both exert major influences on gene expression and enzyme activity phenotypes. A human may inherit a predisposition (genetics) to determine a phenotype (such as a susceptibility to toxicity, and efficacy/safety to a drug), but the magnitude of the phenotype is always a “product” of environment-gene interactions. Despite tremendous challenges, it is expected that the implementation of precision medicine, by integrating genotype, phenotype and the impact of gene-environment interactions, will become pivotal for improving effective human healthcare (Collins and Varmus, 2015; Hamburg and Collins, 2010). In practice, toxicogenomics and pharmacogenomics are the most important components of precision medicine. Applying genotype-guided technologies, toxicogenomics assesses environmental risk factors and provides personalized prevention of exposures to specific hazards, while pharmacogenomics evaluates drug efficacy/safety and to provide personalized treatments (Figure 2). Unfortunately, currently defined genetic variants do not represent the full spectrum of genetic components responsible for drug safety and disease susceptibility. Accurate genotyping of genetic variants in XMET genes is crucial for the implementation of toxicogenomics and pharmacogenomics.

Application and strategies in sequencing genes encoding XMETs

Microarray based XMETs genotyping

Microarray platforms have a proven track record spanning almost two decades with easier usage and low costs. With microarray-based methods, some well-defined genetic variants in XMET genes are genotyped; however, these tests identify only a limited number of genetic variants crossing XMET genes, between 22 (AmpliChip CYP450 test, Roche) and 1936 (DMET™ Plus Premier Pack, Affymetrix), and missing over 90% of the CYPs variants (Brown et al., 2014; Lauschke and Ingelman-Sundberg, 2016; Londin et al., 2014). Although whole-genome single nucleotide polymorphism (SNP) arrays seemed to cover more variants than 1,936 SNPs in DMET Plus (Affimetrix) or 2,088 SNPs in OmniQuad (illumina) for XMETs, they often lack key markers for most XMET loci since SNP coverage for XMET genes is often poor (Brown et al., 2014; Peiffer and Gunderson, 2009). Phillips et al. developed an optimized and validated XMET genotyping panel based on the Illumina GoldenGate platform, encompassing approximately 3,000 variants, which contains an extensive list of genes with a broader applicability for genotyping XMET-related variants (Brown et al., 2014).

Advantages and problems in genotyping XMETs using exome sequencing technologies

NGS techniques are gaining popularity, owing to their advantages in delivering faster, less expensive, and more detailed genomic information. The recent advances in NGS technologies, specifically in whole exome sequencing (WES), make it a practical method for detecting novel variants with lower costs and decreased time required (Lauschke and Ingelman-Sundberg, 2016; Londin et al., 2014; Sboner et al., 2011; Zou et al., 2017); therefore, genetic data acquisition no longer constitutes a major bottleneck for genome-wide association studies. Despite its benefits, some limitations are accompanied with the NGS technology. Notably, the power/ability of WES is highly dependent on the hybridization efficiency of PCR amplification primers or oligonucleotide probes to capture a targeted region. Therefore, some pharmacogenomically relevant variants may not be included in the specifically-designed target regions, resulting in missed coverage for novel variants (Londin et al., 2014). Actually, approximately 30% of variants in XMET genes are located in intronic and intergenic regions that are generally not covered by the exome sequencing technologies that are currently available. Additionally, poor coverage for some genes and regions is observed. For instance, genetic variants in the COMT gene and the VKORC1 promoter region had an average depth of coverage less than 20, compared to variants in other XMET genes with a coverage more than 30, which was reported in the same experiment using the same group of samples (Londin et al., 2014).

Fifty-seven genes located in different regions of the human genome encode the cytochrome P450 enzymes and numerous related pseudogenes. According to the sequence similarity of CYPs, they are grouped into 18 families and 44 subfamilies (Zanger et al., 2008). It is challenging that many sequences within a large fraction of CYP genes are classified as inaccessible by short-read NGS methodologies, and the inaccessible fraction of some important but highly complex genes, such as CYP2D6, with or without gene duplications (Gaedigk, 2013; Lauschke and Ingelman-Sundberg, 2016; Meijerman et al., 2007). Another limitation of WES is that the up-to-date technologies are unable to routinely and precisely characterize copy number variants (CNVs) (Londin et al., 2014). Besides CNVs, exome sequencing also does not provide any additional information into gene rearrangements unless such rearranged fragments are specifically generated for sequencing. The challenges discussed above may create substantial possibilities for false-negative pharmacogenomics findings. Thus, targeted sequencing approaches using specifically amplified PCR fragments or using enriched libraries (these are designed to capture specific fragments/genes) with further advanced NGS sequencing instruments are being pursued, which should eventually provide more sophisticated approaches with improved performance for sequencing difficult genes. Another possible approach to address analytical concordance for SNP identification involves combining data from both microarrays and NGS. Anticipated advances in NGS technologies may make higher read depth whole genome sequencing (WGS) more cost-effective, which could overcome some limitations for targeted sequencing or exome sequencing (Londin et al., 2014).

Advantages and problems in genotyping XMETs using whole genome sequencing technologies

High-depth of coverage in WGS is strongly recommended for DNA resequencing, by which the advantage is to interrogate all types of genetic variants including SNPs, indels, structural variants and CNVs in both coding regions and noncoding regions (Sims et al., 2014). Thus, compared with WES or XMET array, WGS provides increased power to identify more variants in XMET genes (Yang et al., 2016). WGS is able to sequence genes that have a large number of important rare variants, like DPYD and G6PD. However, due to its high cost, the median depth of WGS with the same cost of WEG is generally lower than that for WES (Yang et al., 2016). The lower coverage depth for WGS, compared to WES, is associated with reduced overall sequence accuracy (Sims et al., 2014), leading to a higher false-negative rate in variant calling. However, a greater depth of coverage does not necessarily solve all sequencing problems. Especially, it cannot resolve problems with assembling the gaps associated with repetitive sequences. Instead, the paired-end read with a known distance approach is utilized to place clearly repetitive regions that are smaller than the distance (Schatz et al., 2010; Sims et al., 2014).

Though strategies using WGS are good research approaches to identify genetic variants in XMET genes, WGS is not an economical technology for all clinical applications. The cost of genotyping using NGS is less than that of arraybased technology (Ng et al., 2017). Customized target sequencing has become a cost-effective, highly-efficient and high-throughput methodology with reasonably higher depth (Gordon et al., 2016). Approaches, like PGRNseq, which employs customized capture probes targeting genes of interest, such as XMETs, provide economical tools to improve the pharmacogenomics studies in clinics (Yang et al., 2016). Accordingly, discovery of genetic variants for XMET genes by WGS and assessment of the variant distributions in patients/controls with customized NGS should be a more suitable strategy for pharmacogenomics studies.

Strategies for sequencing XMETs

Improving base accuracy and depth of coverage—The advances in NGS have greatly improved the cost of sequencing, throughput, and speed; however, individual NGS reads generally exhibit limited accuracy and shorter lengths compared to those from traditional Sanger sequencing. Different NGS sequencing platforms utilize alternative sequencing methodologies and data analysis, resulting in distinct tendencies for different types of sequencing errors (Fox et al., 2014). High sequencing accuracy is required for correctly distinguishing important variations in SNPs and indels and for detecting transcriptional modifications due to RNA-editing or alternative mRNA splicing. Two general strategies could be explored to improve NGS sequencing accuracy: (i) increase sequencing depth to reduce the error rate in determining bases; and (ii) improve the accuracy of base-calling algorithms (Ledergerber and Dessimoz, 2011).

Improving pseudogene removal and specifically mapping homologous reads —Mapping and aligning sequence reads to the appropriate reference genome is a basic and important step in analyzing sequencing data. However, many reference genomes for mammals contain large portions of homologous genomic regions, leading to a possibility

that homologous reads could be mapped to multiple different genomic places (multimapping reads). Pseudogenes are important sources of homologous sequences since they are highly similar to their parent genes in the genome. Pseudogenes may possess important regulatory functions, and the human genome harbors many pseudogenes (Pei et al., 2012). To accurately conduct variant (SNPs, indels, inversions, and CNVs) calling, splicing detection and gene/transcript expression quantification (Chen et al., 2017), correction of mapping bias in homologous regions is vital. If we only consider the uniquely mapped reads, we may miss some important information, such as the true read depth of homologous regions, which may lead to inaccurate CNV calling or expression estimation.

Currently, there are mainly three different strategies to resolve the problems introduced by multi-mapping reads: (i) simply ignore the multi-mapping reads, but this may distort gene expression data associated with the ignored reads; (ii) allow the reads to be mapped to all the possible genomic regions; however, it may increase the complexity of shortread mapping problems; and (iii) handle the multi-mapping reads with specific tools, such as mmquant (Zytnicki, 2017) and Rcount (Schmid and Grossniklaus, 2015). Consequently, the correction of mapping bias (e.g. multi-mapping reads) using the third strategy aforementioned would be necessary to precisely analyze the sequencing data (Chen et al., 2016; Roberts et al., 2011).

Improving CNV algorithms—Copy number variations (CNVs) are an important type of genomic variation evolved from duplications, insertions or deletions of genomic sequences whose lengths may vary greatly. CNVs can cause individual differences in physiological phenotypes (Iafrate et al., 2004; Sebat et al., 2004) and may play important roles in the pathogenesis of diverse diseases/cancers (Hastings et al., 2009; Shlien and Malkin, 2009, 2010). Furthermore, CNVs associated with CYP2D6 and glutathione *S*-transferase genes are well known to affect drug safety and efficacy (He et al., 2011).

WGS and WES technologies provide unprecedented opportunities for identifying CNVs with higher coverage and resolution. At present, many tools have been developed to detect and characterize CNVs; these tools can be used for CNV genotyping based on the features and information obtained from WGS or WES data. Different CNV calling tools may differ in terms of accuracy, types of CNVs detected, genotyping speed and computational memory cost, because distinct software packages use disparate algorithms to identify CNVs. Moreover, the length of reads and the type of reads (e.g. single-end and paired-end) can also influence their performance in CNV detection. Currently, the strategies for identifying CNVs can be grouped into five categories based on (i) read depth, e.g. CNVrd2 (Nguyen et al., 2014);(ii) paired-end mapping, e.g. commonLAW (Hormozdiari et al., 2011); (iii) split read, e.g. Gustaf (Trappe et al., 2014);(iv) *de novo* assembly, e.g. TIGRA (Chen et al., 2014); and(v) combinatorial strategy based on more than one of the aforementioned approaches, e.g. Hydra-Multi (Table 1) (Lindberg et al., 2015). Zhao et al. (2013) provided a detailed summary on the available tools in each category for CNV detection. However, the read length for most NGS technologies is still short (<500 bp) and current CNV calling algorithms also present certain shortcomings. For example, each tool may only focus on certain types of CNVs, and none of them could systematically identify all kinds of CNVs. These limitations hinder the comprehensive detection of different CNVs especially CNVs

with very long length and/or the CNVs occurring in repetitive genomic regions. Therefore, continuous improvement of both sequencing technologies and CNV calling algorithms will help to facilitate the accuracy and performance of CNV identification.

Applying third generation sequencing technology in sequencing XMET genes—Third-generation sequencing is providing new insights in genomic research and clinical applications. The comparison of advantages and disadvantages among the first, next-generation and third-generation sequencing is shown in Table 2. In third-generation sequencing technologies, PCR is usually not needed before sequencing, thus shortening sequencing time and increasing length of sequencing reads (Liu et al., 2012). For example, the signal is captured in a real time manner during the extension step of adding nucleotides by enzymes, such as the single-molecule real-time (SMRT) strategy developed by Pacific Bioscience (Liu et al., 2012). Notably, the average read length of Pacific Biosciences PacBio RS is 2,500–3,000 bp, approximately 10 times longer than that of NGS technology. Besides, Oxford Nanopore Technologies (ONT) also developed a strategy to directly sequence DNA molecule by measuring electric currency variations as the bases in a single-stranded are threaded through the nanopore (Weirather et al., 2017). ONT sequencing produces the similar data features with PacBio, thus having the similar advantages and disadvantages as the PacBio system (Weirather et al., 2017). Therefore, the thirdgeneration sequencing technology can solve the problems generated by short-reads NGS in sequencing CYPs by increasing the accuracy of homologous sequencing mapping, providing enriched information for pseudogene removal, and accurately counting copy numbers and genomic locations of CNVs. Although the throughput and base-calling accuracy of the PacBio RS are lower than those of second-generation sequencing technologies (Liu et al., 2012), it has its strength in genome biology studies, especially for sequencing some difficult genome regions, such as XMETs. Sanger sequencing can generate longer and more accurate reads compared to NGS technologies; however, it is costly and time-consuming. To date, despite the relatively high error rate in the third-generation sequencing, it has its advantage to overcome the challenges in resolving ambiguity of highly homologous regions in XMET genes with much longer reads produced by third-generation sequencing technologies. The HLA genes are highly polymorphic, and some of them (A, B, C, DRB1 and DQB1) are longer than 5 kb. Using the NGS technologies is not able to resolve haplotypes of those long and highly polymorphic HLA genes. Ambardar et al. have set up a fulllength HLA typing method based on third-generation (PacBio SMRT) sequencing technology (Ambardar and Gowda, 2018). A similar study demonstrated that sequencing HLA genes by Pac-Bio technology could provide high-resolution allelic information for multiple HLA genes with phased SNPs (Mayor et al., 2015).

To annotate all genetic variants with recommendations from the Clinical Pharmacogenetics Implementation Consortium (CPIC), the PharmCAT (A Pharmacogenomics Annotation Tool) project is under development. Hopefully PharmCAT can provide sophisticated approaches to interpret XMET variant alleles and haplotypes in the near future (Klein and Ritchie, 2017).

Summary and perspectives

The NGS technologies continue to evolve and are accompanied by the innovations in both experimental designs and related bioinformatics algorithms. With the advancement of sequencing technology, new and effective bioinformatics tools will face more challenges. They need to deal with larger amounts of data, and to analyze data more accurately with a higher efficiency. It is worth noting that the human reference genome and related gene annotations are essential to the application of personalized medicine. However, these databases still need to be improved (Chen et al., 2013b). With the improvement in sequencing technologies and computational algorithms, the reference genome will be eventually completed (Chen et al., 2013a). Moreover, we expected that the advance of NGS technologies, with more depths of sequencing coverage, longer reads of the sequencing reaction, higher accuracy of the base calling, better assembly algorithms for pseudogene removal, more precise haplotype construction algorithms, correct annotation of functional relevance to identified variants (Lauschke and Ingelman-Sundberg, 2016) and simplified workflow would facilitate the accuracy for sequencing “complex genes”. NGS technology and related bioinformatics tools provide us an opportunity to explore massive scientific problems related to human diseases and drug sensitivity, and to reveal possible mechanisms of genetic events. However, we still need to carefully carry out sequencing experiments and improve skills and knowledge in interpretation of NGS data. Although the road is full of thorns, we believe that NGS technology innovation will help the practice of precision medicine and promote public health.

Acknowledgements

We thank Drs. William Slikker Jr. and Gray W. Miller for their scientific advice for the preparation of this manuscript. This work was supported by the FDA Project (E0765001) and the National Key Research and Development Program of China (2016YFC0902100 to Geng Chen).

References

Abel HJ, Duncavage EJ, Becker N, Armstrong JR, Magrini VJ, and Pfeifer JD (2010). SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. Bioinformatics 26, 2684–2688. [PubMed: 20876606]

Abyzov A, and Gerstein M (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. Bioinformatics 27, 595–603. [PubMed: 21233167]

Abyzov A, Urban AE, Snyder M, and Gerstein M (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res 21, 974–984. [PubMed: 21324876]

Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, Kitzman JO, Baker C, Malig M, Mutlu O, et al. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41, 1061–1067. [PubMed: 19718026]

Ambardar S, and Gowda M (2018). High-resolution full-length HLA typing method using third generation (Pac-Bio SMRT) sequencing technology. Methods Mol Biol 1802, 135–153. [PubMed: 29858806]

Aune D, Chan DSM, Vieira AR, Navarro Rosenblatt DA, Vieira R, Greenwood DC, Kampman E, and Norat T (2013). Red and processed meat intake and risk of colorectal adenomas: a systematic review and meta-analysis of epidemiological studies. Cancer Causes Control 24, 611–627. [PubMed: 23380943]

Bahcall O (2015). Precision medicine. Nature 526, 335. [PubMed: 26469043]

Brown AMK, Renaud Y, Ross C, Hansen M, Mongrain I, Valois D, Carleton BC, Hayden MR, Dubé MP, Tardif JC, et al. (2014). Development of a broad-based ADME panel for use in pharmacogenomic studies. Pharmacogenomics 15, 1185–1195. [PubMed: 25141894]

Cao J, Yu Y, Huang J, Liu R, Chen Y, Li S, and Liu J (2017). Genome re-sequencing analysis uncovers pathogenecity-related genes undergoing positive selection in Magnaporthe oryzae. Sci China Life Sci 60, 880–890. [PubMed: 28755293]

Chen G, and Shi TL (2013). Next-generation sequencing technologies for personalized medicine: promising but challenging. Sci China Life Sci 56, 101–103. [PubMed: 23393024]

Chen G, Shi T, and Shi L (2017). Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60, 116–125. [PubMed: 27294835]

Chen G, Wang C, Shi L, Qu X, Chen J, Yang J, Shi C, Chen L, Zhou P, Ning B, et al. (2013a). Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA 19, 479–489. [PubMed: 23431329]

Chen G, Wang C, Shi L, Tong W, Qu X, Chen J, Yang J, Shi C, Chen L, Zhou P, et al. (2013b). Comprehensively identifying and characterizing the missing gene sequences in human reference genome with integrated analytic approaches. Hum Genet 132, 899–911. [PubMed: 23572138]

Chen G, Yang J, Chen J, Song Y, Cao R, Shi T, and Shi L (2016). Identifying and annotating human bifunctional RNAs reveals their versatile functions. Sci China Life Sci 59, 981–992. [PubMed: 27650948]

Chen K, Chen L, Fan X, Wallis J, Ding L, and Weinstock G (2014). TIGRA: a targeted iterative graph routing assembler for breakpoint assembly. Genome Res 24, 310–317. [PubMed: 24307552]

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6, 677–681. [PubMed: 19668202]

Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, and Lander ES (2009). Highresolution mapping of copy-number alterations with massively parallel sequencing. Nat Meth 6, 99–103.

Coles B, Nowell SA, MacLeod SL, Sweeney C, Lang NP, and Kadlubar FF (2001). The role of human glutathione S-transferases (hGSTs) in the detoxification of the food-derived carcinogen metabolite N-acetoxy-PhIP, and the effect of a polymorphism in hGSTA1 on colorectal cancer risk. Mutat Res/Fund Mol Mech Mutag 482, 3–10.

Collins FS, and Varmus H (2015). A new initiative on precision medicine. N Engl J Med 372, 793–795. [PubMed: 25635347]

Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, et al. (2009). HLAB*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet 41, 816–819. [PubMed: 19483685]

Erlich H (2012). HLA DNA typing: past, present, and future. Tissue Antig 80, 1–11.

Escalona M, Rocha S, and Posada D (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. Nat Rev Genet 17, 459–469. [PubMed: 27320129]

Evans WE (1999). Pharmacogenomics: translating functional genomics into rational therapeutics. Science 286, 487–491. [PubMed: 10521338]

Evans WE, and Relling MV (2004). Moving towards individualized medicine with pharmacogenomics. Nature 429, 464–468. [PubMed: 15164072]

Fox EJ, Reid-Bayliss KS, Emond MJ, and Loeb LA (2014). Accuracy of next generation sequencing platforms. Next Gener Seq Appl 1, pii: 1000106. [PubMed: 25699289]

Gaedigk A (2013). Complexities of CYP2D6 gene analysis and interpretation. Int Rev Psychiatry 25, 534–553. [PubMed: 24151800]

Girard H, Butler LM, Villeneuve L, Millikan RC, Sinha R, Sandler RS, and Guillemette C (2008). UGT1A1 and UGT1A9 functional variants, meat intake, and colon cancer, among Caucasians and AfricanAmericans. Mutat Res/Fund Mol Mech Mutag 644, 56–63.

Goodwin S, McPherson JD, and McCombie WR (2016). Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351. [PubMed: 27184599]

Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, and Scherer S (2016). PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. Pharmacogenet Genomics 26, 161–168. [PubMed: 26736087]

Guo YL, Shi LM, Hong HX, Su ZQ, Fuscoe J, and Ning BT (2013). Studies on abacavir-induced hypersensitivity reaction: a successful example of translation of pharmacogenetics to personalized medicine. Sci China Life Sci 56, 119–124. [PubMed: 23393027]

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, and Sahinalp SC (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. Bioinformatics 26, 1277–1283. [PubMed: 20385726]

Hamburg MA, and Collins FS (2010). The path to personalized medicine. N Engl J Med 363, 301–304. [PubMed: 20551152]

Handsaker RE, Korn JM, Nemesh J, and McCarroll SA (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nat Genet 43, 269–276. [PubMed: 21317889]

Hastings PJ, Lupski JR, Rosenberg SM, and Ira G (2009). Mechanisms of change in gene copy number. Nat Rev Genet 10, 551–564. [PubMed: 19597530]

He Y, Hoskins JM, and McLeod HL (2011). Copy number variants in pharmacogenetic genes. Trends Mol Med 17, 244–251. [PubMed: 21388883]

Hein DW, Doll MA, Fretland AJ, Leff MA, Webb SJ, Xiao GH, Devanaboyina US, Nangju NA, and Feng Y (2000). Molecular genetics and epidemiology of the NAT1 and NAT2 acetylation polymorphisms. Cancer Epidemiol Biomarkers Prev 9, 29–42. [PubMed: 10667461]

Hormozdiari F, Hajirasouliha I, Dao P, Hach F, Yorukoglu D, Alkan C, Eichler EE, and Sahinalp SC (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics 26, i350–i357. [PubMed: 20529927]

Hormozdiari F, Hajirasouliha I, McPherson A, Eichler EE, and Sahinalp SC (2011). Simultaneous structural variation discovery among multiple paired-end sequenced genomes. Genome Res 21, 2203–2212. [PubMed: 22048523]

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C (2004). Detection of large-scale variation in the human genome. Nat Genet 36, 949–951. [PubMed: 15286789]

Iqbal Z, Caccamo M, Turner I, Flicek P, and McVean G (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44, 226–232. [PubMed: 22231483]

Ivakhno S, Royce T, Cox AJ, Evers DJ, Cheetham RK, and Tavaré S (2010). CNAseg—a novel framework for identification of copy number changes in cancer from second-generation sequencing data. Bioinformatics 26, 3051–3058. [PubMed: 20966003]

Jancova P, Anzenbacher P, and Anzenbacherova E (2010). Phase II drug metabolizing enzymes. Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub 154, 103–116. [PubMed: 20668491]

Klein TE, and Ritchie MD (2017). PharmCAT: A pharmacogenomics clinical annotation tool. Clin Pharmacol Ther 104, 19–22. [PubMed: 29194583]

Koda M, Iwasaki M, Yamano Y, Lu X, and Katoh T (2017). Association between NAT2, CYP1A1, and CYP1A2 genotypes, heterocyclic aromatic amines, and prostate cancer risk: a case control study in Japan. Environ Health Prev Med 22, 72. [PubMed: 29165164]

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, and Gerstein MB (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10, R23. [PubMed: 19236709]

Koturbash I, Tolleson WH, Guo L, Yu D, Chen S, Hong H, Mattes W, and Ning B (2015). microRNAs as pharmacogenomic biomarkers for drug efficacy and drug safety assessment. Biomarkers Med 9, 1153–1176.

Lang NP, Butler MA, Massengill J, Lawson M, Stotts RC, HauerJensen M, and Kadlubar FF (1994). Rapid metabolic phenotypes for acetyltransferase and cytochrome P4501A2 and putative exposure to food-borne heterocyclic amines increase the risk for colorectal cancer or polyps. Cancer Epidemiol Biomarkers Prev 3, 675–682. [PubMed: 7881341]

Lauschke VM, and Ingelman-Sundberg M (2016). Precision medicine and rare genetic variants. Trends Pharmacol Sci 37, 85–86. [PubMed: 26705087]

Lauschke VM, and Ingelman-Sundberg M (2018). How to consider rare genetic variants in personalized drug therapy. Clin Pharmacol Ther 103, 745–748. [PubMed: 29313952]

Ledergerber C, and Dessimoz C (2011). Base-calling for next-generation sequencing platforms. Briefings Bioinf 12, 489–497.

Lee EMJ, Xu K, Mosbrook E, Links A, Guzman J, Adams DR, Flynn E, Valkanas E, Toro C, Tifft CJ, et al. (2016). Pharmacogenomic incidental findings in 308 families: The NIH Undiagnosed Diseases Program experience. Genet Med 18, 1303–1307. [PubMed: 27253732]

Lindberg MR, Hall IM, and Quinlan AR (2015). Population-based structural variation discovery with Hydra-Multi. Bioinformatics 31, 1286–1289. [PubMed: 25527832]

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, and Law M (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012, 251364. [PubMed: 22829749]

Liu Y, Yu Y, Nie X, Zhao L, and Wang X (2018). Association between HLA-B*15:02 and oxcarbazepine-induced cutaneous adverse reaction: a meta-analysis. Pharmacogenomics 19, 547–552. [PubMed: 29629814]

Londin ER, Clark P, Sponziello M, Kricka LJ, Fortina P, and Park JY (2014). Performance of exome sequencing for pharmacogenomics. Personalized Med 12, 109–115.

Lynch AM, Gooderham NJ, Davies DS, and Boobis AR (1998). Genetic analysis of PHIP intestinal mutations in Muta™ Mouse. Mutagenesis 13, 601–605. [PubMed: 9862191]

Lynch T, and Price A (2007). The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. Am Fam Physician 76, 391–396. [PubMed: 17708140]

Manson M, Ball HW, Barrett MC, Clark HL, Judah DJ, Williamson G, and Neal GE (1997). Mechanism of action of dietary chemoprotective agents in rat liver: induction of phase I and II drug metabolizing enzymes and aflatoxin B1 metabolism. Carcinogenesis 18, 1729–1738. [PubMed: 9328168]

Mayor NP, Robinson J, McWhinnie AJM, Ranade S, Eng K, Midwinter W, Bultitude WP, Chin CS, Bowman B, Marks P, et al. (2015). HLA typing for the next generation. PLoS ONE 10, e0127153. [PubMed: 26018555]

Medvedev P, Fiume M, Dzamba M, Smith T, and Brudno M (2010). Detecting copy number variation with mated short reads. Genome Res 20, 1613–1622. [PubMed: 20805290]

Meijerman I, Sanderson LM, Smits PHM, Beijnen JH, and Schellens JHM (2007). Pharmacogenetic screening of the gene deletion and duplications of CYP2D6. Drug Metab Rev 39, 45–60. [PubMed: 17364880]

Miller CA, Hampton O, Coarfa C, and Milosavljevic A (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. PLoS ONE 6, e16327. [PubMed: 21305028]

Ng D, Hong CS, Singh LN, Johnston JJ, Mullikin JC, and Biesecker LG (2017). Assessing the capability of massively parallel sequencing for opportunistic pharmacogenetic screening. Genet Med 19, 357–361. [PubMed: 27537706]

Nguyen HT, Merriman TR, and Black MA (2014). The CNVrd2 package: measurement of copy number at complex loci using highthroughput sequencing data. Front Genet 5, 248. [PubMed: 25136349]

Nijkamp JF, van den Broek MA, Geertman JMA, Reinders MJT, Daran JMG, and de Ridder D (2012). De novo detection of copy number variation by co-assembly. Bioinformatics 28, 3195–3202. [PubMed: 23047563]

Ning B, Nowell S, Sweeney C, Ambrosone CB, Williams S, Miao X, Liang G, Lin D, Stone A, Luke Ratnasinghe D, et al. (2005). Common genetic polymorphisms in the 5′-flanking region of the SULT1A1 gene: haplotypes and their association with platelet enzymatic activity. Pharmacogenet Genomics 15, 465–473. [PubMed: 15970794]

Ning B, Su Z, Mei N, Hong H, Deng H, Shi L, Fuscoe JC, and Tolleson WH (2014). Toxicogenomics and cancer susceptibility: advances with next-generation sequencing. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 32, 121–158. [PubMed: 24875441]

Nourozi MA, Neghab M, Bazzaz JT, Nejat S, Mansoori Y, and Shahtaheri SJ (2017). Association between polymorphism of GSTP1, GSTT1, GSTM1 and CYP2E1 genes and susceptibility to benzeneinduced hematotoxicity. Arch Toxicol 92, 1983–1990. [PubMed: 29204680]

Nowell S, Coles B, Sinha R, MacLeod S, Luke Ratnasinghe D, Stotts C, Kadlubar FF, Ambrosone CB, and Lang NP (2002). Analysis of total meat intake and exposure to individual heterocyclic amines in a case-control study of colorectal cancer: contribution of metabolic variation to risk. Mutat Res/Fund Mol Mech Mutag 506–507, 175–185.

Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, et al. (2012). The GENCODE pseudogene resource. Genome Biol 13, R51. [PubMed: 22951037]

Peiffer DA, and Gunderson KL (2009). Design of tag SNP whole genome genotyping arrays. Methods Mol Biol 529, 51–61. [PubMed: 19381970]

Qi J, and Zhao F (2011). inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. Nucleic Acids Res 39, W567–W575. [PubMed: 21715388]

Roberts A, Trapnell C, Donaghey J, Rinn JL, and Pachter L (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12, R22. [PubMed: 21410973]

Rushmore TH, and Kong AN (2002). Pharmacogenomics, regulation and signaling pathways of phase I and II drug metabolizing enzymes. Curr Drug Metab 3, 481–490. [PubMed: 12369894]

Sboner A, Mu XJ, Greenbaum D, Auerbach RK, and Gerstein MB (2011). The real cost of sequencing: higher than you think! Genome Biol 12, 125. [PubMed: 21867570]

Schatz MC, Delcher AL, and Salzberg SL (2010). Assembly of large genomes using second-generation sequencing. Genome Res 20, 1165–1173. [PubMed: 20508146]

Schmid MW, and Grossniklaus U (2015). Rcount: simple and flexibleRNA-Seq read counting. Bioinformatics 31, 436–437. [PubMed: 25322836]

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. (2004). Large-scale copy number polymorphism in the human genome. Science 305, 525–528. [PubMed: 15273396]

Sheweita S (2000). Drug-metabolizing enzymes mechanisms and functions. Curr Drug Metab 1, 107–132. [PubMed: 11465078]

Shlien A, and Malkin D (2009). Copy number variations and cancer.Genome Med 1, 62. [PubMed: 19566914]

Shlien A, and Malkin D (2010). Copy number variations and cancer susceptibility. Curr Opin Oncol 22, 55–63. [PubMed: 19952747]

Sims D, Sudbery I, Ilott NE, Heger A, and Ponting CP (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15, 121–132. [PubMed: 24434847]

Sindi S, Helman E, Bashir A, and Raphael BJ (2009). A geometric approach for classification and comparison of structural variants. Bioinformatics 25, i222–i230. [PubMed: 19477992]

Sindi SS, Onal S, Peng LC, Wu HT, and Raphael BJ (2012). An integrative probabilistic model for identification of structural variation in sequencing data. Genome Biol 13, R22. [PubMed: 22452995]

Trappe K, Emde AK, Ehrlich HC, and Reinert K (2014). Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. Bioinformatics 30, 3484–3490. [PubMed: 25028727]

Turesky RJ (2007). Formation and biochemistry of carcinogenic heterocyclic aromatic amines in cooked meats. Toxicol Lett 168, 219–227. [PubMed: 17174486]

Wang X, and Zhou XJ (2017). Magnetic resonance imaging in personalized medicine. Sci China Life Sci 60, 1–4. [PubMed: 28058635]

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Buck D, and Au KF (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. F1000 Res 6, 100.

Xie C, and Tammi MT (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinf 10,80.

Xu C, Li CYT, and Kong ANT (2005). Induction of phase I, II and III drug metabolism/transport by xenobiotics. Arch Pharm Res 28, 249–268. [PubMed: 15832810]

Yang L, Price ET, Chang CW, Li Y, Huang Y, Guo LW, Guo Y, Kaput J, Shi L, and Ning B (2013). Gene expression variability in human hepatic drug metabolizing enzymes and transporters. PLoS ONE 8, e60368. [PubMed: 23637747]

Yang W, Wu G, Broeckel U, Smith CA, Turner V, Haidar CE, Wang S, Carter R, Karol SE, Neale G, et al. (2016). Comparison of genome sequencing and clinical genotyping for pharmacogenes. Clin Pharmacol Ther 100, 380–388. [PubMed: 27311679]

Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871. [PubMed: 19561018]

Yoon S, Xuan Z, Makarov V, Ye K, and Sebat J (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19, 1586–1592. [PubMed: 19657104]

Zanger UM, Turpeinen M, Klein K, and Schwab M (2008). Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. Anal Bioanal Chem 392, 1093–1108. [PubMed: 18695978]

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, and Barillot E (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. Bioinformatics 26, 1895–1896. [PubMed: 20639544]

Zhao M, Wang Q, Wang Q, Jia P, and Zhao Z (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 (Suppl 11), S1.

Zou X, Tang G, Zhao X, Huang Y, Chen T, Lei M, Chen W, Yang L, Zhu W, Zhuang L, et al. (2017). Simultaneous virus identification and characterization of severe unexplained pneumonia cases using a metagenomics sequencing technique. Sci China Life Sci 60, 279–286. [PubMed: 27921234]

Zytnicki M (2017). mmquant: how to count multi-mapping reads? BMC Bioinformatics 18, 411. [PubMed: 28915787]

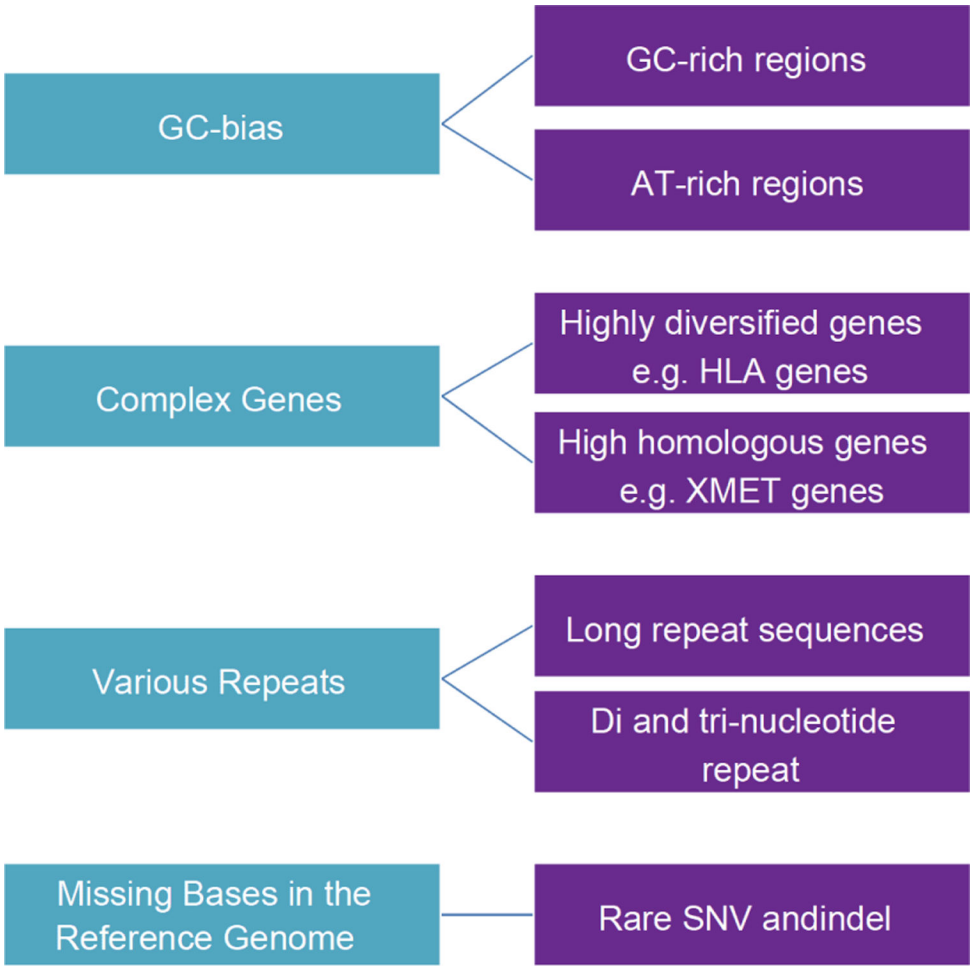

Figure 1.
(Color online) Difficult regions/genes for NGS.

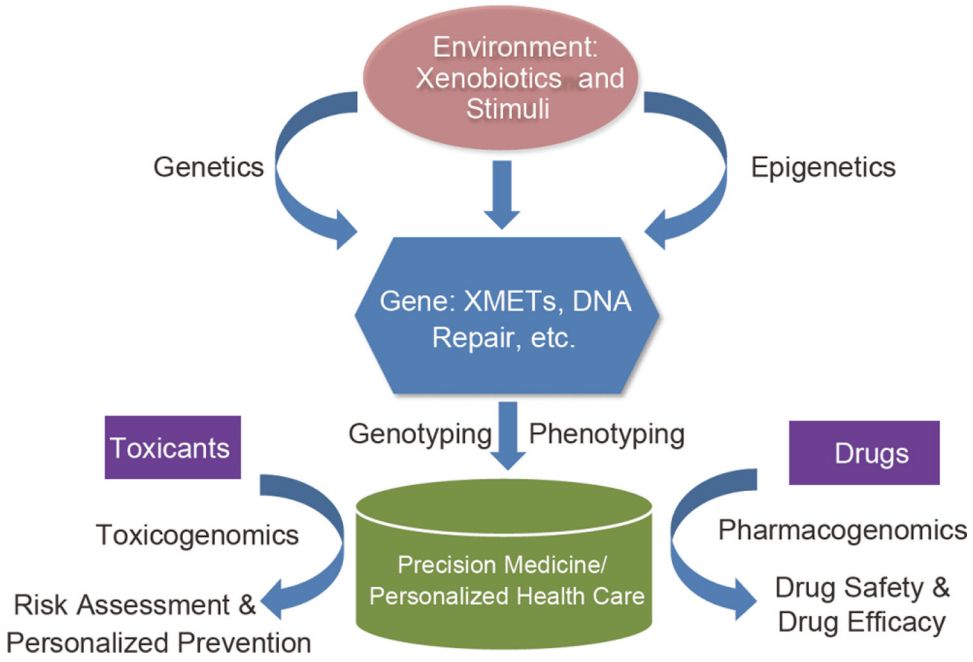

Figure 2.
(Color online) Integration of environment-gene interaction into precision medicine.

Table 1

Categories of CNV tools

Category of strategies	Tools for CNV calling
Read depth based	ReadDepth (Miller et al., 2011), CNVnator (Abyzov et al., 2011), CNV-seq (Xie and Tammi, 2009), CNAseg (Ivakhno et al., 2010), RDXplorer (Yoon et al., 2009), SegSeq (Chiang et al., 2009), mrCaNaVar (Alkan et al., 2009), CNVrd2 (Nguyen et al., 2014)
Paired-end mapping based	BreakDancer (Chen et al., 2009), commonLAW (Hormozdiari et al., 2011), VariationHunter (Hormozdiari et al., 2010), PEMer (Korbel et al., 2009), GASV (Sindi et al., 2009)
Split read based	Pindel (Ye et al., 2009), AGE (Abyzov and Gerstein, 2011), SLOPE (Abel et al., 2010), Gustaf (Trappe et al., 2014)
De novo assembly based	Cortex assembler (Iqbal et al., 2012), Magnolya (Nijkamp et al., 2012), TIGRA (Chen et al., 2014)
Combinatorial strategy based	SVDetect (Zeitouni et al., 2010), CNVer (Medvedev et al., 2010), Genome STRiP (Handsaker et al., 2011), GASVPro (Sindi et al., 2012), nGAP-sv (Qi and Zhao, 2011), NovelSeq (Hajirasouliha et al., 2010), Hydra-Multi (Lindberg et al. 2015)

Table 2

Advantages and disadvantages for sequencing technologies

	Sanger sequencing	Next-generation sequencing	Third-generation sequencing
Methods	Fluorescence	Fluorescence/optics	Fluorescence/optics/electric current
Length	1,000–1,500 bp	150–300 bp	>10,000 bp
Data volume	150 kb	18 Tb	7 Gb–4 Tb
PCR dependence	Yes	Yes	No
Error rate	~0.3%	~0.1%–1%	~15%
Error type	NA	Mismatch	Indel
Speed	Low	High	High
Cost per base	High	Low	Low