frontiers in
GENETICS

REVIEW ARTICLE
published: 26 August 2013
doi: 10.3389/fgene.2013.00160

The role and challenges of exome sequencing in studies of human diseases

Zuoheng Wang[1]*, Xiangtao Liu[2,3,4], Bao-Zhu Yang[2] and Joel Gelernter[2,5,6]

[1] *Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, USA*
[2] *Division of Human Genetics, Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA*
[3] *Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA*
[4] *Department of Statistics, The Ohio State University, Columbus, OH, USA*
[5] *Department of Genetics, Yale University School of Medicine, New Haven, CT, USA*
[6] *Department of Neurobiology, Yale University School of Medicine, New Haven, CT, USA*

Edited by:
Rui Feng, University of Pennsylvania, USA

Reviewed by:
Zheyang Wu, Worcester Polytechnic Institute, USA
Guimin Gao, Virginia Commonwealth University, USA

****Correspondence:***
Zuoheng Wang, Department of Biostatistics, Yale School of Public Health, Yale University, 60 College Street, New Haven, CT 06520, USA
e-mail: zuoheng.wang@yale.edu

Recent advances in next-generation sequencing technologies have transformed the genetics study of human diseases; this is an era of unprecedented productivity. Exome sequencing, the targeted sequencing of the protein-coding portion of the human genome, has been shown to be a powerful and cost-effective method for detection of disease variants underlying Mendelian disorders. Increasing effort has been made in the interest of the identification of rare variants associated with complex traits in sequencing studies. Here we provided an overview of the application fields for exome sequencing in human diseases. We describe a general framework of computation and bioinformatics for handling sequencing data. We then demonstrate data quality and agreement between exome sequencing and exome microarray (chip) genotypes using data collected on the same set of subjects in a genetic study of panic disorder. Our results show that, in sequencing data, the data quality was generally higher for variants within the exonic target regions, compared to that outside the target regions, due to the target enrichment. We also compared genotype concordance for variant calls obtained by exome sequencing vs. exome genotyping microarrays. The overall consistency rate was >99.83% and the heterozygous consistency rate was >97.55%. The two platforms share a large amount of agreement over low frequency variants in the exonic regions, while exome sequencing provides much more information on variants not included on exome genotyping microarrays. The results demonstrate that exome sequencing data are of high quality and can be used to investigate the role of rare coding variants in human diseases.

Keywords: exome sequencing, exome arrays, Mendelian diseases, complex traits, whole-genome sequencing

INTRODUCTION

Determining the genetic basis of human diseases is one of the major research areas in medical science (McCarthy et al., 2008). The allelic spectrum of variants underlying human disorders has long been a topic of discussion and speculation (Pritchard, 2001; Reich and Lander, 2001). Despite significant progress in the identification of large numbers of loci that contribute to complex traits in genome-wide association studies (GWAS), only a small fraction of the observed heritability is explained by the confirmed (genomewide-significant) common variants (Manolio et al., 2009; Schork et al., 2009). A recent study (Yang et al., 2010) demonstrated that the heritability estimation can be improved by using all genomewide common single nucleotide polymorphisms (SNPs) relative to that using only identified genomewide-significant SNPs, and this accounts for some of the heritability that is "missing." The advent of massively parallel sequencing technologies has transformed the field of human genetics and substantially reduced the cost of sequencing large genomic regions relative to the traditional Sanger sequencing (Mardis, 2008; Ansorge, 2009; Metzker, 2010). This allows researchers to investigate variants from a wide range of allelic spectrum, including variants that are too rare for inclusion on microarrays and new mutations; and higher-level structural variants. Thus, sequencing approaches have the potential to explain some of the missing heritability from GWAS for complex traits, through identification of rare variants and structural variations (Manolio et al., 2009; Eichler et al., 2010). However, it is still financially impractical, for most laboratories, to perform whole-genome sequencing for large numbers of subjects at sufficiently high coverage, in order to complete valid large-scale genetic association studies of complex traits.

A more economical approach to gene discovery is to focus on functional coding regions of the human genome. The exome represents about 1% of the human genome with approximately 30 million base pairs, but accounts for about 85% of mutations identified in Mendelian diseases (Ng et al., 2009). Recent developments in high-throughput sequence capture methods have made exome sequencing an attractive and practical approach for investigation of coding variation (Biesecker, 2010; Kaiser, 2010; Mamanova et al., 2010; Ng et al., 2010a,b). During the past 3 years, more than 100 genes have been characterized in rare Mendelian

diseases by the use of whole exome sequencing. Application of this approach for non-Mendelian phenotypes has been, to date, much less widespread.

Traditional microarray-based tag SNP genotyping techniques designed for GWAS target relatively common variants. With the rich information gathered from sequencing over 12,000 individual exomes and whole-genome sequences representing multiple ethnicities and complex traits, the companies that market genotyping arrays (chips), Illumina, and Affymetrix, through a collaboration with leading geneticists, have designed exome chips that contain putative functional exonic variants, with the majority of them focusing on rare markers selected from sequencing studies (Exome chip design[1]). The introduction of exome arrays has provided a fast and economical platform for genotyping the included exonic variants, and has to some extent bridged the gap between traditional genotyping arrays and exome sequencing of very large numbers of samples, although they bring with them their own particular technical issues, most particularly, the inability to query very rare variants or new mutations.

Both exome sequencing and exome genotyping arrays are designed to investigate coding variation. The current approach for exome sequencing is based on a probe hybridization method to select the entire set of human exons as the sequencing target (Hodges et al., 2007; Gnirke et al., 2009). Although the exonic regions are the primary target, the efficiency of different capture technologies can affect the amount of information outside target regions. Currently, there is still a portion of captured DNA fragments falling into non-coding regions such as introns, intron-exon boundary regions, and intergenic regions – some of these regions often contain functional elements. A recent report (Guo et al., 2012) demonstrated that the small amount of sequencing data that lies outside the exonic target regions is of high quality and can be used in genetic studies. In contrast, exome arrays focus on a fixed set of variants by design. Therefore, exome sequencing, compared to the use of exome arrays, generates not only more genetic variations at base-pair resolution in the coding regions, but also additional, albeit limited, variant information outside the primary target regions. In this paper, we first provide an overview of the main application fields for exome sequencing relative to exome genotyping arrays in human diseases. Next, we describe the computational and statistical challenges for handling sequencing data. Then we evaluate the data quality and agreement between these two platforms using our exome sequencing and exome microarray data collected on the same set of subjects. Finally, we discuss some limitations of exome sequencing.

APPLICATIONS OF EXOME SEQUENCING

Next-generation sequencing (NGS) technologies have been applied to several important areas including genomes, transcriptomes, epigenomes, and metagenomes (Zhou et al., 2010). Here, we mainly consider applications of sequencing to the identification of genes and mutations that influence risk for human diseases.

MENDELIAN DISORDERS

The "traditional" approach to elucidating causes of Mendelian disorders – or in any event, the first generalizable approach to locate risk genes without prior knowledge – is based on linkage analysis followed by positional cloning (Botstein and Risch, 2003). Linkage studies require ascertainment of a sufficient number of probands with their families, and thus are not suitable for rare Mendelian diseases where only one or a few individuals may be sampled. In addition, modest-sized linkage studies are not sensitive enough to detect co-segregation within families in case of locus heterogeneity and phenotypic heterogeneity. NGS methods, on the other hand, have the potential to identify all kinds of genetic variation at base-pair resolution throughout the human genome in a single experiment (Bamshad et al., 2011; Gilissen et al., 2011; Ku et al., 2011), and provide an unbiased approach to detecting genetic variation within an individual. Currently sequencing instruments are still limited by throughput and cost efficiency. Exome sequencing, by capturing the protein-coding portion of the genome, generates a full picture of variation at functionally important regions of the genome (excluding regulatory changes), and has now become technically feasible and a more cost-effective strategy to work out the genetic basis of Mendelian disease. It has been a proven tool for the identification of *de novo* mutations underlying some rare monogenic diseases such as Kabuki syndrome (Ng et al., 2010a) and Miller syndrome (Ng et al., 2010b). Since November 2009, exome sequencing has led to the discovery of more than 100 genes in Mendelian diseases (Rabbani et al., 2012). As the sequencing cost per base will drop in the near future, we expect that whole-genome sequencing will be the ultimate approach to detection of *all* genomic variations and help us gain more knowledge on the genetics of Mendelian diseases – but even when the laboratory costs of generating full sequences decrease, there will still be very substantial informatics costs, which are also much lower of exome analysis.

COMPLEX DISEASES

Over the past 8 years, the genetics research community has put a great deal of effort on studies of complex diseases which are caused by the interplay among multiple behavioral, environmental, and genetic factors. Association studies have been applied for decades to investigate the genetics of complex traits (Marian, 2012). With the advancement of high-throughput genotyping technologies, GWAS has been the main tool to find susceptibility genes based on the principle of linkage disequilibrium at the population level (Visscher et al., 2012). The development of SNP arrays genotyping hundreds of thousands or even millions of markers in a single assay has made GWAS feasible in large-scale population genetic studies. Since 2005, more than 8,000 loci have been reported to be associated with various human complex diseases and traits (A catalog of published GWAS[2]). The selection of markers investigated in most GWAS is based on the "common disease, common variant" hypothesis. SNP arrays provide a picture of genome-wide polymorphism in many individuals (The International HapMap Consortium, 2005, 2007), however, they inevitably suffer from ascertainment biases favoring SNPs that are common

[1] http://genome.sph.umich.edu/wiki/Exome_Chip_Design

[2] http://www.genome.gov/gwastudies

in the populations for variant discovery (Akey et al., 2003; Clark et al., 2005). In contrast, gene sequencing provides a more accurate and complete perspective with respect to all polymorphisms in target regions, or whole-genome (Tennessen et al., 2011). As a result, the field is now shifting toward the study of low frequency variants under the hypothesis of "common disease, rare variant," i.e., multiple rare variants with large effect size are in some cases the main determinants of complex disease genetic risk (Marian, 2012). Exome genotyping arrays, based on the knowledge attained from many NGS studies, were designed also to target at a carefully selected subset of rare coding variants. Currently, exome arrays have served as a fast and economical tool for the initial investigation of the role of rare exonic variants in complex diseases (Huyghe et al., 2013), although more comprehensive evaluation of low frequency variants, copy number variants (CNVs), and structural variation, is accomplished much more effectively by NGS.

COMPUTATIONAL AND STATISTICAL CHALLENGE OF SEQUENCING DATA

Next-generation sequencing instruments sequence millions of short DNA fragments in parallel. Compared to gene chip analysis, the data generated by sequencing require more sophisticated bioinformatics and statistical tools. In the identification of variants in NGS studies, the raw data are pre-processed into nucleotide base calls called short reads, varying from dozens to hundreds of base pairs, in the form of a FASTQ file. To call variants from sequencing data, many alignment methods and variant callers have been developed and used to create complex pipelines. A typical pipeline contains an aligner and a variant caller. The aligner maps each of the short reads to positions on a reference genome. The resulting sequence alignment is stored in a sequence alignment/map (SAM) or binary alignment/map (BAM) file (Li et al., 2009a). The variant caller identifies variant sites where the aligned sequences deviate from the known sequences at the reference position. The list of positions is recorded in a variant call format (VCF) file (Danecek et al., 2011). Further steps involve filtering and annotation to reduce variant sites to a smaller set of genes (when the sequence studied is exomic) with possible function and activity. We will now discuss these steps in detail and review the statistical strategies for identifying causal variants in human diseases.

ALIGNMENT

"Alignment" is the step of matching short nucleotide reads to a reference genome. There are various software programs, either commercially available or freely distributed, that can be used to perform sequence reads alignment; to name a few, Bowtie/Bowtie2 (Langmead et al., 2009; Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009, 2010), MAQ (Li et al., 2008), Novoalign[3], and SOAP (Li et al., 2009c). There are many others that are more computationally intensive and are less frequently used. The performance of different alignment methods has been extensively studied (Bao et al., 2011; Ruffalo et al., 2011; Pattnaik et al., 2012). They are based on either hash tables or the Burrows–Wheeler transform (BWT; Burrows and Wheeler, 1994). The former hashes short reads or the reference genome into memory, while the latter compresses data features by creating an index of the reference genome to allow fast access of potential alignment locations (Nielsen et al., 2011). In general, BWT-based methods are faster and more memory-efficient. For instance, the BWA approach, based on BWT, provides a good balance between speed, memory usage, and accuracy, and is currently one of the most commonly used methods for alignment in sequencing projects.

As the current NGS technologies use PCR-like amplification steps in the library preparation, multiple reads originating from the same template could be sequenced. Overrepresentation of certain alleles due to amplification bias introduced during library construction tends to interfere with variant calling. For this reason, it is common to remove PCR duplicates after alignment in exome or whole-genome sequencing studies.

VARIANT CALLING

After alignment of short reads to the reference genome, the next step in the bioinformatics process is variant identification. Currently the sequencing error rate is estimated to be about 1%, which is at a similar scale of the frequency of rare variants or higher. For genotype calling, the presence of sequencing error poses a computational challenge for the identification of true variants. Early generations of genotype calling methods counted allele at each position and used simple cutoff values to determine when to call a SNP. More recent probabilistic methods, such as MAQ (Li et al., 2008) and SOAPsnp (Li et al., 2009b), use fixed prior values for modeling heterozygote probability as well as sequencing error, and make genotype calls based on posterior genotype probabilities. Currently, some widely used variant calling methods include SAMtools (Li et al., 2009a), the Genome Analysis ToolKit (GATK, McKenna et al., 2010), and Atlas2 (Challis et al., 2012). SAMtools builds upon a revised MAQ model to perform computation of genotype likelihood and SNP calling. GATK utilizes the MapReduce (Dean and Ghemawat, 2008) functional programming technique for variant calling, SNP filtering, and quality recalibration. Atlas2 employs a logistic regression model trained on validated whole-exome sequencing data and has better power to assess the quality of potential variants (Ji, 2012).

We conducted a comprehensive evaluation of the variant identification methods using the exome sequencing data described in the next section. Based on our comparisons, GATK in general provided the highest quality of variant identification (Liu et al., Unpublished data).

Insertion and deletion (Indel) mutations are another common form of polymorphism. It requires gapped alignment and pair-end sequence inference. Several software packages have been developed to identify indels, including Pindel, a pattern growth method; and Dindel, a Bayesian approach. A detailed review on Indel calling has been published by Neuman et al. (2012).

There are several issues that can complicate the variant calling step. First, the presence of indels is a major source of false positive in variant identification. Alignment algorithms that allow for gapped alignments are preferred. Second, variable GC content in short reads, error introduced by library preparation due to PCR artifacts, and variable base quality scores can affect variant calling. The original quality scores assigned by the sequencer machine have been shown to be inaccurate and biased. Thus several SNP

[3]http://novocraft.com

calling algorithms, like GATK and SOAPsnp, have recommended recalibration of base quality scores, using various calibrated error models to empirically estimate error rates for each base, in order to improve variant call accuracy.

ANALYZING VARIANTS IN SEQUENCING

The main challenge of analyzing sequencing variants in human diseases is to identify disease-related alleles (which may be new mutations) accounting for a large number of non-pathogenic polymorphisms in the genome (Bamshad et al., 2011). Strategies for finding causal variants differ between Mendelian and complex diseases. Currently, successes in serious Mendelian disorders through exome sequencing rely on various heuristic filtering methods to reduce the number of candidate genes. First, the complete penetrance of a trait is usually assumed, i.e., all carriers of a disease-causing variant will have the phenotype. Any variants present in public databases such as HapMap (The International HapMap Consortium, 2005, 2007), 1000 Genomes Project (Abecasis et al., 2010), and dbSNP (Sherry et al., 2001) will be excluded from further consideration. Then on the basis of the mode of inheritance, for example, a recessive model, the list of candidate variants can be further reduced. This has successfully led to the identification of rare causal variants in more than 10 studies of recessive disorders. However, this type of filtering has certain limitations. Restricting the candidate variants to those not in public databases in the first filtering step could result in exclusion of possible pathogenic variants in the database, an especially noteworthy problem for the mapping of recessive traits. In addition, filtering based on complete penetrance can eliminate variants that are segregating in the population at low frequencies. Therefore more sophisticated analytical and filtering procedures that take into account the minor allele frequency (MAF) of the risk variant hold great promise to finding causal genes in Mendelian disorders (Stitziel et al., 2011).

To identify likely causal variants in complex traits, association tests are commonly employed. Sequencing studies enable us to investigate rare variants association with a trait under the assumption that multiple rare variants constitute the driving force for the trait of interest. The association with rare variants poses new statistical challenges. Power to detect an association with an individual rare variant can be very low because only a small percentage of study subjects carry a rare variant. To increase statistical power, many groups have investigated aggregating sets of rare variants within a gene or genomic region to enrich association signals (Li and Leal, 2008; Madsen and Browning, 2009; Han and Pan, 2010; Morris and Zeggini, 2010; Price et al., 2010; Ionita-Laza et al., 2011; Lin and Tang, 2011; Wu et al., 2011), and recent studies show that power to detect rare variant effects can be greatly enhanced. A comprehensive review on the statistical methodology of sequence-based association studies is described by Ionita-Laza et al. (2013). Another important aspect in sequencing-based association studies is the choice of an appropriate study design. Population-based and family-based designs are the two most commonly used approaches in genetic association studies. For rare variants with large effect size, family-based designs can be advantageous because a particular rare variant found in an affected individual, if it is not a new mutation, is more common in that individual's family than in subjects randomly sampled in the population; this design can therefore potentially enrich for genetic effects. Trio designs, and some other family designs, are also robust to population structure (Ott et al., 2011). However, it can be more difficult to ascertain samples for family-based designs compared to population-based designs. For different study designs, the analytical strategy for rare variant association needs to be chosen accordingly.

Above, we describe a general framework of computation and bioinformatics for handling sequencing data. Next we demonstrate data quality and agreement between exome sequencing and exome microarray (chip) genotypes using our data collected on the same set of subjects in a genetic study of panic disorder.

DATA DESCRIPTION

We studied whole exome sequencing data on 20 patients of panic disorder collected at Connecticut VA Medical Center (VAMC). Twelve of these were from a single pedigree of five generations with more than 70 family members (not all of whom could be genotyped), and the rest were unrelated. All patients gave informed consent approved by the institutional review boards at Yale and CT VAMC. We studied all samples by exome capture using the NimbleGen SeqCap EZ exome v2.0 kit, which targets 44.1 Mb of the genome by design; samples were sequenced at the Yale Center for Genome Analysis (YCGA). DNA fragments from the 20 samples were barcoded and sequenced on five lanes of a flowcell (four samples per lane). The exome sequence data were 74-base paired-end reads generated from the Illumina HiSeq system.

Reads were aligned to the UCSC reference human genome assembly hg19 using the sequence alignment software BWA version 0.6.1 with the default parameters. The mapping files in SAM format were converted to the BAM format and sorted by SAMtools version 0.1.18. Local realignment around the known indels was performed by GATK version 1.6.9 on the sorted BAM files. Picard tools version 1.5.3 was used to remove PCR duplicates. Finally, base quality score recalibration was performed using GATK. These steps generated BAM files ready for variant calling. We used GATK for variant identification. Then the raw variants were filtered using VCFtools version 0.1.7. We further applied genotype filtering using depth $\geq$5 and genotype quality score $\geq$20 (Guo et al., 2012).

The 20 samples were also interrogated for 247,134 variants using the Illumina HumanExome Beadchip genotyping microarray. More than 90% of variants on the exome array fall in the human RefSeq exons. The majority of them are non-synonymous single nucleotide variations. The Illumina exome chip also contains a small fraction of SNPs in splice sites, selected synonymous SNPs, tag SNPs for previous GWAS hits in a variety of diseases, and ancestry informative markers (AIMs). Eight samples failed the genotyping quality control step were excluded from further analysis.

RESULTS

On the 20 samples, we obtained an average of 48.7 (range 31.0–77.6) million reads per subject, with 93$\times$ mean depth in the target regions. The total length of the target region was 47.1 Mb, of which 34.1 Mb were exomic. On average, 95.9% (94.3–97.2%) of reads were mapped to the human reference genome. After

removal of PCR duplicates, 90.7% (87.4–93.0%) of reads were retained. Among those uniquely mapped, 58.8% (55.7–62.9%) of reads were within the exonic regions. This proportion is similar to the numbers reported for Agilent's SureSelect v1 and Illumina's TrueSeq capture kits (Guo et al., 2012). The coverage for the target regions was as follows: 40.6 Mb (57.9%) had coverage of at least 1×, 33.9 Mb (48.4%) had coverage of at least 10×, and 32.4 Mb (46.3%) had coverage of at least 15×. For sequences outside the target region, 209.8 Mb were covered by at least 1 read, 40.5 Mb were covered by at least 10 reads, and 33.9 Mb were covered by at least 15 reads. The comparison of the average read depth inside and outside of the targeted exome is displayed in **Figure 1**. As we expected, the depth of coverage in the exome regions was higher in most regions due to target enrichment. An interesting feature regarding read depth is that it varied across subjects in the target regions, but stayed similar outside the target regions.

After applying GATK and variant filtering, we identified an average of 26,082 (24,122–28,058) variants per subject inside the target regions, with Ti/Tv ratio of 2.85 (2.80–2.95). In addition, we observed an average of 63,760 (51,414–83,835) variants per subject outside the target regions, with Ti/Tv ratio of 2.17 (2.14–2.20). These results are close to the reports that the expected Ti/Tv ratio is around 3.0 for variants inside exons and about 2.0 elsewhere (Bainbridge et al., 2011). The median quality score of variants inside the target regions is 875.4, more than twice of the median quality score of 340.0 outside the target regions. Based on the distribution of variant quality scores inside and outside the target regions (**Figure 2**), the variants identified within the exome regions are of higher quality relative to those outside the target regions.

Besides variant quality score, another way to measure data quality for sequence-based variant calling is to investigate genotype concordance using an alternative genotyping platform. We use the exome microarray data for this purpose. Among the 12 subjects passed quality control on exome arrays, we identified 32,616 (13.2%) variant sites that showed at least one variation, i.e., at least one subject had a heterozygous genotype (denoted by 0/1) or homozygous rare allele genotype (denoted by 1/1). We compared concordance between the array genotypes and the sequence-based genotype calls. We calculated the genotype consistency rate between exome sequence-based and exome chip-based SNP calls for variants overlapping the two platforms in our samples. We used two types of consistency rate: overall consistency and heterozygous variant consistency. Heterozygous consistency rate was defined as the ratio between the number of heterozygous genotypes consistent between exome chip and exome sequencing and the number of heterozygous genotypes on the exome chip that had sequence-based calls with genotype quality score ≥20 and depth ≥5. The results for the 12 subjects are shown in **Table 1**. The overall consistency rate with array-based variant calls

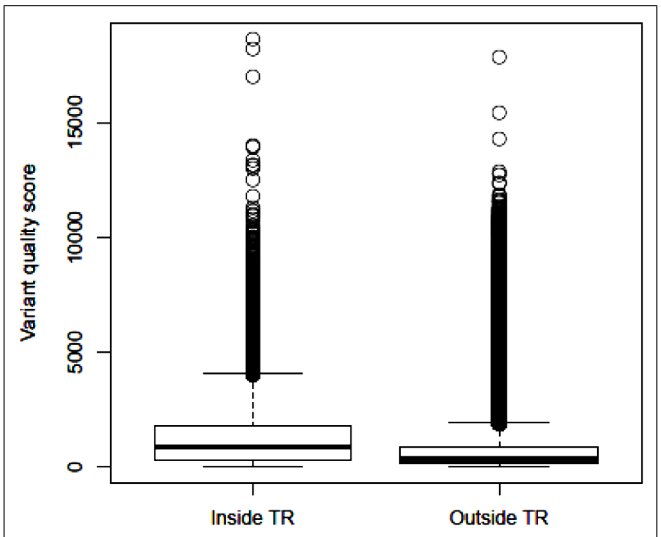

FIGURE 2 | Boxplot of identified variant quality scores inside the target regions and outside the target regions.

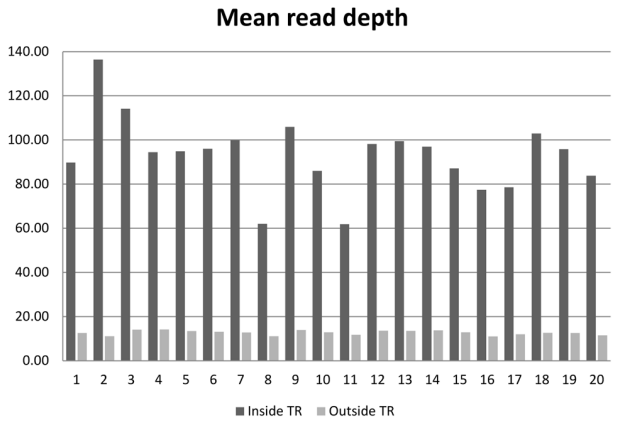

FIGURE 1 | Mean read depth inside the target regions and outside the target regions on the 20 sequenced subjects.

Table 1 | Results of genotype consistency between exome sequencing and exome chip on 12 subjects.

	Consistent genotypes			Consistency rate	
Subject	0/0	0/1	1/1	Overall (%)	Heterozygous (%)
1	218523	4686	2949	99.84	98.28
2	220203	4860	3092	99.83	98.56
4	218043	4608	3016	99.87	97.88
5	218478	4654	2976	99.84	98.21
6	218463	4685	2870	99.83	98.07
7	218553	4765	2999	99.85	97.96
8	214082	4625	2815	99.86	97.55
9	219050	4888	2892	99.87	98.33
15	218233	4719	2995	99.85	98.25
16	217804	4579	3024	99.83	97.99
17	217440	4654	2931	99.84	98.39
19	219067	4566	3025	99.83	98.26

was >99.83% in all samples, and the heterozygous consistency rate was 98.14% (97.55–98.56%). The actual overall consistency rate is higher because we observed a large portion of concordant genotype calls between these two platforms falling in the category of homozygous reference genotypes. On average, more genotype calling errors would occur when the underlying genotype contains the allele that is not the reference allele. Depending on the purpose of the study, for example, in gene-trait association studies, the goal is usually to search for putative rare variants that could be causal for the trait; then, the heterozygous SNP calls would be more informative and the consistency measure based on heterozygous SNPs would be more representative of the true error rate. We also found that the consistency rate in the 1/1 genotype category was similar to the heterozygous consistency rate in our dataset.

Overall, the genotype calls generated by exome sequencing and exome genotyping arrays showed high agreement in all the 12 samples.

DISCUSSION

We have provided an overview of the application of exome-focused NGS technologies in human diseases. The growing number of exome sequencing studies demonstrates the power of this approach in mapping genes involved in Mendelian disorders and suggests utility for complex traits as well. In many successful studies, a small number of individuals was analyzed, and often only affected individuals have been sequenced. However, there are still a large number of Mendelian diseases with unknown genetic causes.

Although exome sequencing has generated high-quality data for single nucleotide variant detection with sufficient depth of coverage, it is still difficult to detect accurately indels with short sequence reads generated by NGS technologies. In addition, exome sequencing is not suitable for the identification of structural variants and chromosomal rearrangements that may involve non-exonic sequence. Furthermore, as the current sequence capturing methods suffer from the problem of uneven and incomplete exonic region capture (Parla et al., 2011), potentially interesting mutations in these exonic regions could be missed. This will likely be solved in the future when the cost of whole-genome sequencing is lower.

Studies of genetically complex traits have also benefited from exome sequencing since the advent of NGS technologies. Although the small sample sizes that can be used in Mendelian diseases are underpowered for detecting association using currently available association tests for complex traits, we can still gain insight by studying small cohorts from the extreme ends of the phenotypic spectrum of common traits, and as costs come down, well powered studies of complex traits via exome sequencing have become feasible. This has been demonstrated by a successful example of a whole exome sequencing study of patients with extremely low levels of low-density lipoprotein (LDL) cholesterol (Musunuru et al., 2010). The findings of risk alleles in GWAS typically cannot pinpoint causal variants, but exome sequencing studies enable more accurate and complete variant discovery (of course this is under the assumption that the risk variant is exomic) and allow for, in theory, the direct association between phenotype and causal variant. They have provided a new mechanistic perspective on the development of the complex disease gene mapping paradigm. Currently, with sequencing data, there is still a strong demand for more powerful and efficient analytic methods for novel gene discovery in the analysis of complex diseases.

We demonstrated the high quality of exome sequencing data in our samples collected from a study of panic disorder. We examined SNP quality within and outside the targeted exome regions. With the NimbleGen SeqCap capturing method, about 59% of the reads in our dataset were mapped within the target regions, meaning and there are still a significant number of reads that map elsewhere. About 30% of reads fall outside >200 bp of the exonic region, and 10% of reads are within 200 bp from the nearest target region. Variant call qualities were generally better for positions within the target regions, due to successful target enrichment. Furthermore, we computed genotype concordance with exome microarray data. The overall consistency rate was >99.83% and the heterozygous consistency rate was 98.14%, which suggests that the two platforms maintained a large amount of agreement over low frequency variants in the exonic regions.

Undoubtedly, the data generated in NGS technologies will continue to grow in terms of the depth per individual and the number of samples per dollar. The role of computation and bioinformatics becomes more and more crucial in the analysis and interpretation of sequencing data. Tremendous effort has been devoted to the development of tools for variant analysis in the process of quality control, alignment, variant identification, and downstream association studies. As whole-genome sequencing becomes prevalent in the next few years, future developments of workflow and pipelines will facilitate researches on human diseases.

ACKNOWLEDGMENTS

This work was partly supported by CTSA KL2 RR024138 (Zuoheng Wang), NIH grants DA030976, DA12849, DA12690, DA18432, AA017535, DA028909, AA11330, MH64122, a VA MERIT award (Joel Gelernter), NIDA career development award DA24758 (Bao-Zhu Yang), Brain and Behavior Research Foundation Young Investigator Award (Bao-Zhu Yang). We thank AnnMarie Lacobelle for DNA preparation and technical support, and Drs. Nicholas Carriero and Robert Bjornson for computational support. Next Generation Sequencing and Genotyping for Exome array were provided by the Yale Center for Genomic Analysis. Data analyses were performed using cluster at Yale University Biomedical High Performance Computing Center.

REFERENCES

Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Akey, J. M., Zhang, K., Xiong, M., and Jin, L. (2003). The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* 20, 232–242. doi: 10.1093/molbev/msg032

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *N. Biotechnol.* 25, 195–203. doi: 10.1016/j.nbt.2008.12.009

Bainbridge, M. N., Wang, M., Wu, Y., Newsham, I., Muzny, D. M., Jefferies, J. L., et al. (2011). Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol.* 12, R68. doi: 10.1186/gb-2011-12-7-r68

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M.

J., Nickerson, D. A., et al. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* 12, 745–755. doi: 10.1038/nrg3031

Bao, S., Jiang, R., Kwan, W., Wang, B., Ma, X., and Song, Y. Q. (2011). Evaluation of next-generation sequencing software in mapping and assembly. *J. Hum. Genet.* 56, 406–414. doi: 10.1038/jhg.2011.43

Biesecker, L. G. (2010). Exome sequencing makes medical genomics a reality. *Nat. Genet.* 42, 13–14. doi: 10.1038/ng0110-13

Botstein, D., and Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.* 33, 228–237. doi: 10.1038/ng1090

Burrows, M., and Wheeler, D. J. (1994). *A Block Sorting Lossless Data Compression Algorithm.* Technical Report 124. Palo Alto: CA: Digital Equipment Corporation.

Challis, D., Yu, J., Evani, U., Jackson, A., Paithankar, S., Coarfa, C., et al. (2012). An integrative variant analysis suite for whole exome next-generation sequencing data. *BMC Bioinformatics* 13:8. doi: 10.1186/1471-2105-13-8

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* 15, 1496–1502. doi: 10.1101/gr.4107905

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dean, J., and Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 107–113. doi: 10.1145/1327452.1327492

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* 11, 446–450. doi: 10.1038/nrg2809

Gilissen, C., Hoischen, A., Brunner, H. G., and Veltman, J. A. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol.* 12, 228. doi: 10.1186/gb-2011-12-9-228

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* 27, 182–189. doi: 10.1038/nbt.1523

Guo, Y., Long, J., He, J., Li, C. I., Cai, Q., Shu, X. O., et al. (2012). Exome sequencing generates high quality data in non-target regions. *BMC Genomics* 13:194. doi: 10.1186/1471-2164-13-194

Han, F., and Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Hum. Hered.* 70, 42–54. doi: 10.1159/000288704

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., et al. (2007). Genome-wide *in situ* exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527. doi: 10.1038/ng.2007.42

Huyghe, J. R., Jackson, A. U., Fogarty, M. P., Buchkovich, M. L., Stančáková, A., Stringham, H. M., et al. (2013). Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat. Genet.* 45, 197–201. doi: 10.1038/ng.2507

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M., and Lange, C. (2011). A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.* 7:e1001289. doi: 10.1371/journal.pgen.1001289

Ionita-Laza, I., Cho, M. H., and Laird, N. M. (2013). Statistical challenges in sequence-based association studies with population- and family-based designs. *Stat. Biosci.* 5, 54–70. doi: 10.1007/s12561-012-9062-9

Ji, H. P. (2012). Improving bioinformatic pipelines for exome variant calling. *Genome Med.* 4, 7. doi: 10.1186/gm306

Kaiser, J. (2010). Human genetics. Affordable 'exomes' fill gaps in a catalogue of rare diseases. *Science* 330, 903. doi: 10.1126/science.330.6006.903

Ku, C. S., Naidoo, N., and Pawitan, Y. (2011). Revisiting Mendelian disorders through exome sequencing. *Hum. Genet.* 129, 351–370. doi: 10.1007/s00439-011-0964-2

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25. doi: 10.1186/gb-2009-10-3-r25

Li, B., and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321. doi: 10.1016/j.ajhg.2008.06.024

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., et al. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.* 19, 1124–1132. doi: 10.1101/gr.088013.108

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., et al. (2009c). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 1966–1967. doi: 10.1093/bioinformatics/btp336

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* 18, 1851–1858. doi: 10.1101/gr.078212.108

Lin, D. Y., and Tang, Z. Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* 89, 354–367. doi: 10.1016/j.ajhg.2011.07.015

Madsen, B. E., and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 5:e1000384. doi: 10.1371/journal.pgen.1000384

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118. doi: 10.1038/nmeth.1419

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* 9, 387–402. doi: 10.1146/annurev.genom.9.081307.164359

Marian, A. J. (2012). Molecular genetic studies of complex phenotypes. *Transl. Res.* 159, 64–79. doi: 10.1016/j.trsl.2011.08.001

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369. doi: 10.1038/nrg2344

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Metzker, M. L. (2010). Sequencing technologies–the next generation. *Nat. Rev. Genet.* 11, 31–46. doi: 10.1038/nrg2626

Morris, A. P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* 34, 188–193. doi: 10.1002/gepi.20450

Musunuru, K., Pirruccello, J. P., Do, R., Peloso, G. M., Guiducci, C., Sougnez, C., et al. (2010). Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. *N. Engl. J. Med.* 363, 2220–2227. doi: 10.1056/NEJMoa1002926

Neuman, J. A., Isakov, O., and Shomron, N. (2012). Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief. Bioinform.* 14, 46–55. doi: 10.1093/bib/bbs013

Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., et al. (2010a). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42, 790–793. doi: 10.1038/ng.646

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., et al. (2010b). Exome sequencing identifies the cause of a Mendelian disorder. *Nat. Genet.* 42, 30–35. doi: 10.1038/ng.499

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276. doi: 10.1038/nature08250

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451. doi: 10.1038/nrg2986

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nat. Rev. Genet.* 12, 465–474. doi: 10.1038/nrg2989

Parla, J. S., Iossifov, I., Grabill, I., Spector, M. S., Kramer, M., and McCombie, W. R. (2011). A comparative analysis of exome capture. *Genome Biol.* 12, R97. doi: 10.1186/gb-2011-12-9-r97

Pattnaik, S., Vaidyanathan, S., Pooja, D. G., Deepak, S., and Panda, B. (2012). Customisation of the exome data analysis pipeline using a combinatorial approach. *PLoS ONE* 7:e30080. doi: 10.1371/journal.pone.0030080

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838. doi: 10.1016/j.ajhg.2010.04.005

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137. doi: 10.1086/321272

Rabbani, B., Mahdieh, N., Hosomichi, K., Nakaoka, H., and Inoue, I. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J. Hum. Genet.* 57, 621–632. doi: 10.1038/jhg.2012.91

Reich, D. E., and Lander, E. S. (2001). On the allelic spectrum of human disease. *Trends Genet.* 17, 502–510. doi: 10.1016/S0168-9525(01)02410-6

Ruffalo, M., LaFramboise, T., and Koyutürk, M. (2011). Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 27, 2790–2796. doi: 10.1093/bioinformatics/btr477

Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* 19, 212–219. doi: 10.1016/j.gde.2009.04.010

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Stitziel, N. O., Kiezun, A., and Sunyaev, S. (2011). Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol.* 12, 227. doi: 10.1186/gb-2011-12-9-227

Tennessen, J. A., O'Connor, T. D., Bamshad, M. J., and Akey, J. M. (2011). The promise and limitations of population exomics for human evolution studies. *Genome Biol.* 12, 127. doi: 10.1186/gb-2011-12-9-127

The International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature* 437, 1299–1320. doi: 10.1038/nature04226

The International HapMap Consortium. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861. doi: 10.1038/nature06258

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90, 7–24. doi: 10.1016/j.ajhg.2011.11.029

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565–569. doi: 10.1038/ng.608

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., and Yu, J. (2010). The next-generation sequencing technology and application. *Protein Cell* 1, 520–536. doi: 10.1007/s13238-010-0065-3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 March 2013; paper pending published: 26 April 2013; accepted: 04 August 2013; published online: 26 August 2013.

*Citation: Wang Z, Liu X, Yang B-Z and Gelernter J (2013) The role and challenges of exome sequencing in studies of human diseases. Front. Genet. **4**:160. doi: 10.3389/fgene.2013.00160*

This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics.