Published online 3 July 2010

Nucleic Acids Research, 2010, Vol. 38, No. 16 ***e164***
doi:10.1093/nar/gkq603

ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data

Kai Wang[1,*], Mingyao Li[2] and Hakon Hakonarson[1,3]

[1]Center for Applied Genomics, Children's Hospital of Philadelphia, [2]Department of Biostatistics and Epidemiology and [3]Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104, USA

Received March 27, 2010; Revised June 2, 2010; Accepted June 18, 2010

ABSTRACT

High-throughput sequencing platforms are generating massive amounts of genetic variation data for diverse genomes, but it remains a challenge to pinpoint a small subset of functionally important variants. To fill these unmet needs, we developed the ANNOVAR tool to annotate single nucleotide variants (SNVs) and insertions/deletions, such as examining their functional consequence on genes, inferring cytogenetic bands, reporting functional importance scores, finding variants in conserved regions, or identifying variants reported in the 1000 Genomes Project and dbSNP. ANNOVAR can utilize annotation databases from the UCSC Genome Browser or any annotation data set conforming to Generic Feature Format version 3 (GFF3). We also illustrate a 'variants reduction' protocol on 4.7 million SNVs and indels from a human genome, including two causal mutations for Miller syndrome, a rare recessive disease. Through a stepwise procedure, we excluded variants that are unlikely to be causal, and identified 20 candidate genes including the causal gene. Using a desktop computer, ANNOVAR requires ~4 min to perform gene-based annotation and ~15 min to perform variants reduction on 4.7 million variants, making it practical to handle hundreds of human genomes in a day. ANNOVAR is freely available at http://www.openbioinformatics.org/annovar/.

INTRODUCTION

High-throughput sequencing data have been produced at unprecedented rates for diverse genomes. There is a strong need for novel informatics and analytical strategies, including methods for sequencing reads alignment, variant identification, genotype calling and association tests, in order to take advantage of the massive amounts of sequencing data. There have been dozens of short read alignment software available now with different functionalities (1), as well as several single nucleotide variants (SNV) and copy number variant (CNV) calling algorithms (2). However, there is a paucity of methods that can simultaneously handle a large number of called variants (typically >3 million variants for a given human genome) and annotate their functional impacts, despite the fact that this is an important task in many sequencing applications. Even when sequencing only exonic regions for Mendelian diseases such as Freeman–Sheldon syndrome, each subject still carries a total of ~20 000 variants, but only two variants *in trans* are the true disease causal mutations (3). Therefore, identifying a small subset of functionally important variants from large amounts of sequencing data is important to pinpoint potential disease causal genes and causal mutations.

Several reasons motivate us to develop a functional annotation pipeline for genetic variants. First, although companies that manufacture sequencing machines or provide sequencing services typically offer software for functional annotation, these software are usually sequencing platform-specific, and cannot be extended to handle users' specific needs (such as using different genome builds or gene annotations). Second, although several databases have been developed for the functional annotation of SNPs or CNVs (4–6), most of them are limited to known variants, typically those reported in dbSNP or CNV databases. We note that some exceptions exist (7), for example, the F-SNP tool (8) and Seattle Seq tool (http://gvs.gs.washington.edu/SeattleSeqAnnotation/) can be used for annotation of novel SNPs. Third, several previously developed mutation prediction algorithms, such as SIFT (9) and PolyPhen (10), require building multiple alignments on sequence databases, can only handle non-synonymous mutations, and are difficult to scale up to many model organism genomes. Nevertheless, for human genomes, SIFT/PolyPhen scores for all possible non-synonymous mutations can be

*To whom correspondence should be addressed. Tel: +1 215 426 1256; Fax: +1 267 426 0363; Email: kai@openbioinformatics.org

computed, so they can be utilized for fast annotation of novel SNVs. Fourth, although it is feasible to build a database with pre-calculated annotation for all 9 billion possible SNVs in the human genome, such databases cannot be easily updated when new annotation information is available, and they cannot handle insertions or deletions. Finally, the development of many current databases and web servers are geared toward the human genome, and cannot be utilized when sequences from non-human genomes need to be annotated. Therefore, there is a strong community need for efficient, configurable, extensible and cross-platform compatible tools to utilize update-to-date information to annotate genetic variants from diverse genomes. The software that we present here, ANNOVAR (Annotate Variation), was developed to fill these unmet needs.

Besides annotating functional effects of variants with respect to genes, ANNOVAR has several other functionalities, including the ability to perform genomic region-based annotations, as well as the ability to compare variants to existing variation databases. Region-based annotations refer to the annotations of variants based on specific genomic elements other than genes, for example, conserved genomic regions, predicted transcription factor binding sites, predicted microRNA target sites and predicted stable RNA secondary structures. These annotations are especially important for whole-genome sequencing data, as the vast majority of variants will be outside of protein coding regions and their functional effects cannot be assessed by gene-based annotations. ANNOVAR can utilize annotation databases from the UCSC Genome Browser as flat text files; however, essentially any annotation database can be handled as long as they conform to Generic Feature Format version 3 (GFF3) standards (http://www.sequenceontology.org/gff3.shtml) for sequence-level feature annotations. Additionally, ANNOVAR can evaluate and filter out subsets of variants that are not reported in public databases such as dbSNP and the 1000 Genomes Project. Typically, rare variants causing Mendelian diseases are less likely to be present in these databases, or are unlikely to be present with high allele frequencies. This rationale has been used to enrich for subsets of variants in previous exome sequencing projects that identified causal mutations for Freeman–Sheldon syndrome (11) and Miller syndrome (3). ANNOVAR offers similar functionality but can extend the comparisons to other public databases such as the 1000 Genomes Project, which offers allele frequency information. Similarly, ANNOVAR can also filter variants against a user-compiled data set, such as all SIFT scores for all possible non-synonymous mutations in the human genome.

We will provide long-term support to the academic community for software usage issues. Additionally, we will continuously update the software to accommodate and take advantage of different sources of functional annotation, for example, annotations based on exome sequencing from the 1000 Genomes Project in the future. We believe that ANNOVAR will be useful to prioritize genetic variants from diverse genomes, and expedite scientific discoveries from the massive amounts of sequencing data produced from high-throughput sequencing platforms.

MATERIALS AND METHODS

Prepare input files with genetic variants

ANNOVAR is a command-line driven software tool and can be used as a standalone application on diverse hardware systems where standard Perl modules are installed. ANNOVAR is open-source, and is freely available at http://www.openbioinformatics.org/annovar/ to the academic community. ANNOVAR takes text-based input files, where each line corresponds to one genetic variant, including SNVs, insertions, deletions or block substitutions. In each line, the first five space- or tab-delimited columns represent chromosome, start position, end position, the reference nucleotide(s) and the observed nucleotide(s). For chromosome positions, ANNOVAR can handle 1-based coordinate system (by default) and half-open zero-based coordinate system (via the use of '–zerostart' argument). Additional columns can be supplied and will be printed out in identical form in output files. For convenience, users can use '0' to fill in the reference nucleotides, if this information is not readily available. Insertions, deletions or block substitutions can be readily represented by this simple file format, by using '–' to represent a null nucleotide. One example is given in Table 1, with extra columns that serve as comments on the variants.

In Table 1, the first variant is a SNV, with a substitution of C in the reference genome to T. The second variant is a single-base insertion, since the reference nucleotide in the reference genome is represented by '–'. The third variant is a single-base deletion, with the observed nucleotide being represented by '–'. The fourth variant is a block substitution, but the reference allele is represented by '0', eliminating the need to provide this allele explicitly on this line. The last variant is a deletion that spans several nucleotides.

Download annotation database

To annotate variants with respect to their functional consequences on genes, ANNOVAR needs to download gene annotation data sets (gene/transcript annotations and FASTA sequences) from the UCSC Genome Browser (12) and save them to local disk. Several different gene annotation systems, including RefSeq genes, UCSC Genes and the Ensembl genes, can be utilized for

Table 1. Example of an input file with five genetic variants

Chromosome	Start	End	Ref	Obs	Comments
16	49303427	49303427	C	T	R702W (*NOD2*)
16	49321279	49321279	–	C	c.3016_3017insC (*NOD2*)
13	19661685	19661685	G	–	35delG (*GJB2*)
1	105293754	105293755	0	ATAAA	Block substitution
1	13133880	13133881	TC	–	2-bp deletion (rs59770105)

annotation. The '–downdb' argument can be utilized for downloading necessary files automatically, if the computer is connected to the Internet. The 'wget' system command will be utilized for downloading, or the Net::Ftp/LWP::UserAgent modules (standard Perl modules installed in most systems by default) can be alternatively utilized. The users can specify different genome builds, such as hg18 (human), mm9 (mouse) or bosTau4 (cow), as long as they are available from the UCSC Genome Browser annotation databases. When performing gene-based annotations by Ensembl gene definitions (13), ANNOVAR will download the FASTA sequences from Ensembl as they were not available from the UCSC Genome Browser.

For region-based annotations, ANNOVAR needs to download annotation databases from the various UCSC Genome Browser tables, based on a user-specified track name. Alternatively, users can specify a custom-built annotation database conforming to Generic Feature Format 3 (GFF3), and ANNOVAR can identify variants overlapping with features annotated in the given GFF3 file. For filter-based annotations, for example, comparing mutations to those detected in the 1000 Genomes Project or dbSNP, ANNOVAR will download specific files from the corresponding websites. ANNOVAR can also download pre-computed SIFT scores for all human non-synonymous mutations, to help annotate human exomes by filter-based annotation procedure.

Scan annotation database

While reading variants from input file, ANNOVAR scans the gene annotation database stored at local disk, and identifies intronic variants, exonic variants, intergenic variants, 5′/3′-UTR variants, splicing site variants and upstream/downstream variants (less than a threshold away from a transcript, by default 1 kb). For intergenic variants, the closest two genes and the distances to them are reported. For exonic variants, ANNOVAR scans annotated mRNA sequences to identify and report amino acid changes, as well as stop-gain or stop-loss mutations. ANNOVAR can also perform region-based annotations on many types of annotation tracks, such as the most conserved elements and the predicted transcription factor binding sites. These annotations must be downloaded by ANNOVAR, before they can be utilized. Finally, ANNOVAR can filter specific variants such as SNPs with >1% frequency in the 1000 Genomes Project, or non-synonymous SNPs with SIFT scores >0.05.

To automate the procedure of reducing large amounts of variants into a small subset of functionally important variants, a script (auto_annovar.pl) is provided in the ANNOVAR package. By default, auto_annovar.pl performs a multi-step procedure by executing ANNOVAR multiple times, each time with several different command line parameters, and generates a final output file containing the most likely causal variants and their corresponding candidate genes. For recessive diseases, this list can be further trimmed down to include genes with multiple variants that are predicted to be functionally important.

Compilation of 'dispensable' genes

Based on the hypothesis that genes with high frequency of non-sense (stop-gain) mutations in population are unlikely to be causal for rare Mendelian diseases, we compiled a list of such 'dispensable' genes using data from the 1000 Genomes Project. For the CEU, YRI and JPT+CHB population separately, we identify genes that have non-sense mutations with combined minor allele frequency (MAF) >1%. For example, if two nonsense mutations in the same gene have MAF of 0.5 and 0.8% in CEU populations, this gene will be regarded as a dispensable gene. This analysis resulted in the identification of a total of 2064 genes from the 1000 Genomes Project. We caution that genes may fall within this list due to sequencing errors or alignment errors; for example, if the gene has many pseudogenes or if it is present within a segmental duplication. This list (~10% of all annotated human genes) is useful as a filtering step to further trim down potential candidate genes for Mendelian diseases.

Compilation of two synthetic data sets

To illustrate the utility of ANNOVAR in identifying causal genes for Mendelian diseases with recessive inheritance, we synthesized a whole-genome data set with ~4.2 million SNVs and ~0.5 million indels. These variants include all variants generated by Illumina on a male Yoruba subject (ftp://ftp.sanger.ac.uk/pub/rd/NA18507/) (14), as well as two known causal mutations for Miller syndrome (G->A mutation at chr16: 70608443 and G->C mutation at chr16: 70612611, representing G152R and G202A in the *DHODH* gene). We tested the variants reduction procedure on this data set using ANNOVAR, to examine whether we can identify a small subset of candidate genes that include the causal gene *DHODH*.

To illustrate the utility of ANNOVAR in identifying causal genes for Mendelian diseases with dominant inheritance, we synthesized whole-exome data sets. Since exome data for four Freeman–Sheldon cases were not available to us, we downloaded the exome data for eight HapMap subjects reported in (11). We then extracted the exome data for the first four subjects, including two Yoruba subjects (NA18507, NA18517) and two European Americans (NA12156 and NA12878). We next added the four known causal mutations to each of the four HapMap subjects (three C–>T mutations at chr17:10485359 and one C–>T mutation at chr17:10485360, representing R672H and R672C mutations in *MYH3*). We tested whether ANNOVAR can identify *MYH3* as the causal gene by examining exomes from these four subjects.

RESULTS AND DISCUSSION

Gene-based, region-based and filter-based annotation of genetic variants

To demonstrate the functionality and output of ANNOVAR, we analyzed the input file shown in Table 1. We applied gene-based annotation procedure using RefSeq gene definitions (15), though the UCSC Genes

definition (16) or Ensembl Gene definition (13) can be used alternatively. Two output files were generated, one of which annotates the location of each variant with respect to genes (one variant per output line), that is, whether it is exonic, intronic, intergenic, splicing site, 5′/3′-UTR, upstream/downstream of genes, or whether it has invalid input format. The other output file contains amino acid changes that may be caused by the mutation. We utilized a standardized nomenclature (17) to annotate non-synonymous SNVs and indels on cDNA or on proteins. For example, the first mutation has a functional consequence as NOD2:NM_022162:exon4:p.R702W, indicating that the mutation causes a non-synonymous change in exon 4 of the NOD2 gene. Since each gene may have multiple splicing isoforms in RefSeq annotations, the RefSeq transcript identifiers are always given after the gene name, and some variants may be annotated with respect to multiple alternative transcripts.

We next examined region-based annotations on the list of variants in Table 1. The '–regionanno' argument needs to be used, and a '–dbtype' argument needs to be specified to select an annotation database. For example, when '–dbtype mce44way' is issued, ANNOVAR will search through the phastConsElements44way annotation. This annotation database contains multiple alignments of 44 vertebrate species and measurements of evolutionary conservation using two methods, phastCons (18) and phyloP (19) on the human genome (UCSC build version: hg18). The second variant (c.3016_3017insC in NOD2) is located in a conserved region, with normalized conservation score of 392. We next used '–dbtype segdup' to identify variants located in segmental duplication regions (20), by examining the genomicSuperDups annotation database. The last variant in Table 1 (rs59770105, a small deletion) is identified within a segmental duplication, with sequence identity of 0.996 with the other copy of the sequence. In addition to the two examples above on region-based annotations, many other annotation databases from UCSC Genome Browser can be also interrogated. Additionally, ANNOVAR implemented the ability to directly interrogate any annotation databases conforming to Generic Feature Format version 3 (GFF3). Several examples of GFF3-based annotation databases are provided in the ANNOVAR website. GFF3 has become the standard annotation format among many established and emerging model organism databases, and provides a convenient standard for exchange of sequence feature annotation. Work with GFF3 files is widely supported by BioPerl and GMOD tools, thus it is relatively simple to convert most sequence feature annotations to GFF3 format. ANNOVAR leverages this standardization, so that users can utilize many custom built annotation databases for annotating genetic variants.

Finally, we examined the variants in Table 1 by filtering them against known variation databases such as dbSNP, the 1000 Genomes Project variation data, or user-supplied list of variants. The '–filter' argument in ANNOVAR is used for this purpose. The third variant in Table 1 (35delG in *GJB2*) is known to be an autosomal recessive mutation for deafness (21). Interestingly, it is annotated in dbSNP but not present in the 1000 Genomes Project. This example illustrates that filtering against dbSNP may sometimes fail to identify causal variants for Mendelian diseases, if the mutation is not private and has been well studied before. One additional useful feature of ANNOVAR is the ability to filter variants against pre-computed functional importance scores, such as SIFT scores (9), for all possible non-synonymous mutations in the human genome. For example, the R702W mutation in *NOD2* was annotated as deleterious by SIFT (score = 0). Since pre-computed SIFT scores can be used, ANNOVAR is very efficient in the annotation, requiring several minutes to handle an exome using a modern desktop computer.

Prioritization of genetic variants to identify candidate genes

To illustrate the utility of ANNOVAR in identifying causal genes responsible for rare Mendelian diseases, we synthesized a whole-genome data set with ~4.2 million SNPs and ~0.5 million indels. These variants include all variants generated by Illumina on a male subject (14), as well as two known causal mutations for Miller syndrome (G152R and G202A in the *DHODH* gene). Miller syndrome is a rare Mendelian disease recently solved by exome-sequencing on four probands and Sanger sequencing on three additional families (3), so the main goal of our experiment is to examine whether and how we can utilize ANNOVAR to trim down the potential candidate genes for a rare recessive disease. In addition, since one of the causal mutations was predicted as 'benign' when a function filter by PolyPhen is imposed (3), we investigated whether alternative and faster filtering strategies can be utilized. We acknowledge that since the full variants data from the original study were not available, here we have to rely on a synthesized variants set to illustrate the usage of ANNOVAR on real data.

An overview of the variants reduction procedure is illustrated in Figure 1. We first performed a gene-based annotation of all ~4.7 million variants, and identified a total of 24 617 exonic SNVs or indels. Given that Miller syndrome is a rare Mendelian disease, in line with the Ng *et al.* study (3), we next focused on 11 166 exonic protein-changing variants only, and identified a subset of 4860 variants falling in highly conserved genomic regions. We note that both causal mutations for Miller syndrome sit in highly conserved regions, with normalized score of 505 and 445, respectively (the normalized scores for all types of annotations range from 0–1000 in UCSC Genome Browser). One of the mutations was predicted as 'benign' by PolyPhen in the Ng *et al.* study and would have been missed, had Ng *et al.* utilized PolyPhen predictions in their filtering procedure (3). We confirmed that it was predicted as benign by SIFT as well (SIFT score = 0.18). We next filtered the variants from the 1000 Genomes project and dbSNP version 130, assuming that variants observed in public databases are less likely to be causal variants for Miller syndrome. This logic is similar to that used in two exome-sequencing studies (3,11), although they did not utilize 1000

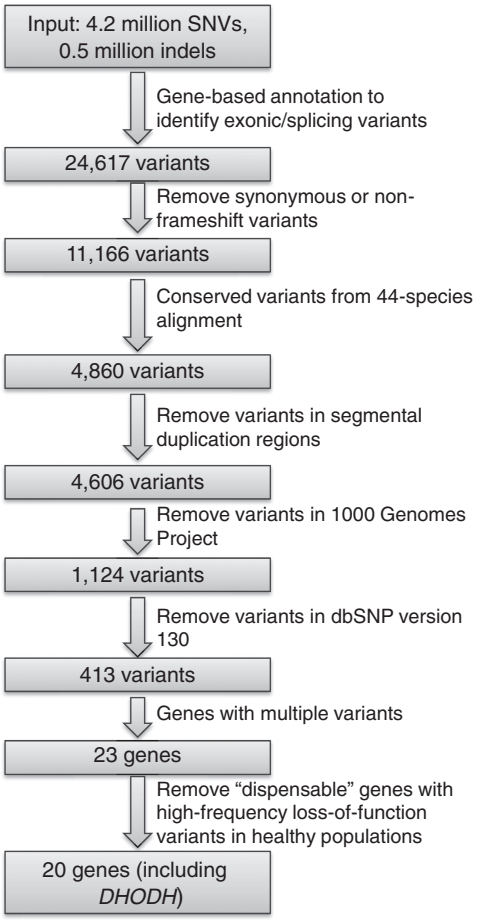

Figure 1. Identification of genes responsible for Miller syndrome using a synthetic data set. The input data set includes all SNVs and indels in subject NA18107 generated by Illumina, as well as two variants known to cause Miller syndrome. The variants reduction method can be implemented by an automation script (auto_annovar.pl) in the ANNOVAR package.

Genomes Project data sets. This process left us with 413 variants. Next, in the reduced variants set, we assessed whether multiple rare variants exist in the same gene as compound heterozygotes. Interestingly, only 23 genes were left by this analysis. Finally, we assessed whether some of these 23 genes belong to a group of 'dispensable' genes, that is, genes with high-frequency nonsense mutations (in >1% subjects) in the 1000 Genomes Project. The underlying rationale is that such genes are unlikely to be causal for a very rare Mendelian disease, or that such genes are highly susceptible to sequencing and alignment errors in short-read sequencing platforms. Three genes (*ZNF717*, *FAT1*, *OR4C3*) were deleted, and we were left with 20 candidate genes, including the causal gene *DHODH*. Each of the methods described above can be performed by different parameters in ANNOVAR, and we also provide a script that can be issued to perform the variants reduction procedure automatically. The procedure takes only ~15 min using a modern desktop computer for ~4.7 million variants.

The above analysis does not necessarily indicate that we can detect causal mutations for a rare Mendelian disease merely by sequencing one subject, and we acknowledge that we did not utilize the true exome-sequencing data. Nevertheless, the results suggest that filtering through a series of steps in ANNOVAR may help drastically reduce the number of candidate genes to a handful of genes that are human-manageable. In this regard, one could imagine that these 20 candidate genes can be sequenced in additional patients affected with Miller syndrome by conventional Sanger sequencing techniques, and the causal gene is likely to be identified directly from these additional sequencing runs.

Analysis of whole-exome variants data on multiple subjects

To examine the utility of ANNOVAR on identifying genes for autosomal dominant diseases, we next simulated the analysis presented in a study that sequenced eight exomes from HapMap subjects and four exomes from patients with Freeman–Sheldon syndrome (11). The full list of variants from eight HapMap exomes has been made publicly available, and the identity of four causal variants were presented in the original manuscript. Therefore, we synthesized four exome data sets, by taking four HapMap exomes and supplementing each with a known causal mutation for the Freeman–Sheldon syndrome. Each exome data set contains from 16 134 to 19 960 exonic variants. We next examined whether the variants reduction procedure can confidently identify the causal gene (*MYH3*) from the four exomes.

Similar to the procedure presented in the original publication (11), we assessed the number of candidate genes by utilizing variants from one, two, three or four subjects. When examining one subject, we were able to identify a set of 159 candidate genes. When examining two subjects, the number of candidate genes dropped substantially to 13. When three subjects were examined, only six candidate genes remained. When four subjects were examined, we can trim down the list of candidate genes to four (*HYDIN*, *KCNJ12*, *COL4A6*, *MYH3*). If we subsequently utilize SIFT scores (9) or PolyPhen scores (10) to evaluate the mutations in these four genes, we can further exclude *KCNJ12* and *COL4A6* as the causal genes. This analysis demonstrated that combined analysis of multiple genomes helped identify the causal gene for dominant Mendelian diseases, when all patients carry causal mutations at the same gene. However, we caution that for complex diseases or for Mendelian diseases where multiple causal genes exist, users need to consider the possibility of genetic heterogeneity in data analysis.

Efficiency of annotation on diverse genomes

One of the advantages of ANNOVAR is the relatively fast speed for annotation. The annotation is mostly based on pre-compiled annotation databases, without the need to generate new multiple sequence alignments or the need to interrogate remote SQL databases. To further demonstrate the performance and efficiency of ANNOVAR, we tested it on several additional data sets (Table 2), using a modern 64-bit Linux computer equipped with 3GHz Intel Xeon CPU and 8GB memory. First, we ran ANNOVAR

Table 2. Benchmark results for gene-based annotation on a computer with 3GHz Intel Xeon CPU

Genome	Data set	No. of variants	Timing	No. of exonic variants	Exonic fraction (%)
Human	Affymetrix 6.0 SNP array	930 006	1 m 2 s	8567	0.92
Human	1000 Genomes Project CEU[a]	9 633 115	8 m 35 s	53 199	0.55
Human	1000 Genomes Project YRI[a]	13 759 844	9 m 19 s	78 398	0.57
Human	1000 Genomes Project JPT+CHB[a]	10 970 708	8 m 32 s	63 793	0.58
Human	dbSNP 130	13 898 531	12 m 38 s	189 383	1.4
Mouse	dbSNP 128	14 864 829	8 m 42 s	157 745	1.1

[a]The list of variants were based on 2009 April release.

on ~1 million SNP markers on the Affymetrix Genome-Wide Human SNP 6.0 arrays, and compared the annotations with those provided by Affymetrix (version na30). We identified 271 SNPs that were annotated as exonic SNPs by ANNOVAR but not by Affymetrix. We manually confirmed that these SNPs are indeed exonic based on the latest UCSC annotation database information. This experiment illustrated the importance of annotating variants on-the-fly rather than relying on pre-calculated annotations that can be easily outdated.

Next, we tested ANNOVAR on ~9 million genetic variants identified in HapMap subjects from the 1000 Genomes Project, and discovered ~53 000, ~78 000 and ~63 000 exonic variants in CEU, YRI and JPT+CHB populations, respectively (Table 2). Compared to 1000 Genomes Project data, analysis of dbSNP data suggested that 1.4% of the variants disrupt exonic regions of genome, indicating a potential ascertainment bias in dbSNP toward functional SNPs (possibly due to the presence of many exon sequencing studies). Furthermore, we tested ANNOVAR on ~15 million SNPs in the mouse genome (that is, variants that differ between mouse strains). We identified 157 745 exonic variants (~1.1%), with slightly higher frequency than those observed in the 1000 Genomes Project. On average, it takes <1 min for every 1 million SNPs, so it is feasible to perform gene-based annotation on many hundreds of genomes in a day using a single personal computer.

In conclusion, ANNOVAR is a rapid, efficient tool to annotate functional consequences of genetic variation from high-throughput sequencing data. Additionally, ANNOVAR provides flexible variants reduction pipeline that helps pinpoint a specific subset of variants most likely to be causal for diseases or traits. With the rapid development and deployment of next-generation sequencing technologies, we expect that ANNOVAR will facilitate taking full advantage of the upcoming massive amounts of sequencing data to expedite scientific discoveries.

ACKNOWLEDGEMENTS

The authors thank Dr Yiran Guo (Children's Hospital of Philadelphia) and Dr Yufeng Shen (Columbia University) for providing valuable comments on ANNOVAR. We thank two anonymous reviewers for their suggestions on utilizing pre-computed SIFT scores for annotations and on handling GFF3-based annotation databases. The authors thank Bentley *et al.* and Ng *et al.* for making their variants data sets publicly available for testing ANNOVAR.

FUNDING

Funding for open access charge: Pilot/Methodological Study Award from National Institues of Health/ National Center for Research Resources Grant UL1 RR025774 (to K.W. and H.H.); R01HG004517 (to M.L.).

Conflict of interest statement. None declared.

REFERENCES

1. Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.
2. Dalca,A.V. and Brudno,M. (2010) Genome variation discovery with high-throughput sequencing data. *Brief. Bioinform.*, **11**, 3–14.
3. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.
4. Gamazon,E.R., Zhang,W., Konkashbaev,A., Duan,S., Kistner,E.O., Nicolae,D.L., Dolan,M.E. and Cox,N.J. (2010) SCAN: SNP and copy number annotation. *Bioinformatics*, **26**, 259–262.
5. Li,S., Ma,L., Li,H., Vang,S., Hu,Y., Bolund,L. and Wang,J. (2007) Snap: an integrated SNP annotation platform. *Nucleic Acids Res.*, **35**, D707–D710.
6. Ge,D., Zhang,K., Need,A.C., Martin,O., Fellay,J., Urban,T.J., Telenti,A. and Goldstein,D.B. (2008) WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res.*, **18**, 640–643.
7. Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief. Bioinform.*, **10**, 35–52.
8. Lee,P.H. and Shatkay,H. (2008) F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.*, **36**, D820–D824.
9. Ng,P.C. and Henikoff,S. (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
10. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
11. Ng,S.B., Turner,E.H., Robertson,P.D., Flygare,S.D., Bigham,A.W., Lee,C., Shaffer,T., Wong,M., Bhattacharjee,A., Eichler,E.E. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.
12. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

13. Curwen,V., Eyras,E., Andrews,T.D., Clarke,L., Mongin,E., Searle,S.M. and Clamp,M. (2004) The Ensembl automatic gene annotation system. *Genome Res.*, **14**, 942–950.
14. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
15. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
16. Hsu,F., Kent,W.J., Clawson,H., Kuhn,R.M., Diekhans,M. and Haussler,D. (2006) The UCSC known genes. *Bioinformatics*, **22**, 1036–1046.
17. den Dunnen,J.T. and Antonarakis,S.E. (2001) Nomenclature for the description of human sequence variations. *Hum. Genet.*, **109**, 121–124.
18. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
19. Pollard,K.S., Hubisz,M.J., Rosenbloom,K.R. and Siepel,A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.*, **20**, 110–121.
20. Bailey,J.A., Yavor,A.M., Massa,H.F., Trask,B.J. and Eichler,E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res.*, **11**, 1005–1017.
21. Gasparini,P., Rabionet,R., Barbujani,G., Melchionda,S., Petersen,M., Brondum-Nielsen,K., Metspalu,A., Oitmaa,E., Pisano,M., Fortina,P. *et al.* (2000) High carrier frequency of the 35delG deafness mutation in European populations. Genetic Analysis Consortium of GJB2 35delG. *Eur. J. Hum. Genet.*, **8**, 19–23.