Supp. Methods S1. Next-Generation Sequencing Methods and Annotation Notes

Solution hybridization exome capture was carried out using the SureSelect Human All Exon 50Mb Kit (Agilent Technologies, Santa Clara, CA). This technique uses biotinylated RNA baits to hybridize to sequences that correspond to exons(1). Manufacturer's protocol version 1.0 compatible with Illumina paired end sequencing was used, with the exception that DNA fragment size and quality was measured using a 2% agarose gel stained with Sybr Gold instead of using an Agilent Bioanalyzer. Additionally, the Illumina library preparation portion of the Sure-Select protocol was performed using the SPRIworks Fragment Library System (Beckman Coulter Genomics, Danvers, MA, USA) according to manufacturer's protocols. Agilent's specifications state that the capture regions total approximately 38 Mb. This kit covers the 1.22% of the human genome corresponding to the Consensus Conserved Domain Sequences database (CCDS) and greater than 1000 non coding RNAs. The 50Mb kit also includes exons defined by the Gencode Project (http://www.sanger.ac.uk/resources/databases/encode/). Flowcell preparation and 76 or 101bp paired end read sequencing were carried out as per protocol for the GAIIx sequencer (2) (Illumina Inc, San Diego CA). Image analysis and base calling on all lanes of data were performed using Illumina Genome Analyzer Pipeline software (GAPipeline versions 1.4.0 or greater) with default parameters.

Read mapping, variant calling and annotation

Reads were aligned to a human reference sequence (UCSC assembly hg18, NCBI build 36) using the package called "efficient large-scale alignment of nucleotide databases" (ELAND). Reads that align uniquely were grouped into genomic sequence intervals of about 100kb, and reads that fail to align were binned with their paired-end mates. Reads in each bin were subjected to a Smith-Waterman-based local alignment algorithm, *cross_match* using the parameters –minscore 21 and –masklevel 0 to their respective 100kb genomic sequence (http://www.phrap.org). Genotypes were called at all positions where there were high-quality sequence bases (Phred-like Q20 or greater) using a Bayesian algorithm (Most Probable Genotype – MPG; Nancy F Hansen, (3)). Genotypes with a MPG score of 10 or greater show >99.89% concordance with SNP Chip data. The targeted regions included the exons of 17,134 genes and total 36,025,890 bases in the human genome.[All exon 38Mb] The targeted regions included exons of 30,241 genes and total 51,499,639 bases in the human genome. [All Exon 50Mb]. Our annotation of cSNVs (coding single nucleotide variants) was based on UCSC all known genes. Missense variants were sorted by the degree of severity of functional disruption prediction using CDPred(4). Variants detected in dbSNP (version 130) were excluded from being potential disease causing cSNVs.

1) Gnirke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27, 182-9 (2009).
2) Bentley, D.R. et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53-9 (2008).
3) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, Abaan HO, Albert TJ; NISC Comparative Sequencing Program, Margulies EH,

Green ED, Collins FS, Mullikin JC, Biesecker LG. Genome Res. 2010 Oct;20(10):1420-31. Epub 2010 Sep 1.

4) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Johnston JJ, Teer JK, Cherukuri PF, Hansen NF, Loftus SK; NIH Intramural Sequencing Center, Chong K, Mullikin JC, Biesecker LG. Am J Hum Genet. 2010 May 14;86(5):743-8. Epub 2010 May 6.

Supp. Methods S2. Variant List Structure and the VarSifter File

The VarSifter data file (*.vs file) is a tab-delimited text file. The first row is a series of headers that label the contents of each column. All rows after the first row are individual variants, which may be substitutions, insertions/deletions, etc. A full explanation of the file structure can be found at the VarSifter website:

http://research.nhgri.nih.gov/software/VarSifter/guide.shtml

The easiest way for a new user to become familiar with the VarSifter format is to open up a small VarSifter file from the provided example file, for example

compexome_37a_homozygous_recessive.vs

with a spreadsheet program like Microsoft Excel. The VarSifter format is a specific example of a more general NextGen-sequencing file type, the "variant list". The fundamental components of a variant list form a "basic variant list" that includes the following information for each variant:

1. The Chromosome number.
2. Position in the chromosome.
    - The position is based on a published reference sequence, e.g. the National Center for Biotechnology Information (NCBI) human genome reference sequence, build 36 or "NCBI hg18 build 36".
    - The position in Varsifter is the left side (towards smaller number in the reference sequence for the given chromosome) flanking base and the right side flanking base. Other means for specifying positions may be used.
3. The original nucleotide(s) or structure found at the position (again based on a reference sequence).
4. The new nucleotide(s) or structure found at the position.

All of the other information in the variant list is added after the genotypes have been done. Common added data (and VarSifter examples include:

1. Gene context (in VarSifter the Gene_name, strand, ref_aa, var_aa, etc).
    - Gene context refers to a specific RNA transcript, which is usually the most common if that information is known.
2. Genotype Quality (in VarSifter the *.score and *.coverage columns).
3. A pathogenicity prediction (in VarSifter the CDPred_score).
    - Many different programs, including PolyPhen and SIFT, can be used to generate scores and different programs produce different types of output.

Other data may include dbSNP allele frequencies and links to the UCSC genome browser or other databases. A basic variant list can be annotated using web-based tools designed specifically for this purpose. Examples include:

1. ANNOVAR (http://www.openbioinformatics.org/annovar/).
2. SeattleSeq (http://gvs.gs.washington.edu/SeattleSeqAnnotation/).

At our site, the VarSifter file contains genotypes for each member of family submitted for exome sequencing (FATHER.NA.*, MOTHER.NA.*, etc). This type of annotation is generally not available at public annotation sites but may be added by custom scripting or incorporated into analysis at later steps in programs such as VAR-MD (soon to be made available) and VAAST

http://www.yandell-lab.org/software/vaast.html).

Readers who would like to manipulate a small VarSifter file to get a feel for the VarSifter program are encouraged to try a partially filtered from the Incremental Filtered VarSifter Files directory provided on the website:

ftp://ftp.nhgri.nih.gov/pub/NIHUDP/ADAMS_METHODS/Example_data_set/

Supp. Methods S3. Procedure for Generating Population Frequency Filter with Galaxy

1. Use Galaxy website
    a. http://main.g2.bx.psu.edu/
2. Obtain dbSNP data set, using specified options, modified as needed for the project at hand
    a. Get Data->UCSC Main table browser
    b. Table Browser Options
        i. Group:->Variation and Repeats
        ii. Track:->All SNPs(132)
        iii. Assembly->Feb. 2009(GRCh37/hg19)
        iv. Output format: BED – browser extensible data
        v. Filter: Create
            1. avHet is >= 0.01
            2. weight = 1 (this will select SNPs that align to a unique position)
            3. valid does include hapmap and by-1000-genomes (not *)
    c. Submit
    d. Get output
3. Send query to Galaxy
4. Lift-Over -> Convert genome coordinates between assemblies and genomes (we need to use this because our data set uses hg18 coordinates)
5. Operate on genomic intervals->Complement intervals of a dataset.
    a. Check genome wide complement
    b. Press Execute button
    c. (This operation creates a “reverse” BED file that *includes* everything that is not a suspect SNP. We make this file to use with VarSifter’s exclude BED file facility)
6. Download BED file

Supp. Table S1. Exome Project Design Considerations

Design Phase	Pre-Sequencing Data Collection	Data Acquisition	Raw Data to Annotated Variant List	Variant Filtering and Analysis	Variant Inspection, Validation, Reanalysis
Review Evidence For Genetic Etiology **Assess Material** Family members tissues/DNA **Technology** Genome/Exome RNA-seq Custom capture **Study Design** Single exome Multiple exome Multiple family **Adjunct Procedure** SNP array RNA array analysis **Data Acquisition** Collaborative Commercial At your facility **Analysis Plan** Alignment Base calling Analysis **Validation Plan** Sanger Functional	**Clinical Data** Family history Phenotypes Penetrance Consanguinity Disease freq. Genetic model **SNP Array Derived** Consanguinity Mosaicism Verify family relationships Dosage variants Recombination mapping **Gene Lists** *Inclusion* Cellular pathways Clinical hypotheses Genes in regions of interest *Exclusion* Highly variable genes **BED Files** Excluded base-pair list Segregation regions Defining other regions of interest	**Sample Library Preparation** Fragmentation End repair Adapter ligation Clonal amplification **Sequencing** Several methods available Generates a raw data file, e.g. FASTQ file ***Quality Assessment*** Library complexity Short read coverage	**Alignment of Short Reads to a Ref. Sequence** Generates an alignment file, e.g. SAM (Sequence Alignment Map) or BAM (Binary Alignment Map) **Genotype Calling** Individual variants Defined by a genomic position and base-pair (or structural) change **Annotation** Localization within known genes Amino acid change Predicted pathogenicity (many other possibilities) ***Quality Assessment*** Alignment Quality Genotype Call Quality Annotation Certainty *(End product is a list of annotated variants)*	**Inclusion/Exclusion by Gene Name** Pathway components Clinical hypotheses Gene exclusion lists **Inclusion/Exclusion Regions of Interest** Haplo -excess/ -insufficiency Segregation/ linkage intervals Regions of homozygozity **Population Frequency** **Variant Characteristics** Frequent false positives Segregation consistency **Genotype Category** Homozygous Heterozygous, etc **Pathogenicity Ranking** *(End product of filtering is a list of candidate variants)*	**Sanger Validation** Individual genotype calls need to be verified. ***Known Variants/ Genes:*** **Search for Previously Reported Variants Associated with Known Syndrome(s)** ***New Variants/ Genes:*** **Correlation with Clinical Phenotype** **Functional Validation** Experimental work to verify pathogenicity and phenotype causation **Iteration** Redo analysis/ revisit assumptions if no viable candidate variants remain

Supp. Table S2a. Candidate List Filtering for 22 Projects--All Family Data Included

Family	Exome Capture	Family Size	Used in Linkage	Affected	Total Variants	Population Frequency	Kill List	HWE	Linkage	Quality	Homoz.	Cmpnd Het
1	38MB	3	3	1	82030	31437	27320	20688	18247	8162	12	5
2	38MB	5	5	2	88274	38096	32518	24890	3838	1157	15	2
3	38MB	3	3	1	90904	35793	31464	24020	21099	9658	10	9
4	38MB	4	4	2	105648	44958	39925	32422	4665	2069	2	5
5	38MB	3	3	1	107065	46944	42052	32471	27778	8686	19	9
6	38MB	5	5	1	110168	49019	42644	33507	16226	3785	3	2
7	38MB	4	4	1	110974	49916	43673	34211	19686	5205	10	2
8	38MB	6	6	2	111357	50426	43571	33897	1569	550	9	1
9	38MB	4	4	1	111656	51433	45448	35173	19424	7209	2	4
10	38MB	4	4	2	111979	51624	45217	36918	6050	1785	5	1
11	38MB	4	3	1	112363	49048	41638	31973	7212	2655	5	5
12	38MB	5	5	1	115479	51464	44489	34035	13941	4171	10	3
13	38MB	3	3	1	117648	53059	46754	36401	31000	12362	14	11
14	38MB	4	4	2	118958	54979	49536	38989	33061	11193	6	3
15	38MB	4	4	1	121066	51996	45362	34976	17223	6413	5	0
16	38MB	5	4	1	133555	63478	56048	44285	31287	8592	17	12
17	38MB	6	6	2	134633	59631	51628	40883	4576	1510	1	0
18	50MB	4	4	1	187489	86594	74172	61738	42290	16062	13	12
19	50MB	3	3	1	190482	87781	75921	63537	26627	11327	8	7
20	50MB	5	5	1	197529	94695	82459	69143	29802	9554	12	14
21	50MB	5	5	1	200111	96418	83851	70381	34692	10435	20	7
22	50MB	4	4	1	219631	107441	94388	80632	50753	19589	11	15

Supp Table S2b. Candidate List Filtering for 22 Projects--Only Probands Included

Proband	Exome Capture	Family Size	Used in Linkage	Affected	Total Variants	Population Frequency	Kill List	HWE	Linkage	Quality	Homoz.	Heteroz.	>1 Het genes
1	38MB	NA	NA	NA	20916	20073	17643	12702	NA	11347	170	87	53
2	38MB	NA	NA	NA	45708	16809	13817	9760	NA	8759	229	94	60
3	38MB	NA	NA	NA	46849	17993	15033	11443	NA	10720	191	144	64
4	38MB	NA	NA	NA	51632	18980	16454	11637	NA	10414	186	93	50
5	38MB	NA	NA	NA	59699	22039	18525	13496	NA	12335	298	113	77
6	38MB	NA	NA	NA	60375	23149	19121	13924	NA	12640	253	194	78
7	38MB	NA	NA	NA	62131	23181	19571	14016	NA	12732	204	107	64
8	38MB	NA	NA	NA	62558	24683	22027	17448	NA	16225	185	225	133
9	38MB	NA	NA	NA	64287	25417	21872	15812	NA	14377	233	193	93
10	38MB	NA	NA	NA	69390	28582	24390	18631	NA	17380	259	233	98
11	38MB	NA	NA	NA	70438	30165	27141	20452	NA	18827	276	157	94
12	38MB	NA	NA	NA	73241	31810	28544	21299	NA	19615	265	155	69
13	38MB	NA	NA	NA	78343	34167	29960	22516	NA	20824	236	138	85
14	38MB	NA	NA	NA	79882	33459	28857	20847	NA	12654	218	149	72
15	38MB	NA	NA	NA	85734	39520	34868	28296	NA	17128	351	508	98
16	38MB	NA	NA	NA	91487	41883	36979	28288	NA	23315	287	180	98
17	38MB	NA	NA	NA	94444	44973	39788	29872	NA	15018	283	115	66
18	50MB	NA	NA	NA	113398	49068	41770	32529	NA	28440	364	231	124
19	50MB	NA	NA	NA	115455	52089	45073	36201	NA	32740	319	267	115
20	50MB	NA	NA	NA	118437	50535	42891	33900	NA	29767	356	233	113
21	50MB	NA	NA	NA	123628	55479	47438	37908	NA	34169	363	237	109
22	50MB	NA	NA	NA	124535	56476	49089	39868	NA	36604	318	396	174

Supp. Table S2 Legend: Cumulative Candidate List Filtering Results, Families Versus Probands-Only

Supp. Table S2a lists filtering data that includes both SNP array and exome data from selected nuclear family members. Supp. Table S2b shows the same projects, with only exome proband data included. The last two data points ("Heterozygotes" and "Homozygotes") are not cumulative, but represent separate application of heterozygote-only, and homozygote-only filters. Note that the implementation of homozygote and heterozygote filters differs between single exome analyses and family based analyses. Mendelian segregation and phase information is not available in the case of single exome analysis. Homozygotes are not checked for inheritance from both parents. The "heterozygote" count is a tabulation of all pair-wise combinations of variants for those cases where more than one heterozygous variant is found in the same gene. Single exome projects start with fewer variants and end with a larger number of candidates for further study. See the text for a further explanation of the various filtration steps.

The table column labels are as follows:

- Family: A set of twenty-two separated families are listed and numbered. The numbers between Supp. Tables S2a and S2b are the same.
- Exome Capture: The Aligent SureSelect kit (38 Mb or 50 Mb) that was used to sequence the family.
- Family Size: The number of individuals in the family.
- Used in Linkage: The number of family members for whom SNP arrays were used to map recombinations.
- Affected: The number of affected family members. In these examples only children were affected.
- Total Variants: The total number of variants generated by the data acquisition (capture, flow cell and genotyping) phase of the exome project.
- Population Frequency: The number of variants left after filtering by population frequency (1% threshold, 1000 genomes and HapMap variants only).
- Kill List: The number of variants remaining after filtering the value from the previous column using our gene exclusion list.
- HWE: The number of variants remaining after filtering the value from the previous column using our base pair exclusion list. The notation HWE (for Hardy Weinberg Equilibrium) is used as the base pair exclusion list comprises variants detected due to being out of HWE (see text).
- Linkage: The number of variants remaining after filtering the value from the previous column using a BED-formatted "include" file of regions that segregated according an autosomal recessive model of inheritance and with the phenotypes assigned to the family by clinical evaluation.

- Quality: The number of variants remaining after filtering the value from the previous column using a filter that demanded that each family member have high quality genotypes (MPG >= 10 and MPG/coverage >= 0.5).
- Homoz.: The number of variants remaining after filtering the value from the Quality column, demanding that the variant be a homozygous non-reference in the affected individuals AND the involved variants segregated in a Mendelian-consistent manner.
- Cmpnd Het: The number of variants remaining after filtering the value from the Quality column, demanding that paired variants were consistent with heterozygous inheritance AND involved variants that segregated in a Mendelian-consistent manner.