Original Article

Healthc Inform Res. 2013 March;19(1):50-55.
http://dx.doi.org/10.4258/hir.2013.19.1.50
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

AnsNGS: An Annotation System to Sequence Variations of Next Generation Sequencing Data for Disease-Related Phenotypes

Young-Ji Na, PhD[1,2], Yonglae Cho, MS[1,2], Ju Han Kim, MD, PhD[1,2]

[1]Seoul National University Biomedical Informatics (SNUBI), Division of Biomedical Informatics, Seoul National University College of Medicine, Seoul;
[2]Systems Biomedical Informatics Research Center, Seoul National University, Seoul, Korea

Objectives: Next-generation sequencing (NGS) data in the identification of disease-causing genes provides a promising opportunity in the diagnosis of disease. Beyond the previous efforts for NGS data alignment, variant detection, and visualization, developing a comprehensive annotation system supported by multiple layers of disease phenotype-related databases is essential for deciphering the human genome. To satisfy the impending need to decipher the human genome, it is essential to develop a comprehensive annotation system supported by multiple layers of disease phenotype-related databases. **Methods:** AnsNGS (Annotation system of sequence variations for next-generation sequencing data) is a tool for contextualizing variants related to diseases and examining their functional consequences. The AnsNGS integrates a variety of annotation databases to attain multiple levels of annotation. **Results:** The AnsNGS assigns biological functions to variants, and provides gene (or disease)-centric queries for finding disease-causing variants. The AnsNGS also connects those genes harbouring variants and the corresponding expression probes for downstream analysis using expression microarrays. Here, we demonstrate its ability to identify disease-related variants in the human genome. **Conclusions:** The AnsNGS can give a key insight into which of these variants is already known to be involved in a disease-related phenotype or located in or near a known regulatory site. The AnsNGS is available free of charge to academic users and can be obtained from http://snubi.org/software/AnsNGS/.

Keywords: High-Throughput Nucleotide Sequencing, DNA Sequence Analysis, Molecular Sequence Annotation, Genome Structural Variation, Disease

Submitted: Feburary 8, 2013
Revised: March 18, 2013
Accepted: March 20, 2013

Corresponding Author
Ju Han Kim, MD, PhD
Seoul National University Biomedical Informatics (SNUBI), Seoul National University College of Medicine, 103 Daehak-ro, Jongno-gu, Seoul 110-799, Korea. Tel: +82-2-740-8320, Fax: +82-2-747-8928, E-mail: juhan@snu.ac.kr

I. Introduction

Next-generation sequencing (NGS) technology now provides a cost-effective approach to the large-scale sequencing of human samples for medical and population genetics. Within the field of personalized genomics, ambitious sequencing projects such as the 1000 Genomes Project [1], the Cancer Genome Atlas (http://cancergenome.nih.gov/), and the International Cancer Genome Consortium (http://www.icgc.org/) are seeking to identify genomic variants among human genomes and to use this knowledge to determine the genetic underpinnings of human diseases by associating variants

with symptoms. Several bioinformatic methods handling NGS data, including sequencing reads alignment, variant detection, and visualization of read sequences, have been developed. However, when considering that an important goal of NGS data is deciphering human genome sequences, it is imperative to annotate detected variants in terms of disease-related phenotypes.

Recently, tools designed for annotating variants in NGS data have been introduced, such as ANNOVAR [2], SeqAnt [3], and GAMES [4]. These tools annotate variants specifically with respect to genes harbouring variants, functional importance scores, and evolutionary conservation. ANNOVAR is a command-line tool that uses information from the UCSC Genome Browser to provide annotations. The SeqAnt is Web-based and can be downloaded, and it also relies on resources from the UCSC Genome Browser. The GAMES supports exome-Seq mutation discovery and functional annotation using the UCSC Genome Browser. All of them can place single nucleotide polymorphisms (SNPs) into functional classes, describe nearby genes, and indicate which SNPs are already described in dbSNP. None of them provides disease-related phenotype information, gives expression probe information for downstream analysis using microarray, or supports gene (or disease)-centric queries (Table 1).

In this article, we focus on disease-related phenotype annotations necessary for the analysis of variants. It is essential to develop a computational environment that can assess the likely functions of the variants observed, and whether or not they are present in existing variant databases. AnsNGS (annotation system of sequence variations for next generation sequencing data) is a new tool for annotating detected variants using disease-associated variants.

II. Methods

1. Input Files of the AnsNGS

Sequencing data can be uploaded and the results can be viewed using an interactive, Web-based graphical user interface (GUI) that has been tested for standard browsers. The AnsNGS takes text-based input files, where each line corresponds to one genetic variant. In each line, the first five tab-delimited columns represent chromosome, start position, end position, the reference nucleotide, and observed nucleotide. Additional columns can be supplied and will be printed in identical form in output files.

2. Genome-Wide Variant Annotation

The AnsNGS is a Web-based system that can comprehensively annotate variants detected with NGS data to provide a multifaceted approach for disease-related phenotype. The AnsNGS queries various tables in the UCSC Genome Browser [5], which provides a MySQL database annotation or any data set conforming to generic feature format ver. 3 (GFF3). As shown in Figure 1, the system consists of six annotators. The six annotators in the AnsNGS are a gene annotator, a microRNA annotator, an SNP annotator, a phenotype annotator, a disease annotator, and a probe annotator.

The gene annotator uses the refFlat table in the UCSC Genome Browser to obtain the genomic location of the nucleotide, the chromosome, the encoded protein(s), the

Table 1. Comparison of functionalities used in variant annotation applications

Functionalities	ANNOVAR	SeqAnt	GAMES	AnsNGS
Gene	RefSeq genes	RefSeq genes	RefSeq genes	RefSeq genes
	UCSC genes			
	Ensembl genes			
microRNA	-	Not used	Not used	miRBase
SNP	dbSNP	dbSNP	dbSNP	dbSNP
Phenotype	-	-	Not used	GAD
Disease	-	-	Not used	OMIM
Probe	-	Not used	Not used	Affymetrix
Conservation	phastCons	phastCons	phyloP	Not used
	phyloP			
Pathway	-	Not used	KEGG	Not used
Gene/disease-centric queries	-	Not used	Not used	Provided

NGS: next-generation sequencing, SNP: single nucleotide polymorphism, GAD: Genetic Association Database, OMIM: Online Mendelian Inheritance in Man, KEGG: Kyoto Encyclopedia of Genes and Genomes.

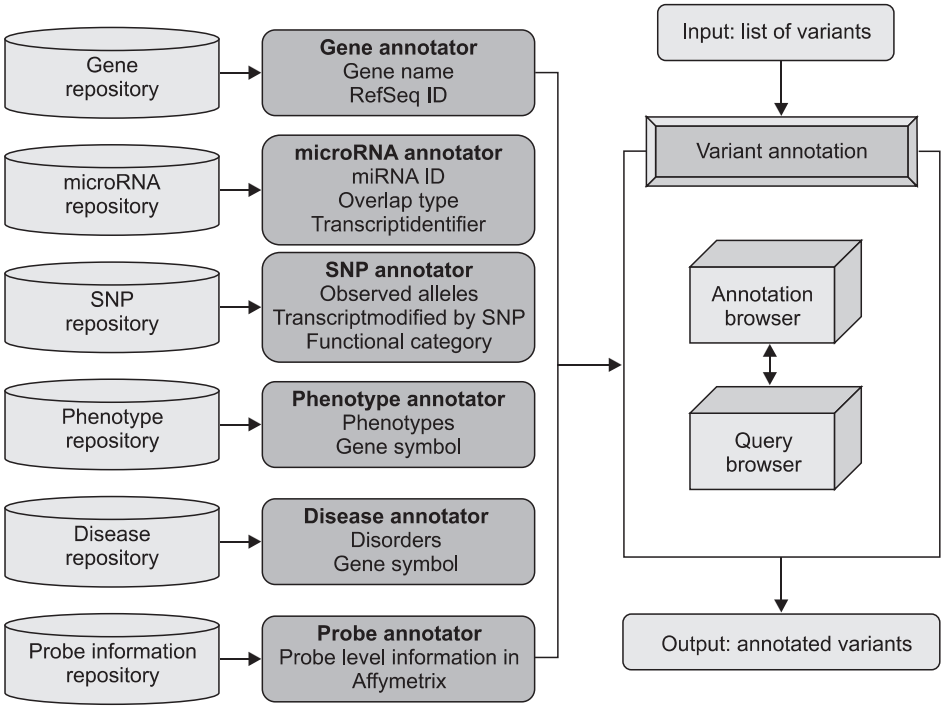

Figure 1. Overview of the AnsNGS. The AnsNGS takes text-based input files, obtains annotation information from the integrated databases, and returns the detailed annotation output to the user. The AnsNGS queries various tables in the UCSC database to extract the information for the reference genome or any data set conforming to Generic Feature Format ver. 3 (GFF3). The system consists of six annotators. The six annotators in the AnsNGS are a gene annotator, a microRNA annotator, an SNP annotator, a phenotype annotator, a disease annotator, and a probe annotator. NGS: next-generation sequencing, SNP: single nucleotide polymorphism.

gene, and the exon count (for all available isoforms). We use this information to determine the position of the mismatch in the gene and to determine whether it is in a coding/non-coding, exon, intron, untranslated region (UTR), or intron/exon junction region (intronic regions contiguous to exon starts and exon ends that are important to evaluate splice-site mutations).

The microRNA annotator employs the miRBase [6] to report microRNAs harbouring variants. The miRBase released in April 2010 includes 940 human microRNAs. From miRBase, the AnsNGS extracts microRNA names, host transcript genes, and genomic locations (e.g., intron, exon, intergenic, 5' UTR and 3' UTR).

The SNP annotator uses the snp130 table in the UCSC Genome Browser. The snp130 table contains almost 19 million SNP annotations, and it includes the first set of SNP calls from the 1000 Genomes Project. The tables are structured to include the position of the SNP on the genome assembly as well as additional information, such as sequences of the observed alleles from rs-fasta files, genotype counts, allele frequencies, kinds of mutation, and the pathogenetic significance of the SNPs, if reported.

The phenotype annotator uses the Genetic Association Database (GAD) which collates known polymorphism-disease associations [7] to provide disease-related phenotype information. Polymorphism-disease association data curated in this way are likely to comprise markers that occur in linkage disequilibrium with the presumed disease associated/functional variants.

The disease annotator takes omimGeneMap and omimMorbidMap tables in the UCSC Genome Browser. The AnsNGS provides annotations of known disease genes and the disease relevance of candidate genes using OMIM [8]. This type of information is critical for the proper analysis of whole-exome sequencing projects, in which mutations in different genes might be biologically involved in the same disease.

To connect NGS data analysis and microarray data analysis for the post hoc approach, reliable functional annotation of microarray probes is essential for the analysis and interpretation of the biological processes. The probe annotator provides annotation linking oligonucleotide probes to target genes, and it is essential for functional biological analysis of Affymetrix microarray experiments.

 http://dx.doi.org/10.4258/hir.2013.19.1.50

3. Annotated Output Files

Annotated output data can be displayed in several ways. The GUI provides a simple and convenient interface for the user to select the annotation information associated with a given variant or to highlight the variants by functional classes. The user is also able to download the annotated variation in a tab delimited text file format. The specific annotation field outputs are found in Table 2.

4. Validation Method

A total of two sequence datasets of various sizes and types were annotated by the AnsNGS. The first consisted of a 48-kb region, including the fragile X mental retardation 1 (FMR1) locus [9]. The second consisted of four HapMap exomes for Freeman-Sheldon syndrome [10]. Sequence data were uploaded to the AnsNGS website. The AnsNGS program determined the annotation information for the variable sites. The resulting output can be viewed on a Web-based GUI that is tested for standard browsers, in a series of downloadable tab-delimited text files (Figure 2).

Table 2. Annotation information used by AnsNGS to provide genomic annotation of NGS in human

Table	Information
Gene	Gene symbol, refseq ID, exon count
microRNA	microRNA ID, transcript gene, genomic location
SNP	dbSNP ID, functions
Phenotype	Broad/narrow phenotype, molecule phenotype, environmental factors
Disease	OMIM ID, description, gene symbol
Probe	Probe ID, gene symbol, chip name

NGS: next-generation sequencing, SNP: single nucleotide polymorphism, OMIM: Online Mendelian Inheritance in Man.

III. Results

As a direct demonstration of the utility of the AnsNGS for detecting functional variants, we annotated the data from a 48-kb resequencing experiment of a single human sample with a known coding sequence mutation at the FMR1 locus [9]. A total of 37 variant sites were identified and annotated in 0.13 seconds (based on searches with a single Intel(R) Xeon(R) CPU, 3.00 GHz), including the I304N mutation, which has been shown to result in intellectual disability. The AnsNGS showed that the variants are associated with mental retardation, X-linked, FRAXE type (OMIM ID: 309548). Disease-related phenotypes were global DNA methylation defect as well as ataxia and cognitive function. In terms of pathophysiology, a loss or shortage of FMRP disrupts the normal functions of nerve cells and, consequently, the nervous system, causing severe learning problems, intellectual disability, and the other features of fragile X syndrome. An *FMR1* gene mutation and the characteristic signs of fragile X syndrome also contribute features of autism spectrum disorders that affect communication and social interaction. In

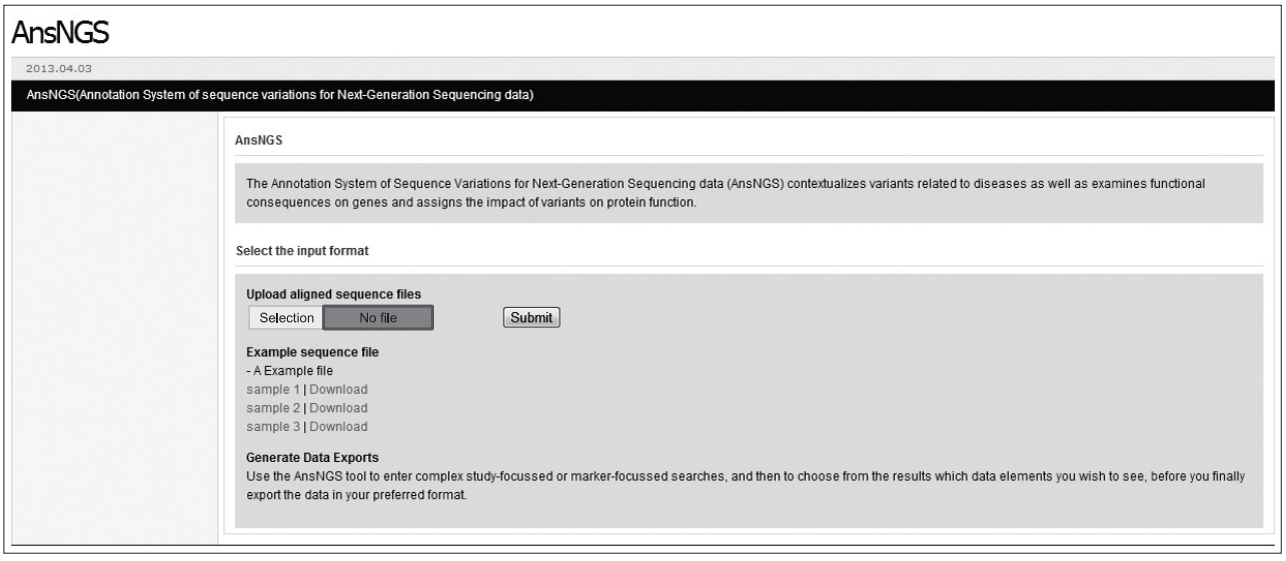

Figure 2. Interface of the AnsNGS. The main page of the AnsNGS is composed of three examples of sequencing data, such as whole genome or whole exome sequencing data. The example files are downloadable. The input file format of the AnsNGS is described in the Methods section. NGS: next-generation sequencing.

Human Genome U133A 2.0 Array (Affymetrix, Santa Clara, CA, USA), genes harbouring variants correspond to Affymetrix probe ID 203689_s_at and 215245_x_at.

To illustrate the utility of the AnsNGS in identifying causal genes for Mendelian diseases with dominant inheritance, we used whole-exome data sets. We downloaded the Freeman-Sheldon exome data for eight HapMap subjects. We then extracted the exome data for the first four subjects, including two Yoruba subjects (NA18507 and NA18517) and two European-Americans (NA12156 and NA12878). The AnsNGS showed that the variants are associated with arthrogryposis (OMIM ID: 160720). Disease-related phenotypes were carcinoma, squamous cell, and esophageal neoplasms in the Medical Subject Heading (MeSH) disease term. In Human Genome U133A 2.0 Array (Affymetrix), genes harbouring variants correspond to Affymetrix probe ID 205940_at. These results demonstrate the utility of annotating genetic variations with terms of disease-related phenotypes.

We have developed and validated the AnsNGS, a novel application for the annotation of NGS data with respect to disease-related phenotypes. Considering that it is important to identify induced variants responsible for a mutant phenotype directly, the AnsNGS is useful to discriminate silent mutations or polymorphisms from variants that are potentially associated with a phenotype or a disease. The AnsNGS generates concise and highly readable reports for a functional selection of important genetic events. In fact, the AnsNGS takes text-based input files, where each line corresponds to one genetic variant, including SNPs, insertions, deletions, or block substitutions.

IV. Discussion

Understanding the underlying complexity of disease phenotypes is a key to identifying causative mutations from NGS data. The annotation data reported by the AnsNGS can meet this requirement. Integrating the AnsNGS into a human DNA sequencing pipeline is powerful and versatile in clinical applications of NGS. The AnsNGS overcomes a significant bottleneck that can slow the wide-scale application of NGS for a host of genetics research and clinical genetics applications. The AnsNGS allows researchers to significantly narrow down the genomic regions of interest to their research, making an efficient and time-saving solution for everyone working in the NGS area and a perfect fit for any NGS analysis pipeline.

In conclusion, the AnsNGS is not only a new application for mining functional SNPs, insertions, and deletions from NGS data, but it also aids in the overall interpretation of large-scale sequencing data. The AnsNGS supports the identification of disease-associated genes and the prioritization of genes relevant to other diseases. The ultimate purpose of the AnsNGS is to provide biological insight into genetic events of disease-linked variants. The AnsNGS can lead to a translational bioinformatics approach and a paradigm shift in the aspect of biological interpretation, particularly in terms of the omics mechanisms of personalized medicine.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0000994). YJN's educational training was supported in part by a grant of the Korea Health 21 R&D Project, Ministry of Health, Welfare and Family Affairs, Republic of Korea (A112020).

References

1. Kaiser J. DNA sequencing: a plan to capture human diversity in 1000 genomes. Science 2008;319(5862):395.
2. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38(16):e164.
3. Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, et al. SeqAnt: a Web service to rapidly identify and annotate DNA sequence variations. BMC Bioinformatics 2010;11:471.
4. Sana ME, Iascone M, Marchetti D, Palatini J, Galasso M, Volinia S. GAMES identifies and annotates mutations in next-generation sequencing projects. Bioinformatics 2011;27(1):9-13.
5. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12(6):996-1006.
6. Griffiths-Jones S. miRBase: the microRNA sequence database. Methods Mol Biol 2006;342:129-38.
7. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet 2004;36(5):431-2.
8. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15(1):57-61.

 http://dx.doi.org/10.4258/hir.2013.19.1.50

9. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. Nat Methods 2007;4(11):907-9.
10. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461(7261):272-6.