Published in final edited form as:
J Med Genet. 2012 July ; 49(7): 433–436. doi:10.1136/jmedgenet-2012-100918.

wANNOVAR: annotating genetic variants for personal genomes via the web

Xiao Chang[1] and **Kai Wang**[1,2]

[1]Zilkha Neurogenetic Institute, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

[2]Department of Psychiatry and Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA

Abstract

Background—High-throughput DNA sequencing platforms have become widely available. As a result, personal genomes are increasingly being sequenced in research and clinical settings. However, the resulting massive amounts of variants data pose significant challenges to the average biologists and clinicians without bioinformatics skills.

Methods and results—We developed a web server called wANNOVAR to address the critical needs for functional annotation of genetic variants from personal genomes. The server provides simple and intuitive interface to help users determine the functional significance of variants. These include annotating single nucleotide variants and insertions/deletions for their effects on genes, reporting their conservation levels (such as PhyloP and GERP++ scores), calculating their predicted functional importance scores (such as SIFT and PolyPhen scores), retrieving allele frequencies in public databases (such as the 1000 Genomes Project and NHLBI-ESP 5400 exomes), and implementing a 'variants reduction' protocol to identify a subset of potentially deleterious variants/genes. We illustrated how wANNOVAR can help draw biological insights from sequencing data, by analysing genetic variants generated on two Mendelian diseases.

Conclusions—We conclude that wANNOVAR will help biologists and clinicians take advantage of the personal genome information to expedite scientific discoveries. The wANNOVAR server is available at http://wannovar.usc.edu, and will be continuously updated to reflect the latest annotation information.

INTRODUCTION

Over the past 5 years, massively parallel DNA sequencing platforms have become widely available.[1] As a result, variants data on genomes from healthy subjects and patients are being generated at an unprecedented rate. However, the development of bioinformatics tools for handling these data lags behind, creating a gap between the generation of massive data and the ability to fully exploit the biological contents of these data. To fill the urgent

Correspondence to, Dr Kai Wang, University of Southern California, 1501 San Pablo St, Los Angeles, CA 90089, USA; kaiwang@usc.edu.

Contributors XC developed and implemented the tool and drafted the manuscript. KW supervised the implementation of the project and revised the manuscript. All the authors read and approved the manuscript.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The two data sets used in the manuscript are available to users in the "Example" section of the wANNOVAR server.

demand, we previously developed the ANNOVAR (ANNOtate VARiation) software for functional annotation of genetic variants from sequence data.[2] ANNOVAR efficiently uses up-to-date information to annotate genetic variants detected from diverse genomes with user-specified versions of genome builds. Although ANNOVAR has become one of the most widely used annotation tools for sequencing data, the requirement to type command line arguments makes ANNOVAR inaccessible to the average biologists and clinicians who would otherwise benefit from its extensive functionality.

Therefore, we developed a web server called wANNOVAR to facilitate web-based personal genome annotation, using ANNOVAR as the backend annotation engine. Users need to simply submit a list of variants (even whole-exome or whole-genome variants), and wANNOVAR can process the submission and generate HTML-based result pages. It allows flexibility by permitting the users to select customised filtering criteria and identify a subset of prioritised variants from thousands or even millions of input variants. Below, we describe the implementation of the wANNOVAR sever and illustrate its utility using two high-throughput sequencing data sets on Mendelian diseases.

METHODS

The web server is composed of a web interface and a background program for executing annotation tasks. Our tests indicated that the server performed well under a light load for user queries. For example, annotating an exome with ~20 000 SNPs and indels takes merely a few minutes in the server. The subroutines for handling user query were written in Perl and were facilitated by the Common Gateway Interface module (CGI.pm). The static and dynamic HTML pages have been tested in different versions of Internet Explorer, Firefox and Google Chrome browsers.

Input fields for the wANNOVAR server include a sample identifier, an email address, a variant file, the reference genome build, the gene definition system and optionally a disease model for running the 'variants reduction' pipeline. The default input format for the variant file is variant call format (VCF),[3] which is a text file that contains meta-information lines, a header line, and data lines containing information about a position in the genome. The server can also handle other input formats, including the ANNOVAR input format, the Complete Genomics ASM.tsv format and the GFF3-SOLiD format. Currently, the input file size is restricted to less than 200 MB, and the input file can be compressed in .gz or .zip format. If all input fields are correctly set, the server will return a webpage with a URL for the results page.

The results page contains a collection of functional annotations for variant calls. Users can download the 'exome summary results' or the 'genome summary results' as Excel-compatible files or tab-delimited files, or choose to view the annotation results in a table on the webpage. The annotations on all variants were grouped into several broad categories including gene annotation, variation databases, functional prediction and region annotations (table 1). Several functional prediction scores for exonic variants from the dbNSFP Database[4] including SIFT,[5] PolyPhen,[6] LRT,[7] MutationTaster[8] and PhyloP,[9] are also provided in the wANNOVAR server to help users judge the functionality of variants using multiple sources of information. As previously described, wANNOVAR can perform a 'variants reduction' procedure to identify a subset of the most likely causal variants/genes for Mendelian diseases, from a large list of variants on personal genomes.[2] For example, users can remove variants observed in public databases such as the 1000 Genomes Project,[10] NHLBI-ESP 5400 exomes[11] and dbSNP[12] with specific minor allele frequency cut-off. The server uses modified versions of dbSNP that excluded all SNPs flagged as 'clinically associated' by dbSNP. We provide several default pipelines for different disease models such as ‘rare recessive Mendelian disease’ and ‘rare dominant Mendelian disease’, but users can also use ‘advanced options’ to specify a custom filtering strategy (table 2).

RESULTS

Analysis of a real exome sequencing data set on Ogden syndrome

To demonstrate the utility of the wANNOVAR server, we analysed variants calls from a family segregating Ogden syndrome ([MIM: 300855]). Thirty years ago, Ogden syndrome was discovered as an X linked lethal infantile disorder, and its genetic basis was recently solved by next-generation sequencing.[13] The disease is characterised by postnatal growth failure with severe delays and dysmorphic features, and is caused by a mutation in the *NAA10* gene, leading to a N-terminal acetyltransferase deficiency. For the family with Ogden syndrome, exon-capture sequencing data was aligned by BWA[14] and genotypes were called by GATK[15] as VCF[3] files in hg19 coordinate. We submitted all chromosome X variants (1318 single nucleotide variants and 161 indels) in the proband to the wANNOVAR server, and tested the ‘variants reduction’ procedure using the default ‘rare recessive Mendelian disease’ pipeline and a custom pipeline (table 2). Compared with the default pipeline, the custom pipeline filter variants set against the two unaffected family members and the deleterious variants were identified using SIFT/PolyPhen scores. Both pipelines identified a hemizygous mutation (p.S37P) within a single candidate gene *NAA10*, and this was precisely the known causal variant in this family.[13] Detailed examination of the ‘exome summary’ table demonstrated that this variant has a SIFT[5] score of 0 (prediction: damaging), PolyPhen[6] score of 0.96 (prediction: probably damaging), LRT[7] score of 1 (prediction: deleterious), Mutation Taster[8] score of 1 (prediction: disease causing), PhyloP[9] score of 0.96 (prediction: conserved) and GERP++[16] score of 3.55 (prediction: highly constrained). The variant is not observed in the 1000 Genomes Project,[10] the dbSNP[12] version 135 (after removing SNPs flagged as ‘clinically associated’) or the NHLBI-ESP 5400 exomes.[11] Therefore, converging bioinformatics evidence supports that this variant may affect protein function.

Analysis of a synthetic whole-genome sequencing data set on Miller syndrome

We next evaluated wANNOVAR on millions of genetic variants from whole-genome sequencing. We used a synthetic data set of a male subject with ~4.2 million single nucleotide variants and ~0.5 million indels,[17] supplemented with two variants (p.G152R and p.G202A) in *DHODH* known to cause Miller syndrome ([MIM: 263750]).[18] This synthetic data set was previously used to illustrate the ‘variant reduction’ procedure.[2] With the default ‘rare recessive Mendelian disease’ pipeline (table 2), the large number of input variants was drastically reduced to 516, and 24 candidate genes were identified including the causal gene *DHODH*. We also tested a custom pipeline that additionally identifies variants in conserved genomic regions[19] and outside of segmental duplication regions[20] (table 2). This custom pipeline identified ten candidate genes including *DHODH*, similar to what has been previously reported.[2] Finally, we tested a different custom pipeline that additionally remove variants with SIFT score >0.05 and PolyPhen2 score <0.85. This custom pipeline identified 14 candidate genes (table 2), but *DHODH* was not among them because one of the mutations (p.G202A) was predicted as tolerated by SIFT (score =0.18) and benign by PolyPhen (score =0.69). However, we note that the variant was correctly predicted as deleterious by LRT, Mutation Taster, PhyloP and GERP++. We caution that these algorithms present predictions that help users prioritise variants/genes, but the true sensitivity/specificity will depend on many factors, and that none of the algorithms constitute proof of being disease causal. In summary, this example has confirmed the utility of the wANNOVAR server in identifying a prioritised list of candidate disease causal genes, yet cautioned the judicious use of function prediction scores.

DISCUSSION

In this manuscript, we presented a web server called wANNOVAR for performing web-based functional annotation of genetic variants from personal genomes. Below we compare the server with other competing approaches and discuss potential future extensions and development.

Several similar web servers exist, including SIFT,[5] PolyPhen[6] and the SeattleSeq server.[21] The wANNOVAR server already incorporates SIFT and PolyPhen2 scores with additional scoring systems (table 1).[4] The wANNOVAR server differs from SeattleSeq in that: (1) it allows flexibility by permitting the users to select gene definition systems, including RefSeq genes,[22] ENSEMBL genes,[23] UCSC genes[24] or GENCODE genes.[25] Compared with the manually compiled RefSeq gene definitions, ENSEMBL genes and UCSC genes are supplemented with computational predictions of transcripts and genes. The GENCODE genes are compiled by a combination of initial manual annotation and experimental validation by the GENCODE consortium, and a refinement of the annotation based on these experimental results. All of the four gene definition systems are widely used in human genomic studies; (2) wANNOVAR produces more annotation results including predicted functional importance scores for non-synonymous variants; (3) wANNOVAR builds in a 'variants reduction' pipeline to facilitate identifying potential disease causal variants and genes from personal genomes.

The wANNOVAR server will be under constant development to improve its functionality. Some of the future plans include: First, we will explore the possibility of allowing FTP access to users with limited internet connection speed for uploading files. Second, we will add more annotation tasks for non-coding variants, splicing variants and UTR variants. Currently, the available annotations are strongly biased towards non-synonymous variants. With the accumulation of cell-type specific data on functional elements from large-scale genomics project, such as the ENCODE project[26] and the development of bioinformatics methods and databases,[27–32] we will be able to provide more annotations for variants outside of coding regions. Third, we will test the use of a backend computing cluster rather than a frontend web server to perform the actual annotation tasks to handle multiple simultaneous user queries. Fourth, we will explore the use of GALAXY[33] and design a plug-in based on ANNOVAR, for better annotating, processing and visualising variants.

In summary, wANNOVAR is an easy-to-use online tool for batch annotation of genetic variants. Given the rapid generation and accumulation of whole-exome or whole-genome sequencing data in research and clinical settings, we expect that wANNOVAR will help biologists and clinicians take advantage of personal genome information in various medical genetics applications.

Acknowledgments

The authors thank James Knowles, Jalas Chaim, Mingyao Li, Gholson Lyon and the three anonymous reviewers for testing the server and providing valuable feedbacks.

Funding The study is supported by start-up funds from the Zilkha Neurogenetic Institute and grant number HG006465 from NIH/NHGRI (K.W.).

REFERENCES

1. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008; 26:1135–1145. [PubMed: 18846087]

2. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. [PubMed: 20601685]
3. VCF4. http://www.1000genomes.org/wiki/Analysis/vcf4.0.
4. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011; 32:894–899. [PubMed: 21520341]
5. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–1081. [PubMed: 19561590]
6. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]
7. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res. 2009; 19:1553–1561. [PubMed: 19602639]
8. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010; 7:575–576. [PubMed: 20676075]
9. Siepel, A.; Pollard, KS.; Haussler, D. New methods for detecting lineage-specific selection; Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006); 2006. p. 190-205.
10. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
11. NHLBI-ESP. https://esp.gs.washington.edu/drupal/.
12. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]
13. Rope AF, Wang K, Evjenth R, Xing J, Johnston JJ, Swensen JJ, Johnson WE, Moore B, Huff CD, Bird LM, Carey JC, Opitz JM, Stevens CA, Jiang T, Schank C, Fain HD, Robison R, Dalley B, Chin S, South ST, Pysher TJ, Jorde LB, Hakonarson H, Lillehaug JR, Biesecker LG, Yandell M, Arnesen T, Lyon GJ. Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. Am J Hum Genet. 2011; 89:28–43. [PubMed: 21700266]
14. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]
15. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
16. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++ PLoS Comput Biol. 2010; 6 e1001025.
17. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara ECM, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM, Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z,

Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

18. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010; 42:30–35. [PubMed: 19915526]
19. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005; 15:1034–1050. [PubMed: 16024819]
20. Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. Segmental duplications: organization and impact within the current human genome project assembly. Genome Res. 2001; 11:1005–1017. [PubMed: 11381028]
21. SeattleSeq. http://gvs.gs.washington.edu/SeattleSeqAnnotation/.
22. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–D65. [PubMed: 17130148]
23. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. The Ensembl automatic gene annotation system. Genome Res. 2004; 14:942–950. [PubMed: 15123590]
24. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics. 2006; 22:1036–1046. [PubMed: 16500937]
25. Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol. 2006; 7(Suppl 1):S4.1–S4.9. [PubMed: 16925838]
26. Raney BJ, Cline MS, Rosenbloom KR, Dreszer TR, Learned K, Barber GP, Meyer LR, Sloan CA, Malladi VS, Roskin KM, Suh BB, Hinrichs AS, Clawson H, Zweig AS, Kirkup V, Fujita PA, Rhead B, Smith KE, Pohl A, Kuhn RM, Karolchik D, Haussler D, Kent WJ. ENCODE whole-genome data in the UCSC genome browser (2011 update). Nucleic Acids Res. 2011; 39:D871–D875. [PubMed: 21037257]
27. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011; 473:43–49. [PubMed: 21441907]
28. Levy S, Hannenhalli S. Identification of transcription factor binding sites in the human genome sequence. Mamm Genome. 2002; 13:510–514. [PubMed: 12370781]
29. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004; 32:5036–5044. [PubMed: 15448185]
30. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–D110. [PubMed: 16381825]
31. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res. 2011; 21:2167–2180. [PubMed: 21875935]
32. Markham NR, Zuker M. UNAFold: software for nucleic acid folding and hybridization. Methods Mol Biol. 2008; 453:3–31. [PubMed: 18712296]

33. Goecks J, Nekrutenko A, Taylor J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010; 11:R86. [PubMed: 20738864]

Table 1

Selected annotation tasks from the wANNOVAR server

Type	Column	Description
Gene annotation	Variant function	Exonic, intronic, intergenic, UTR, etc
	Gene	Impacted gene or neighbouring gene (with distance)
	Exonic variant function	Non-synonymous, synonymous, stopgain, etc
	AAChange	mRNA and amino acid change for coding variants
Variation databases	ESP5400_ALL	Allele frequency in 5400 NHLBI-ESP exomes
	1000G_ALL	Allele frequency in 1000 Genomes Project (currently, version 2012 Feb)
	dbSNP	dbSNP identifier (currently, version 135)
Functional prediction	AVSIFT	Base-level SIFT scores
	LJB_SIFT	1-SIFT scores and predictions (D: damaging, T: tolerated)
	LJB_PolyPhen2	PolyPhen 2 scores and predictions (D: probably damaging; P: possibly damaging; B: bening)
	LJB_LRT	LRT scores and predictions (D: deleterious; N: neutral; U: unknown)
	LJB_MutationTaster	MutationTaster scores and predictions (A: disease_causing_automatic; D: disease_causing; N: polymorphism; P: polymorphism_automatic)
	LJB_PhyloP	PhyloP conservation scores and predictions (C: conserved, N: non-conserved)
	GERP++	GERP++ scores for exonic variants
Region annotation	Conserved	Region-level phastCons LOD scores
	SegDup	Located in segmental duplication region and the sequence identity score

Table 2

Illustration of the “variants reduction” pipeline on the Ogden syndrome data set and the synthetic Miller syndrome data set

Data set	Ogden (exome variants in hg19 coordinate)		Miller (genome variants in hg18 coordinate)		
Variants reduction strategy	Default	Custom	Default	Custom	Custom
Input variants	1479	1479	4702187	4702187	4702187
Identify missense, nonsense and splicing variants	136	136	12410	12410	12410
Identify variants from conserved regions	–	–	–	5395	–
Remove variants in segmental duplications regions	–	–	–	5135	–
Remove variants observed in user-supplied controls	–	16 *	–	–	–
Remove variants observed in the 1000 Genomes Project with MAF>1%	19	3	2275	1116	2275
Remove variants observed in the NHLBI-ESP 5400 exomes with MAF>1%	14	3	1256	740	1256
Remove variants in dbSNP (excluding clinically associated SNPs)	1	1	516	313	516
Remove variants with SIFT score >0.05	–	1	–	–	395
Remove variants with PolyPhen2 score <0.85	–	1	–	–	351
Final list of candidate genes based on disease model	1	1	24	10	14
Correct causal gene identified?	Yes	Yes	Yes	Yes	No

* Two unaffected male family members were used as controls.

SNP, single nucleotide polymorphism.