Chapter 4

SeqAnt 2012: Recent Developments in Next-Generation Sequencing Annotation

Matthew Ezewudo, Promita Bose, Kajari Mondal, Viren Patel,
Dhanya Ramachandran and Michael E. Zwick

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53339

1. Introduction

The discovery of genome-wide genetic variation was central to the field of genomics [1,2]. Now, recent advances in second-generation sequencing technologies and better methods of targeted enrichment mean the detection of genome-wide patterns of genetic variation will soon be a routine operation [3,4]. Yet these advances in DNA sequencing have revealed a new bottleneck: the functional classification and interpretation of newly discovered genetic variation.

The scale of this problem is enormous. The high throughput and low cost of second-generation sequencing platforms now allow geneticists to routinely perform single experiments that identify tens of thousands to millions of variant sites in a single individual, but the methods that exist to annotate these variant sites using information from publicly available databases are too slow to be useful for the large sequencing datasets being generated. Because sequence annotation of variant sites is required before functional characterization can proceed, the lack of a high-throughput pipeline to annotate variant sites efficiently can be a major bottleneck in genetics research and clinical applications of genomics technologies.

To address this problem, we developed the Sequence Annotator (SeqAnt, http://seqant.genetics.emory.edu/), an open source web service and software package that rapidly annotates DNA sequence variants and identifies recessive or compound heterozygous loci in human, mouse, fly, and worm genome sequencing experiments [5]. Variants are characterized with respect to their functional type, frequency, and evolutionary conservation. Annotated variants can be viewed on a web browser, downloaded in a tab-delimited text file, or directly uploaded in a Browser Extensible Document (BED) format to the UCSC Genome Browser. To demonstrate the speed of SeqAnt, we annotated a series of publicly available datasets that ranged in size from 37 to 3,439,107 variant sites; the total time to annotate these data completely ranged from 0.17 seconds to 28 minutes 49.8 seconds.

1.1. Sequence annotation tools

Genome databases accessible via web browsers are very useful in the search for annotation information for DNA sequences. The UCSC Genome Browser web application has been a huge development of great value in analyzing and characterizing sequence information [6]. The application includes a variety of genomic tracks, assemblies, and browsers with genetic information from a host of species. The UCSC Genome Browser, with its various functionalities and annotation options, offers a one-stop shop for researchers, who can work directly on the web application by uploading their data, or they can download source codes of interest from the UCSC Genome Browser and run those locally. Despite its power, however, the main limitation we see in using the UCSC browser for sequence annotation lies in the limited amount of data that can be accessed at a given time, along with the need for human intervention. For example, it is time-consuming for geneticists who want annotation across multiple variant sites at once over different functional classes to use the browser comfortably. Ensembl is yet another superb broad-based web application with an expansive database, offering researchers choices on extracting specific regions of interest and annotating particular regions in the genome [7]. This application has various functionalities and tools that can accept uploaded data, convert formats of documents, and search for sequences of interest; still, like the UCSC browser, it is not the best choice for performing high-throughput sequencing annotation.

SNPnexus is a genetic variation tool developed to help determine functionally relevant SNPs for a given genomic region [8]. It has a user-friendly web interface that accepts inputs in the form of genomic positions, dbSNP id, or chromosomal region. The application database includes two different human genome assemblies: the hg19 and hg18 builds. SNPnexus generates calls on genomic mapping of variant sites, protein function consequences of such variants in the genome, the regulatory elements conserved within the region, and the conservation score of the variant site. The application also provides the genotype and allele frequencies estimation for known SNPs using data from the HapMap Project. This annotation tool, like so many others, is very useful for human variant annotation; however, it does not characterize variants in other species.

Since the development of SeqAnt in 2010, other software tools have come along to perform sequence annotation. Segtor is a tool designed to annotate large sets of genomic coordinates, intervals, single nucleotide variants (SNVs), indels, and translocations [9]. A more recent and very closely related annotation tool is AnnTools [10]. This is an open source web application that accepts user Inputs and queries their database for a full spectrum of variant site annotation, including single nucleotide variants, insertions and deletions, structural variants, and copy number variants. The application has a minimal memory footprint and likewise annotates variants quite rapidly. Nevertheless, AnnTools is restricted to human genome variant annotations and in this sense differs from SeqAnt, which annotates other species besides humans. There are also a number of other variant site annotation tools available either as downloadable command line applications or user interface web applications; these include snpEff (http://snpeff.sourceforge.net), MU2A[11], and Snat [12].

1.2. The distinction of SeqAnt

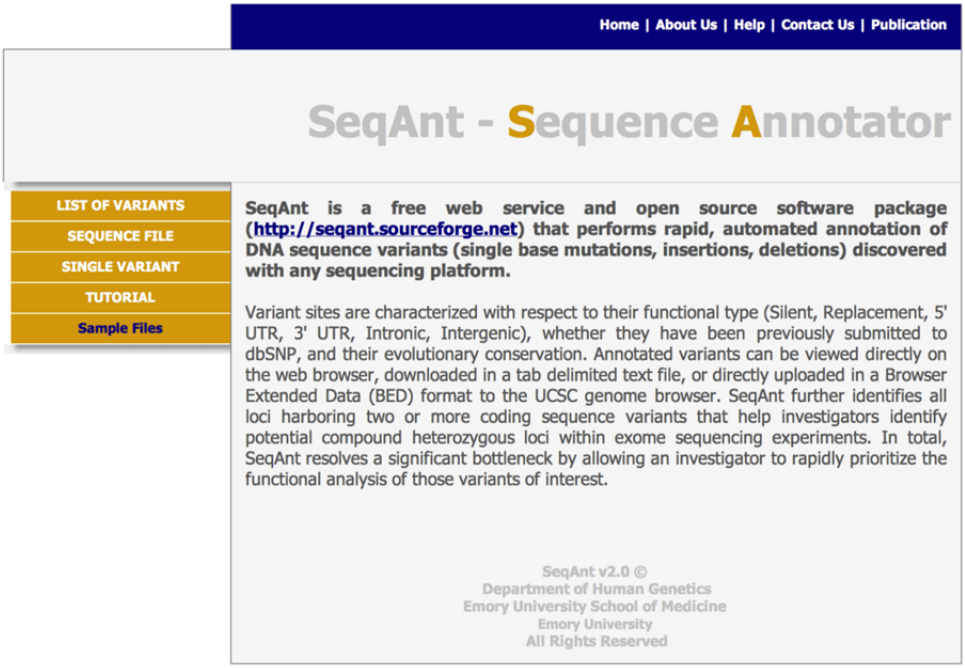
The uniqueness of SeqAnt versus all the other annotation tools we mentioned lies in three factors, which had been the key considerations for developing this technology to begin with. First, SeqAnt delivers annotations for multiple different species, ranging from primates to mammals, and now zebrafish and nematodes. Second, the web application has its own database updated from the UCSC website, which is a collection of binary files that drive the record speed with which large genomic data are annotated. Third, in addition to speed, the memory footprint is quite minimal, as data stored in binary files enable individuals from the public to download both the source file and database and locally run the application without elaborate computing apparatus. Some of the other tools mentioned have one or two of these unique features, but none have the robustness that comes from combining all three approaches to efficiently annotate variants and make meaningful functional calls across species, like SeqAnt does. Overall, we believe these represent important changes to SeqAnt that will be of broad utility to researchers using next-generation sequencing platforms in a wide variety of systems. SeqAnt will continue to be a fully open source web service and software package, and we believe it will prove especially useful for those investigators who lack dedicated bioinformatics personnel or infrastructure in their laboratories.

2. Upgraded features of SeqAnt 2.0

Since the initial publication of SeqAnt, we made a number of improvements that have been incorporated into SeqAnt 2.0 [5]. These modifications fall into four main categories. The first focused on updating the SeqAnt website (http://seqant.genetics.emory.edu). The second includes major changes made to the content and structure of the underlying binary databases that hold the annotation information. The third involves a significant redesign of the directory structure holding the output files. Finally, the last modification included substantial revisions to the number and content of output files themselves. Each of these updates will be described in greater detail in the sections that follow.

2.1. SeqAnt 2.0 - website updates

We undertook a major redesign of the SeqAnt web interface to make it more user-friendly. On the home page, we eliminated redundant tabs and buttons, simplified the overall design, and upgraded the graphic interface's color scheme (Figure 1). This page includes basic information about the original publication of SeqAnt [5], a link to contact the Zwick laboratory, and the web URL for the the SourceForge website (http://seqant.sourceforge.net), where the source code and associated binary libraries can be freely downloaded. From this page, the user is able to quickly access the three main types of input data accepted by SeqAnt. These include **SEQUENCE FILE**, **LIST OF VARIANTS**, and **SINGLE VARIANT**. In addition, the user can choose to view a **TUTORIAL** or select a set of **SAMPLE FILES** to gain experience performing analyses with the SeqAnt.


Home | About Us | Help | Contact Us | Publication
SeqAnt - Sequence Annotator
LIST OF VARIANTS
SEQUENCE FILE
SINGLE VARIANT
TUTORIAL
Sample Files

SeqAnt is a free web service and open source software package (http://seqant.sourceforge.net) that performs rapid, automated annotation of DNA sequence variants (single base mutations, insertions, deletions) discovered with any sequencing platform.

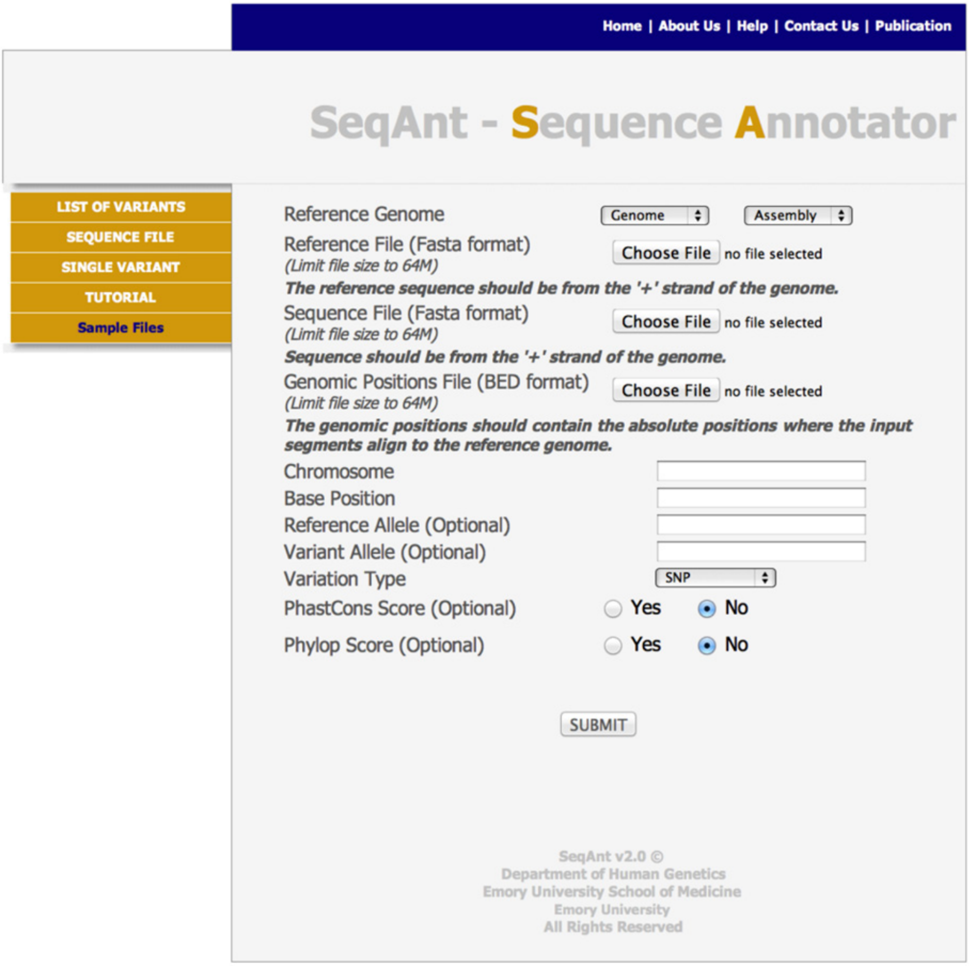
Variant sites are characterized with respect to their functional type (Silent, Replacement, 5' UTR, 3' UTR, Intronic, Intergenic), whether they have been previously submitted to dbSNP, and their evolutionary conservation. Annotated variants can be viewed directly on the web browser, downloaded in a tab delimited text file, or directly uploaded in a Browser Extended Data (BED) format to the UCSC genome browser. SeqAnt further identifies all loci harboring two or more coding sequence variants that help investigators identify potential compound heterozygous loci within exome sequencing experiments. In total, SeqAnt resolves a significant bottleneck by allowing an investigator to rapidly prioritize the functional analysis of those variants of interest.

SeqAnt v2.0 ©
Department of Human Genetics
Emory University School of Medicine
Emory University
All Rights Reserved

Figure 1. Screenshot of new SeqAnt 2.0 home page

Selecting the **SEQUENCE FILE** option returns the web interface shown in Figure 2. A typical use of this feature is when the user wants variation annotation information in a genomic region from a particular chromosome. Three different input files are accepted. The first is a reference sequence file in FASTA format of the entire genomic region being annotated. The second is a sequence file containing multiple FASTA sequences from a sequencing experiment, with each FASTA sequence representing a chromosomal region. The third is a genomic position file in the BED format which represents the coordinates for each of the chromosomal regions in the sequence file. The sequences in both the reference file and the sequence file should be in the positive orientation to ensure accurate annotation. The user is provided the option to choose a reference genome and assembly that will be used for annotating variant sites.

Selecting the **LIST OF VARIANTS** option returns the web interface shown in Figure 3. Only one input file is required to use this feature, the variations list file, which contains a listing of variant sites and the chromosomal regions of these sites, the minor allele and the reference allele. The variant list file is basically a pileup file, with a '.snp' or a '.txt' extension. If the PEMapper option were selected in this interface, the variation list file would be modified to include the sample ID for each individual within the experimental study where the sequence data was generated, if multiple individual samples were being analyzed. This particular (List of Variants) feature is very useful for researchers who want to perform genetic variation analysis (such as whole exome annotation) over a wide expanse of the genome.



Reference Genome Genome Assembly

Reference File (Fasta format)
(Limit file size to 64M) no file selected

The reference sequence should be from the '+' strand of the genome.

Sequence File (Fasta format)
(Limit file size to 64M) no file selected

Sequence should be from the '+' strand of the genome.

Genomic Positions File (BED format)
(Limit file size to 64M) no file selected

The genomic positions should contain the absolute positions where the input segments align to the reference genome.

Chromosome

Base Position

Reference Allele (Optional)

Variant Allele (Optional)

Variation Type SNP

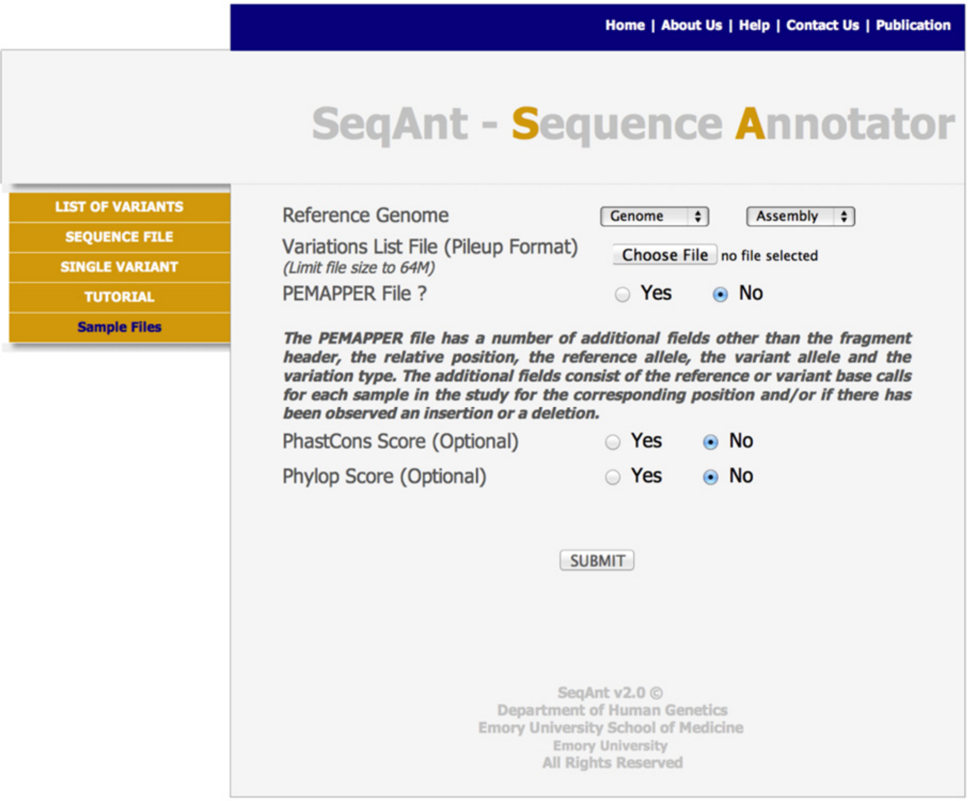
PhastCons Score (Optional) ○ Yes ● No

Phylop Score (Optional) ○ Yes ● No

SeqAnt v2.0 ©
Department of Human Genetics
Emory University School of Medicine
Emory University
All Rights Reserved

Figure 2. Screenshot of the SEQUENCE FILE page

Selecting the **SINGLE VARIANT** option returns the web interface shown in Figure 4. The user is provided the option to choose a reference genome and assembly that will be used for annotating a single variant site.The user then only needs to provide a chromosome and base position to obtain the annotation information.


Home | About Us | Help | Contact Us | Publication
SeqAnt - Sequence Annotator
LIST OF VARIANTS
SEQUENCE FILE
SINGLE VARIANT
TUTORIAL
Sample Files
SUBMIT

Reference Genome Genome Assembly

Variations List File (Pileup Format)
(Limit file size to 64M) Choose File no file selected

PEMAPPER File ? ○ Yes ● No

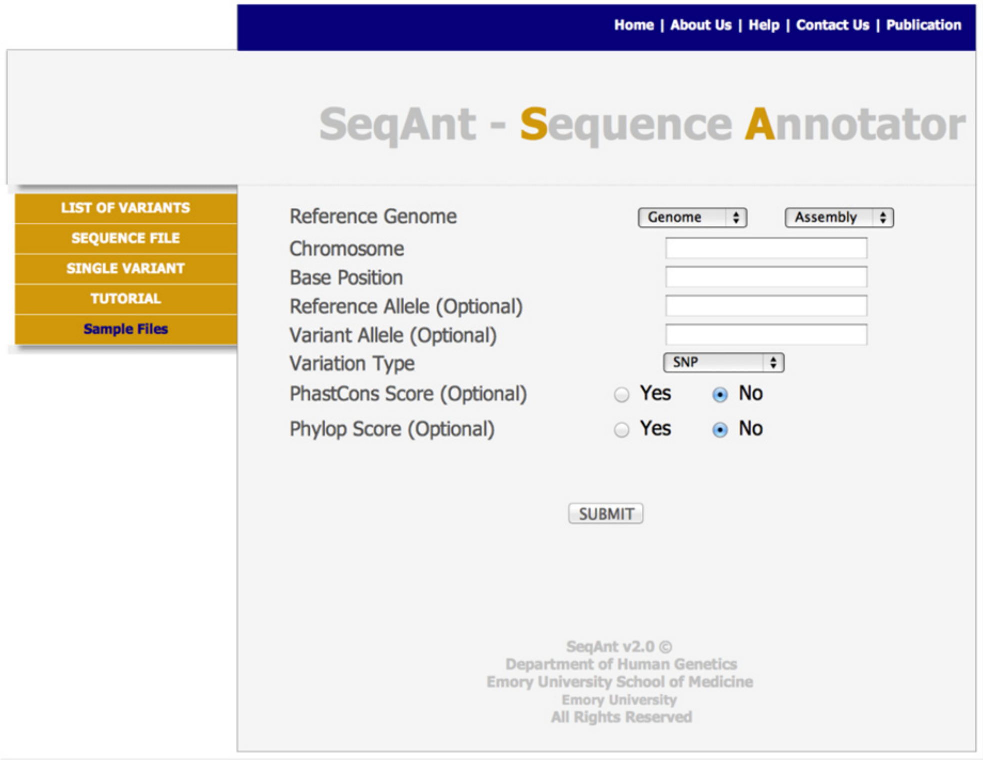
The PEMAPPER file has a number of additional fields other than the fragment header, the relative position, the reference allele, the variant allele and the variation type. The additional fields consist of the reference or variant base calls for each sample in the study for the corresponding position and/or if there has been observed an insertion or a deletion.

PhastCons Score (Optional) ○ Yes ● No

Phylop Score (Optional) ○ Yes ● No

SeqAnt v2.0 ©
Department of Human Genetics
Emory University School of Medicine
Emory University
All Rights Reserved

Figure 3. Screenshot of the LIST OF VARIANTS page


Home | About Us | Help | Contact Us | Publication
SeqAnt - Sequence Annotator
LIST OF VARIANTS
SEQUENCE FILE
SINGLE VARIANT
TUTORIAL
Sample Files
SUBMIT
SeqAnt v2.0 ©
Department of Human Genetics
Emory University School of Medicine
Emory University
All Rights Reserved

Reference Genome: Genome, Assembly

Chromosome

Base Position

Reference Allele (Optional)

Variant Allele (Optional)

Variation Type: SNP

PhastCons Score (Optional) ○ Yes ● No

Phylop Score (Optional) ○ Yes ● No

Figure 4. Screenshot of the SINGLE VARIANT page

2.2. SeqAnt 2.0 - Binary database upgrades

One of the unique features of SeqAnt is the ease and speed with which variant information is accessed from a set of customized binary databases. The SeqAnt binary databases are created from flat text table files obtained from the UCSC Genome Browser website [6]. Five main types of data constitute the SeqAnt binary databases. These include:

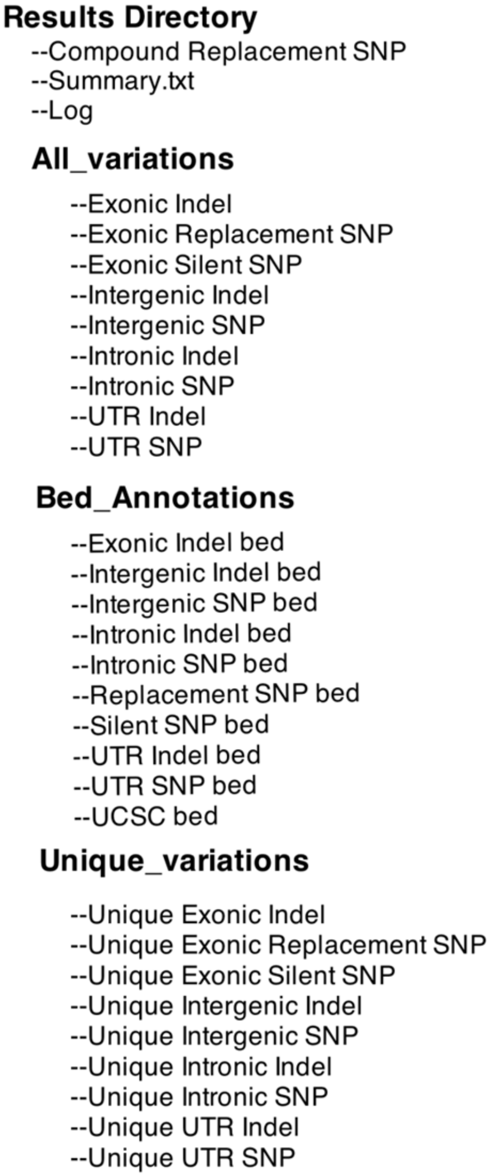
1. Reference Genome Sequence
2. RefGene Annotation
3. dbSNP Variation Data
4. PhastCons Evolutionary Conservation Scores
5. PhyloP Evolutionary Conservation Score

Standard queries, implemented through the web interfaces described above, are able to extract the annotation information from the binary databases. The actual structure of the binary databases is not directly visible to a SeqAnt user, but is worth examining in greater detail. The Reference Genome Sequence provides the basic backbone for other annotation information. Reference sequences for a given species are organized by different builds (i.e. human genome 18, human genome 19). Within each build, data are organized by chromosome, which reflects the structure of the flat files obtained from UCSC. The RefGene Annotation is the collection of information pertaining to known genes for a given species and build. This information is also organized by chromosome. The collection of variant sites in a given species is contained within the dbSNP Variation Data that is also organized by chromosome. Finally, the SeqAnt 2.0 binary databases include two different measures of evolutionary conservation for all sites in a given reference genome sequence. The PhastCons score is best used to detect functional elements in noncoding sequences, whereas the phyloP score provides a measure of the evolutionary conservation of single sites and is most useful for evaluating sites located in coding regions of genes.

Binary files are significantly smaller than their corresponding flat files, so querying binary files uses less memory than the same analysis performed with a flat file. Considering the vast amount of data that has to be accessed during sequence annotation of large genomic regions, the significant difference in the size of the binary files versus flat files helps to account for the speed with which information is processed using binary files. SeqAnt 2.0 updated a number of these specific binary files; a detailed description of the changes follows in the next sections.

2.2.1. Upgrade of dbSNP to SNP132 Track for hg19 Assembly (Homo sapiens)

The original goal of the dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/) was to develop a comprehensive catalog of common (>5% frequency) human genetic variation [13,14]. These variants were subsequently validated by genotyping in multiple human populations, and their patterns of statistical correlation among variants, known as linkage disequilibrium, were revealed in the HapMap project [15,16]. SeqAnt 1.0 included data from the SNP131 track from the dbSNP [17]. SeqAnt 2.0 was updated to the SNP132 build, which was characterized and uploaded to the UCSC Genome Browser in the summer of 2011. SNP132 has an expanded collection of variant sites that can help researchers determine whether an identical variant has been seen before in a different individual.



```
Results Directory
  --Compound Replacement SNP
  --Summary.txt
  --Log
  All_variations
     --Exonic Indel
     --Exonic Replacement SNP
     --Exonic Silent SNP
     --Intergenic Indel
     --Intergenic SNP
     --Intronic Indel
     --Intronic SNP
     --UTR Indel
     --UTR SNP
  Bed_Annotations
     --Exonic Indel bed
     --Intergenic Indel bed
     --Intergenic SNP bed
     --Intronic Indel bed
     --Intronic SNP bed
     --Replacement SNP bed
     --Silent SNP bed
     --UTR Indel bed
     --UTR SNP bed
     --UCSC bed
  Unique_variations
     --Unique Exonic Indel
     --Unique Exonic Replacement SNP
     --Unique Exonic Silent SNP
     --Unique Intergenic Indel
     --Unique Intergenic SNP
     --Unique Intronic Indel
     --Unique Intronic SNP
     --Unique UTR Indel
     --Unique UTR SNP
```

Figure 5. Contents of SeqAnt Output Directory. Directories are in bold; individual files shown in a standard font face.

2.2.2. Addition of PhyloP46way Conservation Score Database for hg19 Assembly (Homo sapiens)

The phyloP Evolutionary Conservation Score data type is a new addition to SeqAnt 2.0. Binary databases, including phylopP scores from a 46-way alignment of vertebrate species to the human genome, were included to complement the PhastCons Evolutionary Conservation Scores previously included in the application. The phyloP scores predict the probability of a given variant site having undergone evolution over time. The absolute phyloP values represent negative log p-values for the null hypothesis that there was no evolution across the regions annotated [18]. Regions that are more conserved tend to have more positive values, whereas sites believed to be fast evolving have negative values. The medium range of these scores for the 46-way alignment from the UCSC Genome Browser is between approximately -3 and +3. It should be noted that, unlike PhastCons, which takes into account flanking bases on a sequence in arriving at its final score for a given variant site, phyloP scores are computed by basically comparing the particular base in the sequence with aligned bases from other species [18]. Variations in highly conserved regions often suggest a significant change that could have functional implications. The PhyloP46way dataset we have on the upgraded SeqAnt web application is the most recent phyloP track in the UCSC, released in December 2009.

2.2.3. Addition of Full Genome Data Set by Chromosome of Zebrafish (danRer6 Assembly)

We selected zebrafish (*Danio rerio*) as the next species to be incorporated into the SeqAnt database because of its emergence as a model organism for a wide range of scientific studies, from behavioral genetics to drug modeling studies and integrative physiology [19,20]. SeqAnt 2.0 has now been updated to include binary files for the genome sequence of zebrafish. We derived binary databases for the first four data types from flat table files on the UCSC Genome Browser website. Flat table files for the phyloP evolutionary conservation score were not available and were therefore not included. The reference genome binaries use the danRer6 assembly, which annotated the datasets by chromosome and was released in December of 2008. The RefGene annotation and dbSNP variation data are relative to the danRer6 assembly. PhastCons evolutionary conservation scores were derived from multiple alignment between seven species and zebrafish. Including the zebrafish in SeqAnt 2.0 should prove valuable for researchers who work with this species.

2.3. SeqAnt 2.0: output directory structure and files

Significant changes to the number and types of output files are reflected in a new output directory structure in SeqAnt 2.0. The output from SeqAnt is contained within a Results directory that includes three subdirectories (Figure 5). This Results directory has the name of the original SeqAnt input file and a subscript '_Annotation_Files'. Within this directory, there are three distinct directories (All_Variations, BED_Annotation, Unique_Variations) holding the output of SeqAnt, which will be described in detail below. This directory also contains three other files of interest to a user. The first is a *.summary.txt file that provides a summary of all the variants annotated by SeqAnt. The second is a Compound.Replacement file that identifies variants, genes, and sample identifiers for those loci with two or more replacement variants. The collected list of variants includes those that could be compound recessive in a given individual, although since the phase of the variants is not determined, this would have to be validated by other means. This file may be useful when looking for genes that harbor variants that may fit a recessive loss-of-function model. The last is a *.log file generated by SeqAnt that records the major events that occur when SeqAnt processes a dataset.

2.3.1. All_variations directory

This directory contains the complete variant annotation files obtained from annotating input files with SeqAnt 2.0 (Figure 5). Two main types of genetic variation are annotated by SeqAnt: single nucleotide variants (SNPs) and insertions/deletions (INDELs). For SNPs, a given variant site when annotated belongs in one of five functional classifications. These include exonic.replacement, exonic.silent, untranslated region (UTR), intronic, or intergenic. For INDELs, a given variant when annotated belongs in one of four functional classifications. These include exonic, UTR, intronic, or intergenic. Overall, there are a total of nine files that contain the variants and their associated annotation information. These annotation files include all possible splice variants impacted by a given variant site. Thus, a given variant site may be listed multiple times in one of the nine output files.

2.3.2. BED_annotation directory

This directory contains files in BED format (http://genome.ucsc.edu/FAQ/FAQformat) that can be visualized on the UCSC Genome Browser or other viewer able to process files in this format. There are ten files total in this directory. Nine of the files include the variants and annotation information as described above; the tenth file (*.ucsc.bed) contains all the annotation information from each of the nine files in a single BED file for the entire genomic region to be visualized. These files can be uploaded to the UCSC browser as custom tracks to be visualized. They can also be visualized in other software packages that process BED files, such as the Integrative Genomics Viewer (Version 2.1) [21].

2.3.3. Unique_variations directory

In contrast to the annotation in the All_Variations directory, the Unique_Variations directory contains nine files that contain a single variant annotation for each SNP or INDEL. Thus, each variant is listed just once, regardless of the number of different splice variants it is predicted to impact. These files allow the user to quickly determine the total number of variants for any specific functional class.

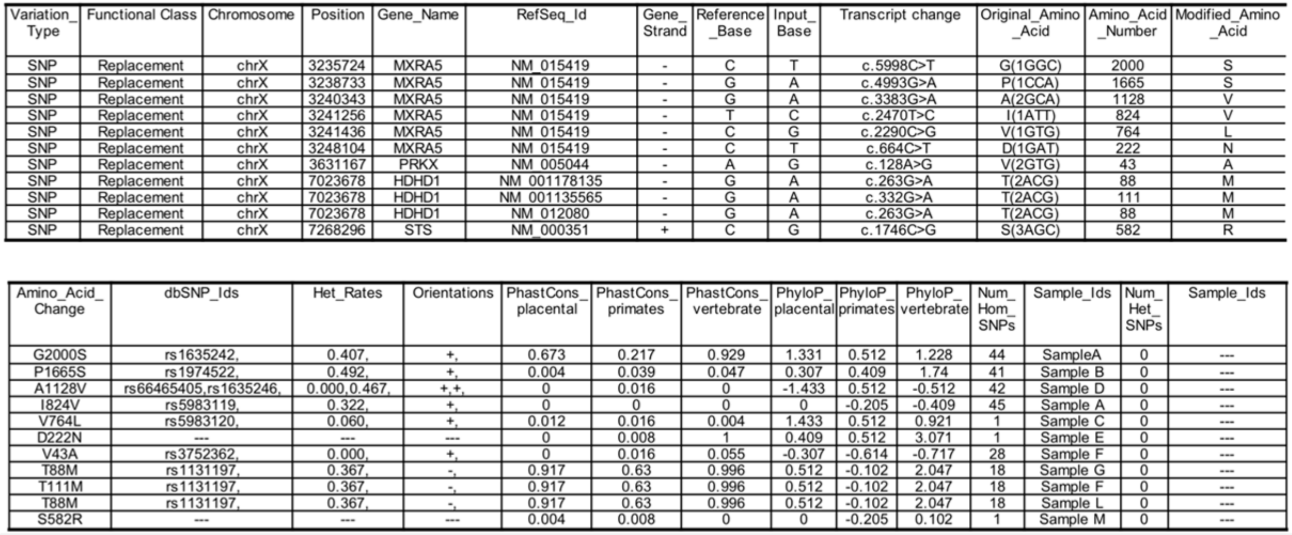
2.4. SeqAnt 2.0 - Output files

2.4.1. Redesign of Result Columns for Annotation Files

We introduced a number of changes to the annotation fields contained within the SeqAnt output files. First, we rearranged the order of columns in the output files to aid users in evaluating their results. Second, we introduced additional feature columns to the output files. These included row 10, which depicts the transcript change that occurs for a coding sequence variant, row 14, which shows the concomitant amino-acid change for a coding sequence variant, and rows 21 and 22, which report the phyloP conservation score values for each variant position annotated. A summary of the annotation information provided by SeqAnt 2.0 is shown below in Table 1. A representation of an example output file is shown in Figure 6 below.

Field ID	Annotation Field	Description
1	Variation_Type	Type of variant
2	Functional Class	Annotated functional category for variant site
3	Chromosome	Chromosome containing variant site
4	Position	Absolute position of variant site on a chromosome
5	Gene_Name	Name of locus containing variant site
6	RefSeq_ID	Ref_Seq ID from UCSC track
7	Gene_Strand	Orientation of locus
8	Reference_Base	Reference allele at variant site
9	Input_Base	Minor allele at variant site
10	Transcript Change	Nucleotide base change on transcript
11	Original_Amino_Acid	Reference amino acid at variant site
12	Amino_Acid_Number	Position of amino acid on peptide chain
13	Modified_Amino_Acid	Modified amino acid due to variant site
14	Amino_Acid_Change	Amino acid change on peptide chain
15	dbSNP_IDs	dbSNP ID If variant site has been reported
16	Het_Rates	dbSNP heterozygosity of reported variant site
17	Orientation	dbSNP orientation of reported variant site
18	PhastCons_placentals	Placental PhastCons score for variant site (46way)
19	PhastCons_primates	Primate PhastCons score for variant site (46way)
20	PhastCons_vertebrate	Vertebrate PhastCons score for variant site (46way)
21	PhyloP_placental	Placental phyloP score for variant site (46way)
22	PhyloP_primates	Primate phyloP score for variant site (46way)
23	PhyloP_vertebrate	Vertebrate phyloP score for variant site (46way)

Table 1. Annotation information output by SeqAnt 2.0



Variation_ Type	Functional Class	Chromosome	Position	Gene_Name	RefSeq_Id	Gene_ Strand	Reference _Base	Input_ Base	Transcript change	Original_Amino _Acid	Amino_Acid _Number	Modified_Amino _Acid
SNP	Replacement	chrX	3235724	MXRA5	NM_015419	-	C	T	c.5998C>T	G(1GGC)	2000	S
SNP	Replacement	chrX	3238733	MXRA5	NM_015419	-	G	A	c.4993G>A	P(1CCA)	1665	S
SNP	Replacement	chrX	3240343	MXRA5	NM_015419	-	G	A	c.3383G>A	A(2GCA)	1128	V
SNP	Replacement	chrX	3241256	MXRA5	NM_015419	-	T	C	c.2470T>C	I(1ATT)	824	V
SNP	Replacement	chrX	3241436	MXRA5	NM_015419	-	C	G	c.2290C>G	V(1GTG)	764	L
SNP	Replacement	chrX	3248104	MXRA5	NM_015419	-	C	T	c.664C>T	D(1GAT)	222	N
SNP	Replacement	chrX	3631167	PRKX	NM_005044	-	A	G	c.128A>G	V(2GTG)	43	A
SNP	Replacement	chrX	7023678	HDHD1	NM_001178135	-	G	A	c.263G>A	T(2ACG)	88	M
SNP	Replacement	chrX	7023678	HDHD1	NM_001135565	-	G	A	c.332G>A	T(2ACG)	111	M
SNP	Replacement	chrX	7023678	HDHD1	NM_012080	-	G	A	c.263G>A	T(2ACG)	88	M
SNP	Replacement	chrX	7268296	STS	NM_000351	+	C	G	c.1746C>G	S(3AGC)	582	R

Amino_Acid_ Change	dbSNP_Ids	Het_Rates	Orientations	PhastCons_ placental	PhastCons_ primates	PhastCons_ vertebrate	PhyloP_ placental	PhyloP_ primates	PhyloP_ vertebrate	Num_ Hom_ SNPs	Sample_Ids	Num_ Het_ SNPs	Sample_Ids
G2000S	rs1635242,	0.407,	+,	0.673	0.217	0.929	1.331	0.512	1.228	44	SampleA	0	---
P1665S	rs1974522,	0.492,	+,	0.004	0.039	0.047	0.307	0.409	1.74	41	Sample B	0	---
A1128V	rs66465405,rs1635246,	0.000,0.467,	+,+,	0	0.016	0	-1.433	0.512	-0.512	42	Sample D	0	---
I824V	rs5983119,	0.322,	+,	0	0	0	0	-0.205	-0.409	45	Sample A	0	---
V764L	rs5983120,	0.060,	+,	0.012	0.016	0.004	1.433	0.512	0.921	1	Sample C	0	---
D222N	---	---	---	0	0.008	1	0.409	0.512	3.071	1	Sample E	0	---
V43A	rs3752362,	0.000,	+,	0	0.016	0.055	-0.307	-0.614	-0.717	28	Sample F	0	---
T88M	rs1131197,	0.367,	-,	0.917	0.63	0.996	0.512	-0.102	2.047	18	Sample G	0	---
T111M	rs1131197,	0.367,	-,	0.917	0.63	0.996	0.512	-0.102	2.047	18	Sample F	0	---
T88M	rs1131197,	0.367,	-,	0.917	0.63	0.996	0.512	-0.102	2.047	18	Sample L	0	---
S582R	---	---	---	0.004	0.008	0	0	-0.205	0.102	1	Sample M	0	---

Figure 6. Snapshot of Exonic Replacement Annotation Output File. The top half shows the data for fields 1 - 13. The bottom half of the figure shows the data from fields 14 - 23. The last four columns report the number of homozygous and heterozygous SNPs and associated sample IDs.

3. An application of SeqAnt 2.0: Targeted next-generation sequencing of *NLGN3* and *NLGN4X* in humans

The targeted sequencing of specific genes or genomic regions is a common experimental design that can benefit from the use of SeqAnt. Here we describe such a study. We sequenced the *NLGN*3 and *NLGN4X* loci in a sample of 144 males with a diagnosis of autism. All the patient samples were obtained from the multiplex Autism Genetic Resource Exchange (AGRE) [22]. Raw base-calling data generated with an Illumina Genome Analyzer (IGA) were used as input for mapping and alignment. The total amount of sequence generated was 7.04 GB. Paired-end reads were mapped and variants were called using PEMapper (Cutler DJ et al, personal communication). In total, 99.7% of target bases had at least 8X coverage, with a median depth of coverage of 452. We identified a total of 208 sites of variation, with 176 single nucleotide polymorphisms and 32 insertions or deletions. Overall levels of variation were estimated at 5.8 x 10^{-4} (Θ_W per site [23]), which matched our expectation for loci from the human X chromosome. We also observed an excess of rare variants, as evidenced by a negative value for the Tajima's D test statistic (-0.27,[24]).

Single nucleotide variants (SNVs) and small insertions and deletions (INDELs) were annotated using SeqAnt [5]. For the SNPs, a total of 68, or 39%, had not been reported before (31 in *NLGN3* and 37 in *NLGN4X,* Table 2). For the INDELs, a total of 24, or 75%, had not been reported before (5 in *NLGN3* and 19 in *NLGN4X*, Table 3). As summarized in Figure 7, almost all common variation (>5% frequency in our sample) is contained in dbSNP, whereas most rare variants (<5%) have not been cataloged there.

Functional class	Total SNPs	SNPs in dbSNP	Novel SNPs	Novel SNPs at Evolutionary Conserved Sites
Replacement	1	1	0	0
Silent	3	3	0	0
UTR	18	10	8	2
Intron	134	78	56	9
Intergenic	20	16	4	0
Total	176	108	68	11

Table 2. Functional annotation of SNPs at the *NLGN3* and *NLGN4X* loci identified by next-generation sequencing of 144 males with autism.

Using SeqAnt to rapidly annotate our sequence data allows us to quickly draw four main conclusions. First, most common variation is already contained in dbSNP, while much of the rare variation remains undiscovered. Second, we did not see any novel replacement variants at either *NLGN3* or *NLGN4X*, suggesting that mutations at these loci are rare causes of autism. Third, we identified novel UTR variants at highly evolutionarily conserved sites, which could contribute to autism susceptibility. We focused on this set of variants for direct functional testing. Finally, we identified novel intronic variants at evolutionarily conserved sites that appear to be located in transcription factor binding sites. These variants are being followed up to determine whether they have a regulatory role that impacts the expression of *NLGN3* or *NLGN4X*. In summary, SeqAnt 2.0 allowed us to rapidly annotate all the sites of variation in our sample and rapidly focus attention on those variants most likely to be autism susceptibility alleles.

Functional Class	Total Indels	Indels in dbSNP	Novel Indels	Novel Indels at Evolutionary Conserved Sites
Coding	0	0	0	0
UTR	1	0	1	1
Intron	25	7	18	0
Intergenic	6	1	5	0
Total	32	8	24	1

Table 3. Functional annotation of INDELs at the *NLGN3* and *NLGN4X* loci identified by next-generation sequencing of 144 males with autism.

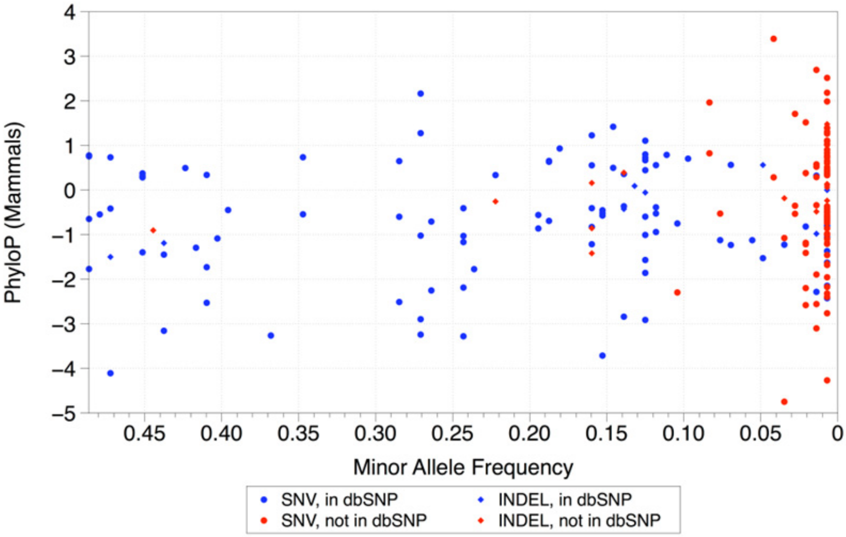


Figure 7. Summary of SNV and indel variation discovered at the *NLGN3* and *NLGN4X* loci in males with ASD. The frequency of SNVs and INDELs (minor alleles) in cases is plotted against their level of evolutionary conservation. Most common variation has already been discovered and exists in public databases (blue; circles and diamonds); most of the rare variation at both loci was discovered in our study and not contained in public databases (red; circles and diamonds).

4. An application of SeqAnt 2.0: Sequencing the *AFF2* locus and X chromosome exome in patients with autism

With improvements in methods of targeted enrichment and next-generation sequencing, the targeted sequencing of all genes on a specific chromosome has become feasible. Specific genes/genomic regions is a common experimental design that benefits from the use of SeqAnt [25]. Here we performed an experiment that combined targeted sequencing with chromosomal exome sequencing. We selected 127 males from the Autism Genetic Resource Exchange (AGRE) multiplex collection and 75 males from the Simons Foundation Autism Research Initiative (SFARI) Simplex Collection, New York, NY, USA (SSC) for target DNA amplification and DNA sequencing. From the AGRE collection, we chose multiplex families with two or more male affected sib-pairs who shared >99% of 76 genotyped SNPs in the *AFF2* genomic region [22]. One male was randomly chosen if both affected siblings were equally affected; otherwise, the male with autism was chosen over those boys with a diagnosis of not quite autism (NQA) or broad spectrum. From the SSC collection, we chose only those boys who were described as autistic and not reported to have any other syndromes. From the SSC collection, we chose 75 male children from different families with a diagnosis of ASD [26].

For the AGRE samples, we prepared target DNA for sequencing the AGRE samples by performing long PCR (LPCR) amplification of the *AFF2* genomic region, followed by sequencing on an Illumina Genome Analyzer. For the SSC samples, we prepared target DNA for Illumina sequencing by using RainDance Technology’s (RDT) microdroplet-based technology to enrich for the human X chromosome exome, as described previously [25]. Following enrichment we performed 70-bp single-end multiplex sequencing on an Illumina Genome Analyzer (IGA). Nearly 20 GB of sequence was generated for AGRE samples, while ~55 GB of sequence was generated for the SSC samples. The *AFF2* reference sequence used for the AGRE samples consists of 10 discontiguous fragments covering 84.8 kb, and the SSC reference sequence consisted of the entire human X chromosome, which spanned 5748 discontiguous fragments covering 4.7 Mb. Raw base-calling data generated with the IGA were mapped and variants called using PEMapper (Cutler DJ et al, personal communication). For AGRE samples, 99% of the bases had more than 8X coverage. Median depth of coverage was in the range of 388-1548. For the SSC samples, between 83% and 97% of the targeted reference bases had more than 8X coverage. Median depth of coverage was in the range of 20-607. We identified a total of 286 sites of variation, with 269 single nucleotide polymorphisms (SNPs) and 17 insertions or deletions (INDELs). Overall levels of variation were similar between the two datasets (Θ_W per site [23]; AGRE - 6.0×10^{-4}, SSC - 6.7×10^{-4}), with an excess of rare variants as evidenced by a negative value for the Tajima’s D test statistics for both sets of samples ([24]; AGRE: -1.46, SSC: -1.41).

We used SeqAnt to annotate the variants found at the *AFF2* locus in the total sample of 202 males with a diagnosis of autism (Mondal et al, in revision). We sought to test the hypothesis that rare variants at the *AFF2* locus can act as autism susceptibility alleles. Annotating our variants using the other web-based tools, like the UCSC Genome Browser or the Ensembl Genome Browser, would have been time-consuming and laborious. SeqAnt helped us rapidly annotate these SNPs and INDELs into different functional classes, as well as reported whether a variant had already been cataloged in the dbSNP database (Tables 4, 5). SeqAnt also reported the PhastCons and phyloP conservation scores, which are important in helping to determine whether a variant might cause a deleterious change in the protein structure/function, since variants in the well-conserved sites are likely to cause such changes. By using this feature of SeqAnt, we could easily identify our list of candidate variants that were rare, as well as likely to cause a damaging change.

Functional Class	Total SNPs	SNPs in dbSNP	Novel SNPs	Novel SNPs at Conserved Sites
Replacement	5	0	5	5
Silent	8	4	4	4
UTR	33	20	13	1
Intron	223	129	94	6
Total	269	153	116	16

Table 4. Functional annotation of single nucleotide polymorphisms at the *AFF2* locus identified by next-generation sequencing of 202 males with autism

Functional Class	Total Indels	Indels in dbSNP	Novel Indels	Novel Indels at Conserved Sites
Exonic	0	0	0	0
UTR	2	0	2	1
Intron	15	7	8	1
Total	17	7	10	2

Table 5. Functional annotation of indels at the *AFF2* locus identified by next-generation sequencing of 202 males with autism

As expected, almost all common variation (>5% frequency in our population) is contained in dbSNP, whereas most rare variants (<5%) are not cataloged in dbSNP (Figure 8). We found that, in our cases, there were five (2.5% of total cases sequenced) singleton nonsynonymous variants. This level of variation in our cases was significantly higher than that seen in a set of 5400 controls. Furthermore, we used SeqAnt to rapidly annotate 1006 X chromosome genes that had been sequenced in the 75 SSC samples, and ultimately showed that the excess mutations at *AFF2* were unusual compared to other X chromosome loci. Thus, the ability to rapidly annotate our sequence variants discovered from sequencing the entire X chromosome exome had a major impact on our ability to assess the role of *AFF2* as an autism susceptibility locus. Finally, SeqAnt helped us identify three rare noncoding UTR sequence variants, one of which was at an evolutionarily conserved site. Subsequent functional testing suggested that the variant at the conserved site acts to influence the level of *AFF2* expression. Thus, for this experiment, SeqAnt allowed us to rapidly focus on those sites of greatest interest for both statistical analyses and direct functional testing.

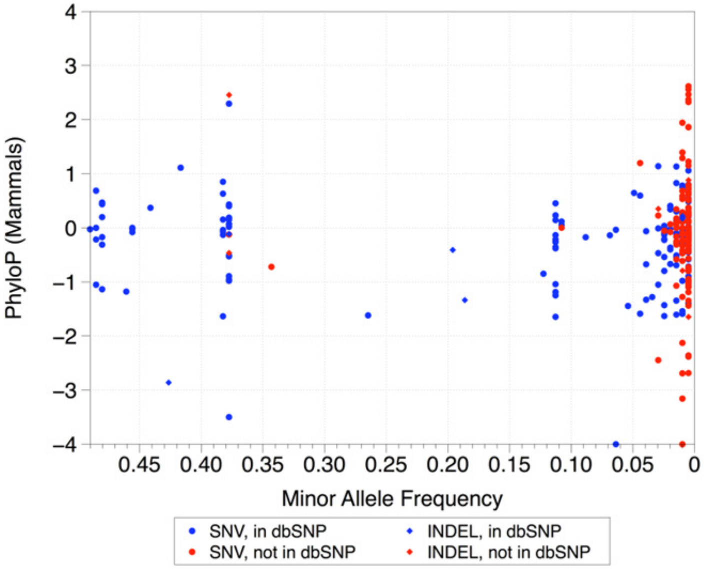


Figure 8. **Summary of SNV and indel variation discovered at the *AFF2* locus in males with ASD.** The frequency of SNVs and indels (minor alleles) in cases is plotted against their level of evolutionary conservation. Most common variation has already been discovered and exists in public databases (blue; circles and diamonds). Most of the rare variation at *AFF2* was discovered in our study and not contained in public databases (red; circles and diamonds).

5. An application of SeqAnt 2.0: Discovering new mutations from forward genetic screens in the mouse

Forward genetic screens in *Mus musculus* have been very informative, revealing unsuspected mechanisms governing basic biological processes [27-32]. In this approach, a potent chemical mutagens, such as *N*-ethyl-*N*-nitrosourea (ENU), is used to randomly induce mutations in mice. The mice are then bred and phenotypically screened to identify lines that disrupt a specific biological process of interest. Although identifying a mutation using the rich resources of mouse genetics is straightforward, it is unfortunately neither fast nor cheap.

To solve this problem, we developed a methodology that combines multiplex chromosome-specific exome capture, next-generation sequencing, rapid mapping, sequence annotation, and variation filtering to detect newly induced causal variants in a dramatically accelerated way [33]. Rapid sequence annotation and variation filtering are critical to this approach. We used SeqAnt as a part of this methodology for rapid annotation of variations obtained from mutant, parental, and background strains in a single experiment. By using SeqAnt, we first annotated all the variants into different functional classes. Next, by comparing variants identified in mutant offspring to those found in dbSNP, the unmutagenized background strains, and parental lines, we could immediately distinguish the induced putative causative mutations from preexisting variations or experimental artifacts (Table 6).

Mutant Line	Functional Classes	Total Homozygous Variants	In dbSNP	In Background Strains, Not in dbSNP	Remaining Variants	Replacement Variants Within Mapped Region
AB5	Replacement	96	80	13	3	1
AB5	Silent	157	143	12	2	-
AB5	UTR	331	191	135	5	-
AB5	Intronic	106	87	17	2	-
AB5	Intergenic	54	50	4	0	-
M2	Replacement	43	8	31	4	2
M2	Silent	19	11	7	1	-
M2	UTR	73	16	55	2	-
M2	Intronic	46	18	20	8	-
M2	Intergenic	40	4	36	0	-
X5	Replacement	128	59	63	6	2
X5	Silent	192	128	63	1	-
X5	UTR	387	231	152	4	-
X5	Intronic	205	116	86	3	-
X5	Intergenic	89	34	55	0	-
Y1	Replacement	17	1	14	2	1
Y1	Silent	5	0	4	1	-
Y1	UTR	14	2	11	1	-
Y1	Intronic	34	0	31	3	-
Y1	Intergenic	7	0	7	0	-

Table 6. Results of filtering homozygous variants sites for each mouse mutant line sequenced.

We demonstrated the use of this approach to find the causative mutations induced in four novel ENU lines identified from a recent ENU screen. In all four cases, after applying our method and combining with standard mapping data used to initially localize the variant to a chromosome, we found two or fewer putative mutations (and sometimes only a single one). Confirming that the variant was in fact causative was then easily achieved via standard segregation approaches. SeqAnt gave us the ability to rapidly annotate and screen variants of lesser interest (silent, UTR, intronic, intergenic), so we could instead focus our attention on those variants (replacement) that were most likely to account for the mutant phenotype.

6. An application of SeqAnt 2.0: Exome sequencing to discover mutations affecting neutrophil function in very-early-onset pediatric Crohn's disease

Children with very-early-onset (VEO) pediatric Crohn's disease (CD) are found to have high levels of neutrophil dysfunction. Neutrophils are an abundant type of white blood cell that play an essential role in innate immunity. We therefore hypothesized that children with very-early-onset Crohn's disease would exhibit an increased frequency of genetic mutations affecting neutrophil function. For an initial study we selected 45 VEO CD patients (median (range) age: 8.5 (5-10) years) with CBir1 sero-reactivity and moderate-to-severe clinical disease activity at diagnosis. We used the Roche NimbleGen SeqCap EZ Human Exome Library v2.0 on genomic DNA extracted from whole blood to capture the whole exome for each patient. Barcodes were used to prepare the libraries for whole-exome capture, which allowed us to sequence two whole exomes per lane of next-generation sequencing. We performed multiplexed 100 base-pair paired-end sequencing on an Illumina HiSeq 2000 instrument. We used PEMapper (Cutler and Zwick, in revision) to map raw sequence reads and identify variants sites relative to the ~30.8 Mb human exome reference sequence (NCBI37/hg19).

We then used SeqAnt to annotate all variant sites for functional significance, frequency, presence in databases like dbSNP, and measures of evolutionary conservation. Our central hypothesis was that early-onset (pediatric) forms of IBD would be substantially influenced by deleterious mutations found in the neutrophil pathway. If true, a straightforward evolutionary model of mutation-selection balance predicts that these variants ought to be rare in the general population, found at highly evolutionarily conserved sites, and have large effects on gene function. Thus, variants found in coding regions (replacement, nonsense, exonic insertions/deletions) that putatively alter protein structure and function will be the strongest candidates as contributors to IBD in pediatric patients. A number of lines of evidence specifically implicate loci involved in neutrophil functional pathways. We therefore proposed a strategy of first discovering variation in genes known to function in the neutrophil pathway, followed by direct functional testing of alleles from specific patients.

Gene	Location	Variants	Type	Position	Function	Frequency in VEO CD Patients	Frequency in Control Population
CSF2RA	chrX (p22.33)	0	-	-	GM-CSF signaling	-	-
CSF2RB	chr22 (q12.3)	1	SNP	37331455	GM-CSF signaling	0.02	0.0024
CYBB	chrX (p11.4)	1	SNP	37663322	oxidative burst	0.02	0.0032
DUOX1	chr15 (q21.1)	2	SNP SNP	45448069 45431655	enterocyte, H202	0.02 0.02	0.0003 0.0002
DUOX2	chr15 (q21.1)	1	Indel	45393428-30	enterocyte, H2O2	0.02	-
FCGR1A	chr1 (q21.2)	0	-	-	phagocytosis	-	-
FCGR2A	chr1(q23.3)	0	-	-	phagocytosis	-	-
FCGR2B	chr1 (q23.3)	0	-	-	phagocytosis	-	-
FCGR3A	chr1 (q23.3)	0	-	-	phagocytosis	-	-
FCGR3B	chr1 (q23.3)	0	-	-	phagocytosis	-	-
IL27RA	chr19 (p13.12)	1	Indel	14159807	IL-27 signaling	0.02	-
JAK2	chr9 (p24.1)	0	-	-	GM-CSF signaling	-	-
MPO	chr17 (q22)	0	-	-	bacterial killing	-	-
NCF1	chr7 (q11.23)	0	-	-	oxidative burst	-	-
NCF2	chr1 (q25.3)	0	-	-	oxidative burst	-	-
NCF4	chr22 (q12.3)	1	SNP	37273825	oxidative burst	0.02	0.0001
NLRP12	chr19 (q13.42)	0	-	-	chemotaxis	-	-
NOS2	chr17 (q11.2-q12)	3	Indel Indel Indel	26087106 26096042 26085975-76	reactive nitrogen intermediates	0.2 0.2 0.57	- - -
NOX1	chrX (q21.1)	0	-	-	oxidative burst	-	-
NOX3	chr6 (q25.3)	0	-	-	oxidative burst	-	-
NOX4	chr11 (q14.3)	2	SNP	89088208	oxidative burst	0.02	-

Gene	Location	Variants	Type	Position	Function	Frequency in VEO CD Patients	Frequency in Control Population
			SNP	89182666		0.02	0.0022
NOX5	chr15 (q23)	0	-	-	oxidative burst	-	-
PRAM1	chr19 (p13.2)	1	Indel	8564497-500	adhesion	0.02	-
RAC1	chr7 (p22.1)	0	-	-	oxidative burst	-	-
RAC2	chr22 (q12.3)	0	-	-	oxidative burst	-	-
SELPLG	chr12 (q24.11)	1	SNP	109017468	adhesion	0.11	-
SLC11A1	chr2 (q35)	2	Indel SNP	219247739 219254723	bacterial killing	0.02 0.02	- -
STAT3	chr17 (q21.2)	2	SNP SNP	40481429 40477064	IL-27 signaling	0.02 0.02	- 0.0002
STAT5A	chr17 (q21.2)	1	SNP	40461109	GM-CSF signaling	0.02	-
STAT5B	chr17 (q21.2)	0	-	-	GM-CSF signaling	-	-
VAV1	chr19 (p13.3)	0	-	-	oxidative burst	-	-
VAV2	chr9 (q34.2)	0	-	-	oxidative burst	-	-
VAV3	chr1 (p13.3)	0	-	-	oxidative burst	-	-

Table 7. Genetic variants found in genes that regulate neutrophil function.

We used SeqAnt to annotate all the sequence variations from the 45 exomes and identified a total of 60,682 variant sites of interest in coding regions (54,313 replacement SNPs, 2953 indels covering 6369 bases). For our exploratory genome-wide analysis of SNPs, we restricted our analysis to those variants with phyloP scores greater than 2.0, which corresponds to the top 1% of conserved sites in the human genome. Remaining were 12,575, of which 51% (6490) were not cataloged in dbSNP 132 and might constitute novel mutations contributing to early-onset IBD. We then restricted our analysis to 33 neutrophil genes. Table 6 contains a list of these 33 neutrophil genes with the number of rare putative functional variants (replacement SNPs or exonic indels). These variants are to be followed up using direct functional assays to assess function. Again, SeqAnt enabled us to rapidly annotate all variants, ignore those variants of lesser interest, and focus our attention on those most likely to contribute to the VEO CD in our sequenced patients.

7. Future directions

We have shown many useful features of SeqAnt and how it can be applied in a variety of experiments, yet we continue to develop SeqANt and plan to expand its functionalities going forward. Our goal is to create a one-stop online tool that readily accepts raw sequencing data and generates output through the annotation and functional characterization stages. Moreover, because our software and libraries are open source, they can be downloaded and optimized locally as part of a next-generation sequencing pipeline. SeqAnt is a truly dynamic application that is updated regularly to keep up with the constant flow of new sequencing data, genome assemblies, and improved annotation information available from public databases like those found at the UCSC Genome Browser.

Genomic sequence annotation requires an up-to-date and comprehensive database of DNA sequence information for a given organism. Our first aim is to continue adding to our database organisms whose genomic information could be annotated. We plan on including several other mammals, vertebrates, invertebrates, and ultimately bacteria strains in the near future. This will give researchers a web application they can use to speed their genetic studies of such organisms. We are also in the process of updating the dbSNP information contained in the SeqAnt database.

Another area of future focus is to broaden the types of input and output files that SeqAnt could work with, while embracing standards in broad use in the bioinformatics community. We intend to include the capability to directly annotate .vcf files as a standard input file format. Presently, all our output files are either text files or BED files. We also plan to provide the option of having the annotation output in .vcf format. Furthermore, we intend to modify SeqAnt to make the .map and .ped files (PLINK formats) from the snp variant file, which will be beneficial for substructure analysis and several other analyses that can be done using PLINK.

The inclusion of additional custom tracks from the UCSC browser to annotate for conserved and putatively functional sites will also be a future area of SeqAnt development. Our hope is that this will improve the effectiveness of downstream functional analysis. We also plan to have the application hosted in a cloud computing environment, side by side with other bioinformatics tools. This is relevant not only because of the wider accessibility it guarantees, but there is often the added ease of using other tools in the same environment to generate and modify input and output files from SeqAnt for further analysis.

SeqAnt was set up to be a dynamic application, and our improvements to this software make it possible to apply SeqAnt to different genomic variant analysis situations. Inevitable advances in sequencing technologies will spur continued demand for tools that can make sense out of the enormous raw sequence data generated, and we will work continually to make SeqAnt adaptable to these improvements and even more accessible to the wider public.

8. Conclusion

Great advances in targeted enrichment methods and DNA sequencing are beginning to allow individual investigators to sequence significant portions of many genomes; the bottleneck this has revealed lies with the annotation and interpretation of the resulting genomic variation data. SeqAnt is a software tool that directly addresses this bottleneck in a wide variety of potential applications. SeqAnt is an open source application that contains a number of unique features. The first is its ability to annotate data from many organisms, not just humans. Second, it is able to perform this analysis with a minimal memory footprint. Third, it completes this analysis in record time, thereby removing a significant bottleneck facing a researcher using the latest next-generation sequencing platforms.

The modifications we made to the application ensure we have the latest data tracks for the species we currently have in the SeqAnt binary databases. Furthermore, we have expanded the number of species that can now be annotated. Finally, with the addition of the PhyloP46Way conservation track, researchers can more confidently assess the evolution and significance of a particular variant site when the phyloP scores are viewed side by side with the PhastCons score values.

We have applied SeqAnt to various studies in our lab, from the work analysis of data on targeted sequencing of particular genes to the analysis of whole-exome data. We also used SeqAnt in the variant annotation of mouse genome and the adaptation of HapMap data for analyzing human exomes. The results from these various applications establish SeqAnt as a user-friendly tool that could help researchers in their work over a wide range of endeavors.

SeqAnt will continue to be an open source web application, which we will constantly update to meet the demands of changing and improving genomic and sequencing technologies. The future of genomics and variation studies lies in our ability to properly use the massive amounts of information we have obtained from DNA sequencing. Sequence annotation tools like SeqAnt that can efficiently turn such data into useable information will play a key role in this future.

Author details

Matthew Ezewudo, Promita Bose, Kajari Mondal, Viren Patel, Dhanya Ramachandran,
and Michael E. Zwick[*]
Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, 30322, USA

Acknowledgement

This work was supported by the National Institutes of Health/National Institutes of Mental Health (NIH/NIMH) and Gift Fund (grant number: MH076439, MEZ); the Simons Foundation Autism Research Initiative (MEZ); and the Training Program in Human Disease Genetics (grant number: 1T32MH087977, DR). We thank members of the Cutler and Zwick labs and Jennifer G. Mulle for discussion, Cheryl T. Strauss for editing, and the Emory-Georgia Research Alliance Genome Center (EGC), supported in part by PHS Grant UL1 RR025008 from the Clinical and Translational Science Award program, National Institutes of

[*] Corresponding Author

Health, National Center for Research Resources, for performing the Illumina sequencing discussed in this chapter. The ELLIPSE Emory High Performance Computing Cluster was used for the development of SeqAnt.

9. References

[1] Lander ES. 1996. The new genomics: global views of biology. *Science (New York, NY)* 274: 536-539.

[2] Chakravarti A. 2011. Genomic contributions to Mendelian disease. *Genome Res* 21: 643-644.

[3] Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, Coop G, Przeworski M. 2009. Broad-scale recombination patterns underlying proper disjunction in humans. *PLoS Genet* 5: e1000658.

[4] Bhangale TR, Rieder MJ, Nickerson DA. 2008. Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* 40: 841-843.

[5] Shetty AC, Athri P, Mondal K, Horner VL, Steinberg KM, Patel V, Caspary T, Cutler DJ, Zwick ME. 2010. SeqAnt: a web service to rapidly identify and annotate DNA sequence variations. *BMC Bioinformatics* 11: 471.

[6] Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, James Kent W. 2012. The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 40: D918-23.

[7] Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GR, Ruffier M, Schuster M, Sobral D, Spudich G, Tang YA, Trevanion S, Vandrovcova J, Vilella AJ, White S, Wilder SP, Zadissa A, Zamora J, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Parker A, Proctor G, Vogel J, Searle SM. 2011. Ensembl 2011. *Nucleic Acids Res* 39: D800-6. PMC3013672.

[8] Chelala C, Khan A, Lemoine NR. 2009. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* 25: 655-661.

[9] Renaud G, Neves P, Folador EL, Ferreira CG, Passetti F. 2011. Segtor: rapid annotation of genomic coordinates and single nucleotide variations using segment trees. *PLoS ONE* 6: e26715.

[10] Makarov V, O&apos,Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. 2012. AnnTools: A Comprehensive and Versatile Annotation Toolkit for Genomic Variants. *Bioinformatics*

[11] Garla V, Kong Y, Szpakowski S, Krauthammer M. MU2A – Reconciling the genome and transcriptome to determine the effects of base substitutions.

[12] Jiang J, Jiang L, Zhou B, Fu W, Liu J-F, Zhang Q. 2011. Snat: a SNP annotation tool for bovine by integrating various sources of genomic information. *BMC genetics* 12: 85.

[13] Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Altshuler D, Group ISNPMW. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409: 928-933.

[14] Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ. 2004. Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics* 20: 1022-1032.

[15] Consortium IH, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MMY, Tsui SKW, Xue H, Wong JT-F, Galver LM, Fan J-B, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier J-F, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok P-Y, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui L-C, Mak W, Song YQ, Tam PKH, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PIW, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe&apos,er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CDM, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L&apos,Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li

Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851-861.

[16] Consortium IH. 2005. A haplotype map of the human genome. *Nature* 437: 1299-1320.

[17] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29: 308-311.

[18] Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20: 110-121.

[19] Briggs JP. 2002. The zebrafish: a new model organism for integrative physiology. *Am J Physiol Regul Integr Comp Physiol* 282: R3-9.

[20] Norton W, Bally-Cuif L. 2010. Adult zebrafish as a model organism for behavioural genetics. *BMC Neurosci* 11: 90. PMC2919542.

[21] Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*

[22] Geschwind DH, Sowinski J, Lord C, Iversen P, Shestack J, Jones P, Ducat L, Spence SJ, Committee AGRES. 2001. The autism genetic resource exchange: a resource for the study of autism and related neuropsychiatric conditions. *Am J Hum Genet* 69: 463-466.

[23] Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* 7: 256-276.

[24] Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585-595.

[25] Mondal K, Shetty AC, Patel V, Cutler DJ, Zwick ME. 2011. Targeted sequencing of the human X chromosome exome. *Genomics* 98: 260-265.

[26] Fischbach GD, Lord C. 2010. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron* 68: 192-195.

[27] Caspary T, Anderson KV. 2006. Uncovering the uncharacterized and unexpected: unbiased phenotype-driven screens in the mouse. *Dev Dyn* 235: 2412-2423.

[28] Cook MC, Vinuesa CG, Goodnow CC. 2006. ENU-mutagenesis: insight into immune function and pathology. *Curr Opin Immunol* 18: 627-633.

[29] Acevedo-Arozena A, Wells S, Potter P, Kelly M, Cox RD, Brown SD. 2008. ENU mutagenesis, a way forward to understand gene function. *Annu Rev Genomics Hum Genet* 9: 49-69.

[30] Beutler B, Moresco EM. 2008. The forward genetic dissection of afferent innate immunity. *Curr Top Microbiol Immunol* 321: 3-26.

[31] Caspary T. 2010. Phenotype-driven mouse ENU mutagenesis screens. *Methods Enzymol* 477: 313-327.

[32] Stottmann RW, Moran JL, Turbe-Doan A, Driver E, Kelley M, Beier DR. 2011. Focusing forward genetics: a tripartite ENU screen for neurodevelopmental mutations in the mouse. *Genetics* 188: 615-624. PMC3176541.

[33] Sun M, Mondal K, Patel V, Horner VL, Long AB, Cutler DJ, Caspary T, Zwick ME. 2012. Multiplex Chromosomal Exome Sequencing Accelerates Identification of ENU-Induced Mutations in the Mouse. *G3 (Bethesda, Md)* 2: 143-150.