The American Journal of Human Genetics, Volume *104*

Supplemental Data

A CCR4-NOT Transcription Complex, Subunit 1, *CNOT1*, Variant Associated with Holoprosencephaly

Paul Kruszka, Seth I. Berger, Karin Weiss, Joshua L. Everson, Ariel F. Martinez, Sungkook Hong, Kwame Anyane-Yeboa, Robert J. Lipinski, and Maximilian Muenke

SUPPLEMENTAL MATERIAL AND METHODS

Table S1. Holoprosencephaly cohort.

	Trios (n=134)
Average age	2.8 years
Gender	42% male
HPE type	
Alobar	18%
Semilobar	49%
Lobar	23%
Middle interhemispheric variant (MIHV)	5%
Microform	4%
Ethnicity	
Caucasian	61%
African American	1%
Latin American	25%
Asian	6%
Native American	0%
Middle Eastern	3%

Table S2. *De novo* variants in proband 1 and 2.

	Chromosome	Position (hg19)	Variant type	Gene	Variant	ExAC frequency	CADD score
Proband 2	Chr16	58610468	nonsynonymous	*CNOT1*	c.1603C>T:p.Arg535Cys (GenBank: NM_001265612.1)	0	35
Proband 1	Chr16	58610468	nonsynonymous	*CNOT1*	c.1603C>T:p.Arg535Cys	0	35
Proband 1	Chr17	78210857	synonymous	*SLC26A11*	c.867A>G:p.(=) (GenBank: NM_000199.4)	0	
Proband 1	Chr19	48722162	synonymous	*CARD8*	c.1119T>C:p.(=) (GenBank: NM_001184902.1)	0	
Proband 1	Chr4	141889002	nonsynonymous	*RNF150*	c.510G>A:p.Met170Ile (GenBank: NM_020724.2)	0	29

SUPPLEMENTAL METHODS

Exome Sequencing

Exome sequencing, assembly, genotyping, and annotation were carried out by the National

Intramural Sequencing Center (NISC). Genomic DNA (approximately 1 µg) was fragmented to an average size of 150 bp and subjected to DNA library creation using established Illumina paired-end protocols. Capture utilized the NimbleGen SeqCap EZ Version 3.0+ UTR (Roche NimbleGen, Madison, WI). Captured regions totaled approximately 96 Mb. Flow cell preparation and 125-bp paired end read sequencing were performed as per the HiSeq2000 Sequencer protocol (Illumina, San Diego, CA).

Read mapping, variant calling and annotation. Fastq files were then aligned to reference genome human_g1k_v37_decoy using bwa mem and sam output was compressed to bam format using picard SamFormatConverter. The aligned bamfile was sorted and indexed using samtools and ReadGroups based on Sample ID were added with Picard AddOrReplaceReadGroups command. Picard's MarkDuplicates command was then applied. At this point the file was processed through a GATK 3.6 pipeline based on the recommended best practices using the genome capture intervals utilized by the EXaC consortium for exome targets with interval padding of 100 basepairs. Targets were realigned using the RealignerTargetCreator tools with the GATK resource bundle's 1000G_phase1.indels.b37.vcf and Mills_and_1000G_gold_standard.indels.b37.vcf. IndelRealigner was then applied with the intervals identified. The BaseRecalibrator was then applied using the GATK resource bundle's known sites from dbsnp_138.hg19.vcf.gz, 000G_phase1.indels.b37.vcf, and Mills_and_1000G_gold_standard.indels.b37.vcf. PrintReads was then used to generate the recalibrated file utilizing the data table generated from the previous step. GATK HaplotypeCaller was then used to generate g.vcf files for each reprocessed bam file. All g.vcf files were simultaneously passed to GATK's GenotypeGVCFs to generate a combined joint called vcf file. Variant Quality Score Recalibration pipeline was then applied. First the SNP VQSR was performed using the GATK VariantRecalibrator using annotations of QD, MQRankSum, ReadPosRankSum, FS, MQ, and InbreedingCoeff. Resources utilized from the GATK resource bundle included hapmap_3.3.b37.vcf , 1000G_omni2.5.b37.vcf,

1000G_phase1.snps.high_confidence.b37.vcf, dbsnp_138.b37.vcf, and dbsnp_138.b37.excluding_sites_after_129.vcf. Indel VariantRecalibrator was performed using annotations of FS, ReadPosRankSum, InbreedingCoeff, MQRankSum, and QD. Resources used from GATK resource bundle included Mills_and_1000G_gold_standard.indels.b37.vcf, Axiom_Exome_Plus.genotypes.all_populations.poly.vcf, and dbsnp_138.b37.vcf. The SNP recalibration was then applied using APplyRecalibration with a filter level of 99.6 and the INDEL recalibration was applied with a filter level of 95.0. Genomic posteriors were calculated for each call using the GATK CalculateGenotypePosteriors with supporting data from GATK resource bundle 1000G_phase3_v4_20130502.sites.vcf.gz and a pedigree file containing relationships between parents and probands in the trios. Genotypes with GQ less than 20 were labeled with lowGQ using GATK VariantFiltration. GATK VariantAnnotator was then applied to label PossibleDeNovo variants.

Variant sites with multiple alleles were split into single line entries in the vcf file using bcftools norm. Indels were left aligned and normalized using bcftools. VCF file was then annotated using Annovar's table_annovar command to annotated with refGene annotations, frequency information from exac, 1000 genomes, kaviar, and haplotype reference consortium, and scores from GERP, CADD 1.3, DANN, FATHMM, EIGEN, GWAVA, and DBNSFP30a.

A custom perl script then processed the table output to label inheritance calls and filter out low quality and common variants. Filter settings included QD>2, DP>5930 (to restrict to reads with an average call depth of 10x across all samples), ExcessiveHets<10, Maximum frequency in Exac or other population database or subpopulation database of 0.001, Passing the VQSR filter, Function of Exonic or splicing, and exonic function not being a synonymous SNV. This variant list was further filtered to identify presumed de novo variants where the trio genotypes are called such that both parents are homozygous reference, and the proband was called heterozygous. Further quality filtering was applied to

ensure that all genotypes in the trio have a QD greater than 20, Depth greater than 10. Also required that both parents have a variant call depth of 0 while the proband have a variant call depth greater than 4, a reference call depth greater than 4, and the reference calls must make up between 20% to 80% of the calls that that locus. To remove calls resulting from systematic sequencing noise, we excluded variants with an allele count greater than 3 in our samples.

Copy number variation calling

Copy number variation (CNV) prediction from exome data was done using the XHMM (eXome-Hidden Markov Model) caller. We used GATK to generate the depth of coverage statistics required for XHMM from the BAM files of our HPE cohort and a control set. GATK output was then run through the XHMM pipeline, generating a VCF file containing each predicted CNV. We then annotated each CNV for genes contained and cytogenetic region using Annovar. A custom perl script was used to evaluate for any de-novo CNVs in the HPE probands. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

Sanger Sequencing

Variant sequence verification was performed using standard methods (Sanger et al. 1977). Sequencing was performed with v3.1 BigDye Terminator Cycle Sequencing Kit (Life Technologies, Grand Island, NY) in the ABI 3730xl Sequencer (Life Technologies) according to the manufacturer's protocol. Sequence data were aligned to the published reference genomic sequences (GenBank) using Sequencher 5.0.1 (Gene Codes Corp., Ann Arbor, MI).