###Final Take-Home Project###

##My name is Zhongyi Zhang.##

#BIOINFO

I choose the topic 1 --> The Pevsner autism one.

1) Call variants on Pevsner autism bam (Web Document 9.7 at http://www.bioinfbook.org/php/ (Links to an external site.)Links to an external site. Links to an external site.then annotate with snpEff + Clinvar and upload (unannotated) VCF to VEP for gnomAD (not ExAC) population frequencies. Compare variants in the 101 target gene list with gnomAD frequencies expected autism frequencies. Use hg19.fa (see Supplement 1 info below), not Pevsner's WebDocument\_9-6\_101autism.fa. Use CDC 2016 frequencies: https://www.cdc.gov/ncbddd/autism/data.html (Links to an external site.)Links to an external site. ; Use the frequency for the most recent year available, 2014 (birth year 2006).

---------------------------------------------------------------------------------------------------------------------------------------------------------------------------

[student@msbi32400lab5 ~]$ cd /data/final/data/

[student@msbi32400lab5 data]$ chmod +x twoBitToFa

[student@msbi32400lab5 data]$ ./twoBitToFa hg19.2bit hg19.fa

[student@msbi32400lab5 data]$ samtools faidx hg19.fa

Now in the data folder, I have hg19.2bit, hg19.fa, hg19.fa.fai, twoBitToFa, and WebDocument\_9-7\_mysample1.bam

[student@msbi32400lab5 data]$ samtools mpileup -uf hg19.fa WebDocument\_9-7\_mysample1.bam | bcftools call -mv > WebDocument\_9-7\_mysample1\_var.raw.vcf

Note: none of --samples-file, --ploidy or --ploidy-file given, assuming all sites are diploid

[warning] samtools mpileup option `u` is functional, but deprecated. Please switch to using bcftools mpileup in future.

[mpileup] 1 samples in 1 input files

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

[student@msbi32400lab5 data]$ less WebDocument\_9-7\_mysample1\_var.raw.vcf

##fileformat=VCFv4.2

##FILTER=<ID=PASS,Description="All filters passed">

##samtoolsVersion=1.9+htslib-1.9

##samtoolsCommand=samtools mpileup -uf hg19.fa WebDocument\_9-7\_mysample1.bam

##reference=file://hg19.fa

##contig=<ID=chrM,length=16571>

##contig=<ID=chr1,length=249250621>

##contig=<ID=chr2,length=243199373>

##contig=<ID=chr3,length=198022430>

##contig=<ID=chr4,length=191154276>

##contig=<ID=chr5,length=180915260>

##contig=<ID=chr6,length=171115067>

##contig=<ID=chr7,length=159138663>

##contig=<ID=chr8,length=146364022>

##contig=<ID=chr9,length=141213431>

##contig=<ID=chr10,length=135534747>

##contig=<ID=chr11,length=135006516>

##contig=<ID=chr12,length=133851895>

##contig=<ID=chr13,length=115169878>

##contig=<ID=chr14,length=107349540>

##contig=<ID=chr15,length=102531392>

##contig=<ID=chr16,length=90354753>

##contig=<ID=chr17,length=81195210>

##contig=<ID=chr18,length=78077248>

##contig=<ID=chr19,length=59128983>

##contig=<ID=chr20,length=63025520>

##contig=<ID=chr21,length=48129895>

##contig=<ID=chr22,length=51304566>

##contig=<ID=chrX,length=155270560>

##contig=<ID=chrY,length=59373566>

##ALT=<ID=\*,Description="Represents allele(s) other than observed.">

##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">

WebDocument\_9-7\_mysample1\_var.raw.vcf

...

[student@msbi32400lab5 data]$ bcftools filter -s LowQual -e '%QUAL<20' WebDocument\_9-7\_mysample1\_var.raw.vcf > WebDocument\_9-7\_mysample1\_var.flt.vcf

[student@msbi32400lab5 data]$ less WebDocument\_9-7\_mysample1\_var.flt.vcf

##fileformat=VCFv4.2

##FILTER=<ID=PASS,Description="All filters passed">

##samtoolsVersion=1.9+htslib-1.9

##samtoolsCommand=samtools mpileup -uf hg19.fa WebDocument\_9-7\_mysample1.bam

##reference=file://hg19.fa

##contig=<ID=chrM,length=16571>

##contig=<ID=chr1,length=249250621>

##contig=<ID=chr2,length=243199373>

##contig=<ID=chr3,length=198022430>

##contig=<ID=chr4,length=191154276>

##contig=<ID=chr5,length=180915260>

##contig=<ID=chr6,length=171115067>

##contig=<ID=chr7,length=159138663>

##contig=<ID=chr8,length=146364022>

##contig=<ID=chr9,length=141213431>

##contig=<ID=chr10,length=135534747>

##contig=<ID=chr11,length=135006516>

##contig=<ID=chr12,length=133851895>

##contig=<ID=chr13,length=115169878>

##contig=<ID=chr14,length=107349540>

##contig=<ID=chr15,length=102531392>

##contig=<ID=chr16,length=90354753>

##contig=<ID=chr17,length=81195210>

##contig=<ID=chr18,length=78077248>

##contig=<ID=chr19,length=59128983>

##contig=<ID=chr20,length=63025520>

##contig=<ID=chr21,length=48129895>

##contig=<ID=chr22,length=51304566>

##contig=<ID=chrX,length=155270560>

##contig=<ID=chrY,length=59373566>

##ALT=<ID=\*,Description="Represents allele(s) other than observed.">

##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">

...

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

[student@msbi32400lab5 data]$ java -Xmx2G -jar /data/snpEff/snpEff.jar eff -canon -noLog hg19 WebDocument\_9-7\_mysample1\_var.flt.vcf > /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.vcf

[student@msbi32400lab5 data]$ less /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.vcf

##fileformat=VCFv4.2

##FILTER=<ID=PASS,Description="All filters passed">

##samtoolsVersion=1.9+htslib-1.9

##samtoolsCommand=samtools mpileup -uf hg19.fa WebDocument\_9-7\_mysample1.bam

##reference=file://hg19.fa

##contig=<ID=chrM,length=16571>

##contig=<ID=chr1,length=249250621>

##contig=<ID=chr2,length=243199373>

##contig=<ID=chr3,length=198022430>

##contig=<ID=chr4,length=191154276>

##contig=<ID=chr5,length=180915260>

##contig=<ID=chr6,length=171115067>

##contig=<ID=chr7,length=159138663>

##contig=<ID=chr8,length=146364022>

##contig=<ID=chr9,length=141213431>

##contig=<ID=chr10,length=135534747>

##contig=<ID=chr11,length=135006516>

##contig=<ID=chr12,length=133851895>

##contig=<ID=chr13,length=115169878>

##contig=<ID=chr14,length=107349540>

##contig=<ID=chr15,length=102531392>

##contig=<ID=chr16,length=90354753>

##contig=<ID=chr17,length=81195210>

##contig=<ID=chr18,length=78077248>

##contig=<ID=chr19,length=59128983>

##contig=<ID=chr20,length=63025520>

##contig=<ID=chr21,length=48129895>

##contig=<ID=chr22,length=51304566>

##contig=<ID=chrX,length=155270560>

##contig=<ID=chrY,length=59373566>

##ALT=<ID=\*,Description="Represents allele(s) other than observed.">

##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">

##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">

##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">

##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">

##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">

##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">

##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">

##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">

##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">

##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">

##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">

##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">

##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">

##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">

##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">

##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">

##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">

##bcftools\_callVersion=1.9+htslib-1.9

##bcftools\_callCommand=call -mv; Date=Wed Mar 20 23:39:03 2019

##FILTER=<ID=LowQual,Description="Set if true: %QUAL<20">

##bcftools\_filterVersion=1.9+htslib-1.9

##bcftools\_filterCommand=filter -s LowQual -e %QUAL<20 WebDocument\_9-7\_mysample1\_var.raw.vcf; Date=Wed Mar 20 23:49:34 2019

##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"

##SnpEffCmd="SnpEff hg19 WebDocument\_9-7\_mysample1\_var.flt.vcf "

##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation\_Impact | Gene\_Name | Gene\_ID | Feature\_Type | Feature\_ID | Transcript\_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.le

ngth | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO' ">

##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene

\_Name | Gene\_ID | Number\_of\_transcripts\_in\_gene | Percent\_of\_transcripts\_affected'">

##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format

: 'Gene\_Name | Gene\_ID | Number\_of\_transcripts\_in\_gene | Percent\_of\_transcripts\_affected'">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample2

chrM 410 . A T 59.0 PASS DP=3;VDB=0.76;SGB=-0.453602;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,1,1;MQ=44;ANN=T|upstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.-240A>T|||||240|,T|upstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.-1263A>T|||||1263|,T|intergenic\_region|MODIFIER|CHR\_START-RNR1|CHR\_START-RNR1|intergenic\_region|CHR\_START-RNR1|||n.410A>T|||||| GT:PL 1/1:89,6,0

chrM 5581 . C T 7.30814 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=36;ANN=T|downstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.\*3978C>T|||||3978|,T|downstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.\*2351C>T|||||2351|,T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.5581C>T|||||| GT:PL 1/1:36,3,0

chrM 6588 . C T 27.4222 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=T|downstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.\*4985C>T|||||4985|,T|downstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.\*3358C>T|||||3358|,T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.6588C>T|||||| GT:PL 1/1:57,3,0

chrM 9378 . G A 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=A|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.9378G>A|||||| GT:PL

1/1:60,3,0

chrM 14770 . A G 30.4183 PASS DP=3;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=G|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.14770A>G|||||| GT:PL

1/1:60,3,0

chrM 14906 . A G 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=G|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.14906A>G|||||| GT:PL

1/1:39,3,0

chrM 15933 . C T 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.15933C>T|||||| GT:PL

1/1:39,3,0

chrM 15943 . T C 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=C|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.15943T>C|||||| GT:PL

...

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

[student@msbi32400lab5 data]$ java -Xmx2G -jar /data/snpEff/SnpSift.jar annotate -noLog /data/snpEff/data/hg19/clinvar/clinvar\_20190211.vcf.gz /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.vcf > /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.clinvar.vcf

[student@msbi32400lab5 data]$ less /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.clinvar.vcf

##fileformat=VCFv4.2

##FILTER=<ID=PASS,Description="All filters passed">

##samtoolsVersion=1.9+htslib-1.9

##samtoolsCommand=samtools mpileup -uf hg19.fa WebDocument\_9-7\_mysample1.bam

##reference=file://hg19.fa

##contig=<ID=chrM,length=16571>

##contig=<ID=chr1,length=249250621>

##contig=<ID=chr2,length=243199373>

##contig=<ID=chr3,length=198022430>

##contig=<ID=chr4,length=191154276>

##contig=<ID=chr5,length=180915260>

##contig=<ID=chr6,length=171115067>

##contig=<ID=chr7,length=159138663>

##contig=<ID=chr8,length=146364022>

##contig=<ID=chr9,length=141213431>

##contig=<ID=chr10,length=135534747>

##contig=<ID=chr11,length=135006516>

##contig=<ID=chr12,length=133851895>

##contig=<ID=chr13,length=115169878>

##contig=<ID=chr14,length=107349540>

##contig=<ID=chr15,length=102531392>

##contig=<ID=chr16,length=90354753>

##contig=<ID=chr17,length=81195210>

##contig=<ID=chr18,length=78077248>

##contig=<ID=chr19,length=59128983>

##contig=<ID=chr20,length=63025520>

##contig=<ID=chr21,length=48129895>

##contig=<ID=chr22,length=51304566>

##contig=<ID=chrX,length=155270560>

##contig=<ID=chrY,length=59373566>

##ALT=<ID=\*,Description="Represents allele(s) other than observed.">

##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">

##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">

##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">

##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">

##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">

##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">

##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">

##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">

##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">

##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">

##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">

##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">

##INFO=<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better)">

##INFO=<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)">

##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">

##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">

##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">

##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">

##bcftools\_callVersion=1.9+htslib-1.9

##bcftools\_callCommand=call -mv; Date=Wed Mar 20 23:39:03 2019

##FILTER=<ID=LowQual,Description="Set if true: %QUAL<20">

##bcftools\_filterVersion=1.9+htslib-1.9

##bcftools\_filterCommand=filter -s LowQual -e %QUAL<20 WebDocument\_9-7\_mysample1\_var.raw.vcf; Date=Wed Mar 20 23:49:34 2019

##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"

##SnpEffCmd="SnpEff hg19 WebDocument\_9-7\_mysample1\_var.flt.vcf "

##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation\_Impact | Gene\_Name | Gene\_ID | Feature\_Type | Feature\_ID | Transcript\_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.le

ngth | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO' ">

##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene\_Name | Gene\_ID | Number\_of\_transcripts\_in\_gene | Percent\_of\_transcripts\_affected'">

##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Format: 'Gene\_Name | Gene\_ID | Number\_of\_transcripts\_in\_gene | Percent\_of\_transcripts\_affected'">

##SnpSiftVersion="SnpSift 4.3t (build 2017-11-24 10:18), by Pablo Cingolani"

##SnpSiftCmd="SnpSift Annotate /data/snpEff/data/hg19/clinvar/clinvar\_20190211.vcf.gz /data/final/results/WebDocument\_9-7\_mysample1\_var\_flt\_snpEff.vcf"

##INFO=<ID=DBVARID,Number=.,Type=String,Description="nsv accessions from dbVar for the variant">

##INFO=<ID=ALLELEID,Number=1,Type=Integer,Description="the ClinVar Allele ID">

##INFO=<ID=CLNSIG,Number=.,Type=String,Description="Clinical significance for this single variant">

##INFO=<ID=CLNVCSO,Number=1,Type=String,Description="Sequence Ontology id for variant type">

##INFO=<ID=CLNREVSTAT,Number=.,Type=String,Description="ClinVar review status for the Variation ID">

##INFO=<ID=RS,Number=.,Type=String,Description="dbSNP ID (i.e. rs number)">

##INFO=<ID=CLNDNINCL,Number=.,Type=String,Description="For included Variant : ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB">

##INFO=<ID=ORIGIN,Number=.,Type=String,Description="Allele origin. One or more of the following values may be added: 0 - unknown; 1 - germline; 2 - somatic; 4 - inherited; 8 - paternal; 16 - maternal; 32 - de-novo; 64 - biparental; 128 - uniparental; 256 - not-tested; 512 - tested-inconclusive; 1073741824 - other">

##INFO=<ID=MC,Number=.,Type=String,Description="comma separated list of molecular consequence in the form of Sequence Ontology ID|molecular\_consequence">

##INFO=<ID=CLNDN,Number=.,Type=String,Description="ClinVar's preferred disease name for the concept specified by disease identifiers in CLNDISDB">

##INFO=<ID=CLNVC,Number=1,Type=String,Description="Variant type">

##INFO=<ID=CLNVI,Number=.,Type=String,Description="the variant's clinical sources reported as tag-value pairs of database and variant identifier">

##INFO=<ID=AF\_EXAC,Number=1,Type=Float,Description="allele frequencies from ExAC">

##INFO=<ID=AF\_ESP,Number=1,Type=Float,Description="allele frequencies from GO-ESP">

##INFO=<ID=CLNSIGINCL,Number=.,Type=String,Description="Clinical significance for a haplotype or genotype that includes this variant. Reported as pairs of VariationID:clinical significance.">

##INFO=<ID=CLNDISDB,Number=.,Type=String,Description="Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN">

##INFO=<ID=GENEINFO,Number=1,Type=String,Description="Gene(s) for the variant reported as gene symbol:gene id. The gene symbol and id are delimited by a colon (:) and each pair is delimited by a vertical bar (|)">

##INFO=<ID=CLNDISDBINCL,Number=.,Type=String,Description="For included Variant: Tag-value pairs of disease database name and identifier, e.g. OMIM:NNNNNN">

##INFO=<ID=AF\_TGP,Number=1,Type=Float,Description="allele frequencies from TGP">

##INFO=<ID=CLNSIGCONF,Number=.,Type=String,Description="Conflicting clinical significance for this single variant">

##INFO=<ID=CLNHGVS,Number=.,Type=String,Description="Top-level (primary assembly, alt, or patch) HGVS expression.">

##INFO=<ID=SSR,Number=1,Type=Integer,Description="Variant Suspect Reason Codes. One or more of the following values may be added: 0 - unspecified, 1 - Paralog, 2 - byEST, 4 - oldAlign, 8 - Para\_EST, 16 - 1kg\_failed, 1024 - other">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample2

chrM 410 . A T 59.0 PASS DP=3;VDB=0.76;SGB=-0.453602;MQSB=1;MQ0F=0;AC=2;AN=2;DP4=0,0,1,1;MQ=44;ANN=T|upstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.-240A>T|||||240|,T|upstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.-1263A>T|||||1263|,T|intergenic\_region|MODIFIER|CHR\_START-RNR1|CHR\_START-RNR1|intergenic\_region|CHR\_START-RNR1|||n.410A>T|||||| GT:PL 1/1:89,6,0

chrM 5581 . C T 7.30814 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=36;ANN=T|downstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.\*3978C>T|||||3978|,T|downstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.\*2351C>T|||||2351|,T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.5581C>T|||||| GT:PL 1/1:36,3,0

chrM 6588 . C T 27.4222 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=T|downstream\_gene\_variant|MODIFIER|RNR1|RNR1|transcript|NR\_137294.1|pseudogene||n.\*4985C>T|||||4985|,T|downstream\_gene\_variant|MODIFIER|RNR2|RNR2|transcript|NR\_137295.1|pseudogene||n.\*3358C>T|||||3358|,T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.6588C>T|||||| GT:PL 1/1:57,3,0

chrM 9378 . G A 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=A|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.9378G>A|||||| GT:PL

1/1:60,3,0

chrM 14770 . A G 30.4183 PASS DP=3;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=G|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.14770A>G|||||| GT:PL

1/1:60,3,0

chrM 14906 . A G 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=G|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.14906A>G|||||| GT:PL

1/1:39,3,0

chrM 15933 . C T 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=T

|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.16173C>T|||||| GT:PL

1/1:60,3,0

chrM 16322 . T C 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=C|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.16322T>C|||||| GT:PL

1/1:60,3,0

chrM 16329 . C T 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=T|intergenic\_region|MODIFIER|RNR2-CHR\_END|RNR2-CHR\_END|intergenic\_region|RNR2-CHR\_END|||n.16329C>T|||||| GT:PL

1/1:60,3,0

chr1 844860 . T C 3.22451 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=C|intergenic\_region|MODIFIER|FAM41C-LOC100130417|FAM41C-LOC100130417|intergenic\_region|FAM41C-LOC100130417|||n.844860T>C|||||| GT:PL 1/1:30,3,0

chr1 844878 . T G 7.30814 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=G|intergenic\_region|MODIFIER|FAM41C-LOC100130417|FAM41C-LOC100130417|intergenic\_region|FAM41C-LOC100130417|||n.844878T>G|||||| GT:PL 1/1:36,3,0

chr1 844906 . T G 8.13869 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,0,1;MQ=60;ANN=G|intergenic\_region|MODIFIER|FAM41C-LOC100130417|FAM41C-LOC100130417|intergenic\_region|FAM41C-LOC100130417|||n.844906T>G|||||| GT:PL 1/1:37,3,0

chr1 963700 . GCCCCCCCCCCCCCCCC GCCCCCCCCC 104.0 PASS INDEL;IDV=12;IMF=0.705882;DP=17;VDB=0.117011;SGB=-0.670168;MQ0F=0;AC=2;AN=2;DP4=0,0,0,10;MQ=29;ANN=GCCCCCCCCC|intron\_variant|MODIFIER|AGRN|AGRN|transcript|NM\_001305275.1|protein\_coding|2/37|c.463+5868\_463+5874delCCCCCCC||||||WARNING\_TRANSCRIPT\_MULTIPLE\_STOP\_CODONS

GT:PL 1/1:134,30,0

chr1 963824 . G A 39.1655 PASS DP=5;VDB=0.764235;SGB=-0.511536;RPB=0.666667;MQB=0.833333;BQB=0.5;MQ0F=0;ICB=1;HOB=0.5;AC=1;AN=2;DP4=0,2,0,3;MQ=47;ANN=A|intron\_variant|MODIFIER|AGRN|AGRN|transcript|NM\_001305275.1|protein\_coding|2/37|c.463+5982G>A||||||WARNING\_TRANSCRIPT\_MULTIPLE\_STOP\_CODONS GT:PL 0/1:72,0,44

chr1 1057422 . G C 9.88514 LowQual DP=1;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=C|intergenic\_region|MODIFIER|C1orf159-LINC01342|C1orf159-LINC01342|intergenic\_region|C1orf159-LINC01342|||n.1057422G>C|||||| GT:PL 1/1:39,3,0

chr1 1057508 . C G 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=G|intergenic\_region|MODIFIER|C1orf159-LINC01342|C1orf159-LINC01342|intergenic\_region|C1orf159-LINC01342|||n.1057508C>G|||||| GT:PL 1/1:60,3,0

chr1 1057524 . C T 30.4183 PASS DP=2;SGB=-0.379885;MQ0F=0;AC=2;AN=2;DP4=0,0,1,0;MQ=60;ANN=T|intergenic\_region|MODIFIER|C1orf159-LINC01342|C1orf159-LINC01342|intergenic\_region|C1orf159-LINC01342|||n.1057524C

>T|||||| GT:PL 1/1:60,3,0

chr1 1100319 . C A 40.4148 PASS DP=4;VDB=0.22;SGB=-0.453602;MQ0F=0;AC=2;AN=2;DP4=0,0,2,0;MQ=60;ANN=A|upstream\_gene\_variant|MODIFIER|MIR200B|MIR200B|transcript|NR\_029639.1|pseudogene||n.-2165C>A|||||2165|,A|upstream\_gene\_variant|MODIFIER|MIR200A|MIR200A|transcript|NR\_029834.1|pseudogene||n.-2924C>A|||||2924|,A|upstream\_gene\_variant|MODIFIER|MIR429|MIR429|transcript|NR\_029957.1|pseudogene||n.-4066C>A|||||4066|,A|intergenic\_region|MODIFIER|LINC01342-MIR200B|LINC01342-MIR200B|intergenic\_region|LINC01342-MIR200B|||n.1100319C>A|||||| GT:PL 1/1:70,6,0

...

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

e!GRCh37

VEP - Variant Effect Predictor website. I click the "Web interface Launch Ve!P".

Then I go to the website "http://grch37.ensembl.org/Homo\_sapiens/Tools/VEP?db=core" and upload the file "WebDocument\_9-7\_mysample1\_var.flt.vcf" (The Unannotated one). (I clicked GRCh37 website. It changed to GRCh37.p13)

For Transcript database to use section, I choose the "RefSeq transcripts".

I clicked and checked HGVS for identifiers section.

I choose the "gnomAD (exomes) allele frequencies" and "1000 Genomes global minor allele frequency" for "Frequency data for co-located variants" in the "Variants and frequency data" section.

For Addional annotations, I checked the "Exon and intron numbers" and cancelled the "Transcript support level" and "APPRIS".

I keep the "Prediction and score" choice for "SIFT" and "PolyPhen".

Lastly, I put "Showing one selected consequence per variant allele" for "Filters".

Wait for the website, and clicked "View Results" once finished.

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Summary Satatistics:

Category Count

Variants processed 19124

Variants filtered out 0

Novel / existing variants 2830 (14.8) / 16294 (85.2)

Overlapped genes 4468

Overlapped transcripts 4575

Overlapped regulatory features 838

Consequences (all):

intron\_variant: 44%

intergenic\_variant: 35%

regulatory\_region\_variant: 6%

non\_coding\_transcript\_variant: 4%

upstream\_gene\_variant: 3%

downstream\_gene\_variant: 3%

synonymous\_variant: 1%

3\_prime\_UTR\_variant: 1%

missense\_variant: 1%

Others

Coding consequences

synonymous\_variant: 49%

missense\_variant: 27%

inframe\_deletion: 9%

inframe\_insertion: 6%

frameshift\_variant: 5%

protein\_altering\_variant: 3%

stop\_gained: 1%

start\_lost: 0%

coding\_sequence\_variant: 0%

Coding consequences:

synonymous\_variant: 49%

missense\_variant: 27%

inframe\_deletion: 9%

inframe\_insertion: 6%

frameshift\_variant: 5%

protein\_altering\_variant: 3%

stop\_gained: 1%

start\_lost: 0%

coding\_sequence\_variant: 0%

{

By checking the website: https://www.cdc.gov/ncbddd/autism/data.html

Identified Prevalence of Autism Spectrum Disorder

ADDM Network 2000-2014 Combining Data from All Sites

The most recent:

Surveillance Year: 2014

Birth Year: 2006

Number of ADDM Sites Reporting: 11

Combined Prevalence per 1,000 Children (Range Across ADDM Sites): 16.8 (13.1-29.3)

This is about 1 in X children…: 1 in 59

I use 1/59 = 0.0169.

Filter the VEP results looking for variants that have a gnomAD\_AF value of LESS than 0.0169 (= 1/59) .

This would remove a lot of the more common polymorphisms.

}

I set a filter to restrict results by clicking the option "Show one selected consequence per variant allele". I used the filter function on the website. I set "gnomAD AF - < - 0.0169, Consequence - is not - synonymous\_variant, and Consequence - is not - intron\_variant". The number of variants decreased from 19,124 to less than 100. I downloaded as the vep.txt file.

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

From the VEP website:

Uploaded variant Location Allele Consequence Impact Symbol Gene Feature type Feature Biotype Exon Intron HGVSc HGVSp cDNA position CDS position Protein position Amino acids Codons Existing variant Distance to transcript Feature strand Symbol source GIVEN REF USED REF BAM EDIT SIFT PolyPhen AF gnomAD AF gnomAD AFR AF gnomAD AMR AF gnomAD ASJ AF gnomAD EAS AF gnomAD FIN AF gnomAD NFE AF gnomAD OTH AF gnomAD SAS AF Clinical significance Somatic status Phenotype or disease Pubmed

. X:153626695-153626695 A 5\_prime\_UTR\_variant MODIFIER RPL10 6134 Transcript NM\_001256577.1 protein\_coding 1/6 - NM\_001256577.1:c.-64G>A - 125 - - - - rs116301817 - 1 EntrezGene G G - - - 0.0479 0.01301 0.1684 0.006181 0 0 0 0.000203 0.005287 0.0004711 - - - -

. X:135092595-135092595 T splice\_region\_variant, intron\_variant LOW SLC9A6 10479 Transcript NM\_001042537.1 protein\_coding - 6/15 NM\_001042537.1:c.900-6C>T - - - - - - rs17001258 - 1 EntrezGene C C OK - - 0.0193 0.004768 0.06247 0.002747 0 0 0 4.992e-05 0.002221 5.223e-05 benign - 1 25741868, 18414213

. X:125299567-125299567 T missense\_variant MODERATE DCAF12L2 340578 Transcript NM\_001013628.2 protein\_coding 1/1 - NM\_001013628.2:c.341G>A NP\_001013650.1:p.Gly114Asp 514 341 114 G/D GGC/GAC rs1442331593, COSM3424442, COSM3424443 - -1 EntrezGene C C OK

0

0.999

- 5.603e-06 0 0 0 0 0 1.252e-05 0 0 - 0, 1, 1 0, 1, 1 -

. X:109561057-109561073 CCGCCGCCGCCGC inframe\_deletion MODERATE AMMECR1 9949 Transcript NM\_015365.2 protein\_coding 1/6 - NM\_015365.2:c.240\_242del NP\_056180.1:p.Gly82del 308-323 227-242 76-81 CGGGGG/CGGGG TGCGGCGGCGGCGGCGGG/TGCGGCGGCGGCGGG rs763207969 - -1 EntrezGene CCGCCGCCGCCGCCGC CCGCCGCCGCCGCCGC OK - - - 0.003486 0.001672 0.002107 0.00423 0.001995 0.005597 0.004468 0.004729 0.002127 - - - -

. X:92928114-92928114 C missense\_variant MODERATE NAP1L3 4675 Transcript NM\_004538.5 protein\_coding 1/1 - NM\_004538.5:c.190A>G NP\_004529.2:p.Ser64Gly 569 190 64 S/G AGC/GGC rs62641622 - -1 EntrezGene T T OK

0.43

0

0.0190 0.00395 0.05532 0.002083 0 0.0001636 0 0.0001702 0.00323 0.0001078 - - - -

. 22:51041768-51041790 GAGGAGGAGGAGGAGGAGG inframe\_deletion MODERATE MAPK8IP2 23542 Transcript NM\_012324.3 protein\_coding 3/12 - NM\_012324.3:c.308\_310del NP\_036456.1:p.Glu103del 406-427 289-310 97-104 EEEEEEEG/EEEEEEG GAGGAGGAGGAGGAGGAGGAGGGA/GAGGAGGAGGAGGAGGAGGGA rs572434194, TMP\_ESP\_22\_51041769\_51041771 - 1 EntrezGene GAGGAGGAGGAGGAGGAGGAGG GAGGAGGAGGAGGAGGAGGAGG - - - 0.0020 0.004492 0.001967 0.00308 0.00405 0.002702 0.008237 0.005012 0.004154 0.004388 - - - -

. 22:50502592-50502614 AGCAGCAGCAGCAGCAGCA inframe\_deletion MODERATE MLC1 23209 Transcript NM\_015166.3 protein\_coding 11/12 - NM\_015166.3:c.927\_929del NP\_055981.1:p.Leu310del 1542-1563 908-929 303-310 VLLLLLLL/VLLLLLL GTGCTGCTGCTGCTGCTGCTGCTA/GTGCTGCTGCTGCTGCTGCTA rs768073887 - -1 EntrezGene AGCAGCAGCAGCAGCAGCAGCA AGCAGCAGCAGCAGCAGCAGCA - - - - 0.0005551 0.0005741 0.0004271 0.0002106 0.0003562 0.001585 0.0006283 0.0001921 0.0001682 - - - -

. 22:42262948-42262969 GCAGCAGCAGCAGCAGCA inframe\_deletion MODERATE SREBF2 6721 Transcript NM\_004599.3 protein\_coding 2/19 - NM\_004599.3:c.221\_223del NP\_004590.2:p.Ser74del 395-415 203-223 68-75 GSSSSSSN/GSSSSSN GGCAGCAGCAGCAGCAGCAGCAAT/GGCAGCAGCAGCAGCAGCAAT rs143615881 - 1 EntrezGene GCAGCAGCAGCAGCAGCAGCA GCAGCAGCAGCAGCAGCAGCA OK - - - 0.0001343 6.68e-05 0 0 0 0.0002804 0.0002253 0 3.289e-05 - - - -

. 22:41573266-41573266 C missense\_variant MODERATE EP300 2033 Transcript NM\_001429.3 protein\_coding 31/31 - NM\_001429.3:c.5551A>C NP\_001420.2:p.Thr1851Pro 5946 5551 1851 T/P ACT/CCT rs199771020 - 1 EntrezGene A A -

0.12

0.276

- 0.0003044 0 0 0 0 0.002117 0.0002651 0.000188 0 - - - -

. 22:28194909-28194914 TG inframe\_deletion MODERATE MN1 4330 Transcript NM\_002430.2 protein\_coding 1/2 - NM\_002430.2:c.1620\_1622del NP\_002421.3:p.Gln550del 2573-2577 1618-1622 540-541 QQ/Q CAACAG/CAG rs757115560 - -1 EntrezGene TGTTG TGTTG OK - - - 0.0004336 0.006846 0.0002748 0 8.518e-05 0 4.396e-05 0.0004864 0.0002392 - - - -

. 20:46279860-46279862 CAACA inframe\_insertion MODERATE NCOA3 8202 Transcript NM\_181659.2 protein\_coding 20/23 - NM\_181659.2:c.3788\_3789insACA NP\_858045.1:p.Gln1276dup 4048-4049 3787-3788 1263 Q/QQ CAG/CAACAG rs753491875 - 1 EntrezGene CA CA OK - - - 0.005401 0.001528 0.006247 0.001775 0.00829 4.64e-05 0.0001981 0.004148 0.02915 - - - -

. 20:35422792-35422792 G missense\_variant MODERATE SOGA1 140710 Transcript NM\_080627.2 protein\_coding 14/15 - NM\_080627.2:c.3693G>C NP\_542194.2:p.Gln1231His 4033 3693 1231 Q/H CAG/CAC rs34459518 - -1 EntrezGene C C -

0.72

0

0.0280 0.00662 0.09593 0.003982 0 5.84e-05 0 0.0002363 0.002959 9.781e-05 - - - -

. 20:32664866-32664866 A missense\_variant MODERATE RALY 22913 Transcript NM\_016732.2 protein\_coding 8/10 - NM\_016732.2:c.691G>A NP\_057951.1:p.Gly231Ser 1263 691 231 G/S GGC/AGC rs11538302 - 1 EntrezGene G G OK

0.74

0

0.2512 0.01196 0.03719 0.004714 0.007229 0.04632 0.00813 0.008807 0.01257 0.003907 - - - -

. 20:3765807-3765807 C missense\_variant MODERATE CENPB 1059 Transcript NM\_001810.5 protein\_coding 1/1 - NM\_001810.5:c.1324T>G NP\_001801.1:p.Leu442Val 1531 1324 442 L/V TTG/GTG rs1480653341, COSM5956546 - -1 EntrezGene A A OK

0.5

0

- 0.01043 0.01152 0.01514 0.008424 0.02653 0.00102 0.008319 0.01564 0.01373 - 0, 1 0, 1 -

. 19:50367486-50367486 T missense\_variant MODERATE PNKP 11284 Transcript XM\_005258474.1 protein\_coding 6/16 - XM\_005258474.1:c.586T>A XP\_005258531.1:p.Tyr196Asn 714 586 196 Y/N TAC/AAC rs3739186 - -1 EntrezGene A A - - - 0.0122 0.002538 0.03411 0.002025 0 0 0 0.000188 0.001823 0.0001299 benign, likely\_benign - 1 18414213, 20445776, 17055652

. 19:50040307-50040307 A missense\_variant MODERATE RCN3 57333 Transcript NM\_020650.2 protein\_coding 4/7 - NM\_020650.2:c.463G>A NP\_065701.2:p.Val155Met 910 463 155 V/M GTG/ATG rs77227069 - 1 EntrezGene G G OK

0.11

0.154

0.0122 0.00322 0.04249 0.001808 0 7.005e-05 0 0.0001766 0.002289 4.226e-05 - - - -

. 19:49656688-49656688 T missense\_variant MODERATE HRC 3270 Transcript NM\_002152.2 protein\_coding 1/6 - NM\_002152.2:c.1807G>A NP\_002143.1:p.Glu603Lys 1994 1807 603 E/K GAG/AAG rs61732829 - -1 EntrezGene C C OK

0

0.972

0.0294 0.006418 0.08967 0.003913 0 5.801e-05 0 0.0002487 0.003133 0.0001651 - - - -

. 19:3747910-3747910 A missense\_variant MODERATE TJP3 27134 Transcript XM\_005259539.1 protein\_coding 18/20 - XM\_005259539.1:c.2498G>A XP\_005259596.1:p.Gly833Asp 2517 2498 833 G/D GGC/GAC rs10408494, COSM6875124 - 1 EntrezGene G G -

0.04

0.367

0.0487 0.01139 0.1588 0.008266 0.00255 0 0 0.0004345 0.006431 0.0001951 - 0, 1 0, 1 -

. 18:9887041-9887041 C missense\_variant MODERATE TXNDC2 84203 Transcript NM\_001098529.1 protein\_coding 2/2 - NM\_001098529.1:c.565T>C NP\_001091999.1:p.Ser189Pro 1014 565 189 S/P TCC/CCC rs749249191, COSM6286428, COSM6286429 - 1 EntrezGene T T OK

1

0.003

- 4.169e-05 0 0 0 0 0 9.828e-06 0.0002169 0.0002589 - 0, 1, 1 0, 1, 1 -

. 17:46669562-46669565 GG 3\_prime\_UTR\_variant MODIFIER HOXB5 3215 Transcript NM\_002147.3 protein\_coding 2/2 - NM\_002147.3:c.\*8del - 875-877 - - - - rs150874789 - -1 EntrezGene GGG GGG OK - - 0.0064 0.001192 0.01751 0.0005362 0 0 0 5.406e-05 0.0003658 6.509e-05 - - - -

. 17:37321347-37321347 A missense\_variant MODERATE ARL5C 390790 Transcript NM\_001143968.1 protein\_coding 2/6 - NM\_001143968.1:c.92C>T NP\_001137440.1:p.Thr31Ile 493 92 31 T/I ACC/ATC rs9912267 - -1 EntrezGene G G -

0

1

0.0196 0.003494 0.05941 0.002338 0 0 0 0.0001209 0.002067 0.0002629 - - - -

. 16:3929941-3929941 T 5\_prime\_UTR\_variant MODIFIER CREBBP 1387 Transcript NM\_004380.2 protein\_coding 1/31 - NM\_004380.2:c.-24G>A - 181 - - - - rs28407999 - -1 EntrezGene C C - - - 0.0052 0.00113 0.01714 0.0008545 0 0 0 0 0.0003855 0 - - - -

. 16:2229752-2229752 T missense\_variant MODERATE CASKIN1 57524 Transcript NM\_020764.3 protein\_coding 18/20 - NM\_020764.3:c.3617C>A NP\_065815.1:p.Ala1206Glu 3649 3617 1206 A/E GCG/GAG rs374957682 - -1 EntrezGene G G -

0.71

0.001

- 6.351e-06 0 0 0 0 0 1.552e-05 0 0 - - - -

. 15:74888024-74888040 CAGCAGCAGCAGC inframe\_deletion MODERATE ARID3B 10620 Transcript NM\_001307939.1 protein\_coding 9/9 - NM\_001307939.1:c.1609\_1611del NP\_001294868.1:p.Ser538del 1827-1842 1596-1611 532-537 ASSSSS/ASSSS GCCAGCAGCAGCAGCAGC/GCCAGCAGCAGCAGC rs749888609 - 1 EntrezGene CAGCAGCAGCAGCAGC CAGCAGCAGCAGCAGC OK - - - 0.000168 7.043e-05 0.0001312 0.0002258 0.0003156 0 0.0001945 0 0.0002162 - - - -

. 14:51383717-51383717 T missense\_variant MODERATE PYGL 5836 Transcript NM\_002863.4 protein\_coding 8/20 - NM\_002863.4:c.962G>A NP\_002854.3:p.Arg321His 1089 962 321 R/H CGT/CAT rs116465563, COSM1222828 - -1 EntrezGene C C OK

0.04

0.001

0.0084 0.00197 0.02973 0.0006849 0 0 0 1.791e-05 0.0007291 3.249e-05 uncertain\_significance, benign 0, 1 1, 1 25741868

. 14:45665690-45665690 T missense\_variant MODERATE FANCM 57697 Transcript NM\_020937.2 protein\_coding 21/23 - NM\_020937.2:c.5656C>T NP\_065988.1:p.His1886Tyr 5755 5656 1886 H/Y CAC/TAC rs79343837 - 1 EntrezGene C C -

0.43

0

0.0100 0.002157 0.03 0.00137 0 0 0 0.0001792 0.001095 0 benign, likely\_benign - 1 25741868

. 14:23574095-23574128 GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA downstream\_gene\_variant MODIFIER C14orf119 55017 Transcript NM\_017924.3 protein\_coding - - - - - - - - - rs781677359 4431 1 EntrezGene GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA GCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA OK - - - 0.002423 0.006929 0.004072 0.001988 0.00234 0 0.001082 0.003671 0.001674 - - - -

. 14:23573709-23573709 T downstream\_gene\_variant MODIFIER C14orf119 55017 Transcript NM\_017924.3 protein\_coding - - - - - - - - - rs141661506 4044 1 EntrezGene C C OK - - 0.0038 0.0007629 0.0126 0.0005894 0 0 0 0 0.001106 0.000133 - - - -

. 14:23549948-23549948 T missense\_variant MODERATE ACIN1 22985 Transcript NM\_014977.3 protein\_coding 6/19 - NM\_014977.3:c.770G>A NP\_055792.1:p.Arg257Lys 1098 770 257 R/K AGA/AAA rs11555803 - -1 EntrezGene C C OK

0

0.995

0.0228 0.005872 0.08025 0.00408 0.0001015 0 0 0.000357 0.003114 0.0002925 - - - -

. 12:63544601-63544601 T missense\_variant MODERATE AVPR1A 552 Transcript XM\_005269002.1 protein\_coding 2/3 - XM\_005269002.1:c.25G>A XP\_005269059.1:p.Gly9Ser 3057 25 9 G/S GGT/AGT rs2228154 - -1 EntrezGene C C - - - 0.0593 0.01276 0.1828 0.009233 0 0 8.325e-05 0.0006444 0.004312 0.0002024 - - - -

. 12:49426878-49426878 T missense\_variant MODERATE KMT2D 8085 Transcript NM\_003482.3 protein\_coding 39/54 - NM\_003482.3:c.11610G>A NP\_003473.3:p.Met3870Ile 11610 11610 3870 M/I ATG/ATA rs73302195, COSM6494163 - -1 EntrezGene C C OK

0.01

0

0.0192 0.003141 0.04933 0.002546 0 0 0.0007248 0.0001046 0.003132 4.382e-05 benign 0, 1 1, 1 18414213, 24728327

. 12:9085324-9085342 GCAGCAGCAGCA inframe\_deletion MODERATE PHC1 1911 Transcript NM\_004426.2 protein\_coding 8/15 - NM\_004426.2:c.1284\_1289del NP\_004417.2:p.Gln438\_Gln439del 1428-1445 1272-1289 424-430 AQQQQQQ/AQQQQ GCGCAGCAGCAGCAGCAGCAA/GCGCAGCAGCAGCAA rs368945524, TMP\_ESP\_12\_9085325\_9085330 - 1 EntrezGene GCAGCAGCAGCAGCAGCA GCAGCAGCAGCAGCAGCA OK - - 0.0172 0.003794 0.05054 0.004644 0 0 0 0.0002276 0.002208 0.0001632 benign - - 25741868

. 12:7045885-7045899 CAGCA inframe\_deletion MODERATE ATN1 1822 Transcript NM\_001007026.1 protein\_coding 5/10 - NM\_001007026.1:c.1461\_1469del NP\_001007027.1:p.Gln500\_Gln502del 1693-1706 1456-1469 486-490 QQQQQ/QQ CAGCAACAGCAGCAG/CAGCAG rs782073335, TMP\_ESP\_12\_7045886\_7045900 - 1 EntrezGene CAGCAACAGCAGCA CAGCAACAGCAGCA OK - - - 3.52e-05 0 3.341e-05 0.0002211 0.0001374 0 1.874e-05 0 3.597e-05 - - - -

. 12:2062350-2062356 TGGTGGTGG protein\_altering\_variant MODERATE DCP1B 196513 Transcript NM\_152640.3 protein\_coding 7/9 - NM\_152640.3:c.755\_756insCCA NP\_689853.3:p.His251dup 830-835 750-755 250-252 LHQ/LHHQ CTCCACCAG/CTCCACCACCAG rs745491834 - -1 EntrezGene TGGTGG TGGTGG - - - - 0.0001831 0.0001325 2.989e-05 0 0.001225 0.0002039 3.667e-05 0.0001852 0.0003586 - - - -

. 12:2062323-2062352 TGCTGCTGCTGCTGCTGCTGCTG inframe\_deletion MODERATE DCP1B 196513 Transcript NM\_152640.3 protein\_coding 7/9 - NM\_152640.3:c.777\_782del NP\_689853.3:p.Gln260\_Gln261del 834-862 754-782 252-261 QQQQQQQQQQ/QQQQQQQQ CAGCAGCAGCAGCAGCAGCAGCAGCAGCAA/CAGCAGCAGCAGCAGCAGCAGCAA rs748841037 - -1 EntrezGene TGCTGCTGCTGCTGCTGCTGCTGCTGCTG TGCTGCTGCTGCTGCTGCTGCTGCTGCTG - - - - 0.0001423 0.0001489 9.8e-05 0 0.0003898 0 0.0002023 0 0 - - - -

. 11:116729161-116729161 C missense\_variant MODERATE SIK3 23387 Transcript XM\_005271481.1 protein\_coding 21/25 - XM\_005271481.1:c.3020A>G XP\_005271538.1:p.Tyr1007Cys 3020 3020 1007 Y/C TAT/TGT rs55730930 - -1 EntrezGene T T - - - 0.0020 0.003479 0.001047 0.003068 0.001512 0.001973 0.0006879 0.005776 0.00299 0.000341 - - - -

. 11:113857661-113857661 A missense\_variant MODERATE HTR3A 3359 Transcript NM\_213621.3 protein\_coding 7/8 - NM\_213621.3:c.1145G>A NP\_998786.2:p.Arg382His 1378 1145 382 R/H CGT/CAT rs35815285, CM011786 - 1 EntrezGene G G OK

0.04

0.681

0.0343 0.007659 0.1013 0.006373 0.0008123 5.799e-05 0 0.0006451 0.005835 0.0002599 - - 0, 1 -

. 11:71906793-71906793 C splice\_donor\_variant HIGH FOLR1 2348 Transcript NM\_016725.2 protein\_coding - 4/4 NM\_016725.2:c.493+2T>C - - - - - - rs144637717, CS118196 - 1 EntrezGene T T OK - - 0.0034 0.003359 0.0005227 0.00137 0.006195 5.798e-05 0.000583 0.002177 0.003829 0.0141 uncertain\_significance - 1, 1 25741868, 23757202

. 11:65632076-65632076 C missense\_variant MODERATE MUS81 80198 Transcript XM\_005274307.1 protein\_coding 11/16 - XM\_005274307.1:c.1171A>C XP\_005274364.1:p.Asn391His 1520 1171 391 N/H AAC/CAC rs115472389 - 1 EntrezGene A A - - - 0.0160 0.004055 0.05959 0.002144 0 0 0 5.373e-05 0.0007294 0.0001299 - - - -

. 11:16205506-16205506 G splice\_region\_variant, intron\_variant LOW SOX6 55553 Transcript NM\_001145819.1 protein\_coding - 5/15 NM\_001145819.1:c.748-6T>C - - - - - - rs144474428 - -1 EntrezGene A A - - - 0.0018 0.0005428 0.007726 0.0003886 0 0 0 0 0 6.504e-05 - - - -

. 8:110656909-110656909 T 5\_prime\_UTR\_variant MODIFIER SYBU 55638 Transcript NM\_001099745.1 protein\_coding 2/8 - NM\_001099745.1:c.-21G>A - 332 - - - - rs772441845 - -1 EntrezGene C C OK - - - 1.992e-05 0 0 0 0 0.0001889 0 0 5.265e-05 - - - -

. 7:154684363-154684363 T 3\_prime\_UTR\_variant MODIFIER DPP6 1804 Transcript NM\_130797.3 protein\_coding 26/26 - NM\_130797.3:c.\*173C>T - 3174 - - - - rs76896340 - 1 EntrezGene C C - - - 0.0341 0.007137 0.09708 0.004775 0 0 0 0.0002845 0.005128 0.0003664 - - - -

. 7:116502628-116502628 C 5\_prime\_UTR\_variant MODIFIER CAPZA2 830 Transcript NM\_006136.2 protein\_coding 1/10 - NM\_006136.2:c.-38T>C - 66 - - - - rs192510686 - 1 EntrezGene T T OK - - 0.0639 0.00859 0.1944 0.009646 0 0 0 0.0004272 0.008382 0.0002924 - - - -

. 7:45905737-45905758 TGCTGCTGCTGCTGCTGC intergenic\_variant MODIFIER - - - - - - - - - - - - - - rs770987587 - - - - - - - - - 0.004316 0.01579 0.002283 0.002439 0.005319 0.01181 0.003706 0.003846 0.004252 - - - -

. 7:6661739-6661757 AGCAGCAGCAGCAGC inframe\_deletion MODERATE ZNF853 54753 Transcript NM\_017560.1 protein\_coding 3/3 - NM\_017560.1:c.1133\_1135del NP\_060030.1:p.Gln378del 1397-1414 1118-1135 373-379 EQQQQQL/EQQQQL GAGCAGCAGCAGCAGCAGCTG/GAGCAGCAGCAGCAGCTG rs767304139 - 1 EntrezGene AGCAGCAGCAGCAGCAGC AGCAGCAGCAGCAGCAGC - - - - 0.0001382 0.0002936 0 0 0 0.001041 7.74e-05 0 4.456e-05 - - - -

. 6:157099402-157099449 CAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGCAGCAGCA inframe\_insertion MODERATE ARID1B 57492 Transcript XM\_005267069.1 protein\_coding 1/20 - XM\_005267069.1:c.443\_444insGCAGCA XP\_005267126.1:p.Gln157\_Gln158dup 424-470 421-467 141-156 QQQQQQQQQQQQQQQQ/QQQQQQQQQQQQQQQQQQ CAGCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGCAGCAGCAA/CAGCAGCAGCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGCAGCAGCAA rs770869529, rs587779743, TMP\_ESP\_6\_157099403\_157099405 - 1 EntrezGene CAGCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGCAGCAGCA CAGCAGCAGCAGCAGCAGCAGCAACAGCAGCAGCAGCAGCAGCAGCA - - - - 0.003938 0.05508 0.00392 0 0.0005683 0.0001044 0.000876 0.003576 0.000562 benign, likely\_benign - - 25741868

. 6:42897357-42897384 TGCTGCTGCTGCTGCTGCTGCTGC inframe\_deletion MODERATE CNPY3 10695 Transcript NM\_001318842.1 protein\_coding 1/7 - NM\_001318842.1:c.74\_76del NP\_001305771.1:p.Leu25del 491-517 50-76 17-26 LLLLLLLLLP/LLLLLLLLP TTGCTGCTGCTGCTGCTGCTGCTGCTGCCG/TTGCTGCTGCTGCTGCTGCTGCTGCCG rs780206711 - 1 EntrezGene TGCTGCTGCTGCTGCTGCTGCTGCTGC TGCTGCTGCTGCTGCTGCTGCTGCTGC OK - - - 0.006573 0.007997 0.0042 0.005101 0.006626 0.01023 0.007486 0.006124 0.004697 - - - -

. 6:16327906-16327911 TG inframe\_deletion MODERATE ATXN1 6310 Transcript XM\_005249287.1 protein\_coding 6/7 - XM\_005249287.1:c.633\_635del XP\_005249344.1:p.His211del 1407-1411 631-635 211-212 HQ/Q CATCAG/CAG rs770305044, TMP\_ESP\_6\_16327907\_16327915 - -1 EntrezGene TGATG TGATG - - - - 0.006684 0.01706 0.007278 0.001162 0.02027 0.00363 0.005152 0.006052 0.00593 uncertain\_significance - - 25741868

. 6:16327903-16327905 TGATG inframe\_insertion MODERATE ATXN1 6310 Transcript XM\_005249287.1 protein\_coding 6/7 - XM\_005249287.1:c.638\_639insTCA XP\_005249344.1:p.Gln212\_Gln213insHis 1413-1414 637-638 213 Q/HQ CAG/CATCAG rs780549091 - -1 EntrezGene TG TG - - - - 0.01591 0.02119 0.007344 0.01184 0.0285 0.01286 0.01478 0.01627 0.02608 benign - 1 -

. 6:16327897-16327897 A missense\_variant MODERATE ATXN1 6310 Transcript XM\_005249287.1 protein\_coding 6/7 - XM\_005249287.1:c.645G>T XP\_005249344.1:p.Gln215His 1421 645 215 Q/H CAG/CAT rs184327938 - -1 EntrezGene C C -

0.15

0

- 0.008306 0.003497 0.005155 0.001193 0.09345 0.003778 0.005072 0.007778 0.007714 likely\_benign - - -

. 5:128864375-128864375 G splice\_region\_variant, intron\_variant LOW ADAMTS19 171019 Transcript NM\_133638.4 protein\_coding - 6/22 NM\_133638.4:c.1328+5A>G - - - - - - rs73246875 - 1 EntrezGene A A - - - 0.0397 0.007866 0.1192 0.005782 0.0005382 0 0 0.0003649 0.003067 3.854e-05 - - - -

. 4:56304529-56304548 CTGCTGCTGCTGCTGC inframe\_deletion MODERATE CLOCK 9575 Transcript NM\_004898.3 protein\_coding 22/23 - NM\_004898.3:c.2278\_2280del NP\_004889.1:p.Gln760del 2679-2697 2262-2280 754-760 TQQQQQQ/TQQQQQ ACGCAGCAGCAGCAGCAGCAG/ACGCAGCAGCAGCAGCAG rs774776137 - -1 EntrezGene CTGCTGCTGCTGCTGCTGC CTGCTGCTGCTGCTGCTGC - - - - 0.0001383 0 3.149e-05 0 0 0.0001397 0.00026 0.000193 0 - - - -

. 3:65425585-65425608 TGCTGCTGCTGTTGCTG inframe\_deletion MODERATE MAGI1 9223 Transcript XM\_005265563.1 protein\_coding 9/24 - XM\_005265563.1:c.1239\_1244del XP\_005265620.1:p.Gln422\_Gln423del 1750-1772 1222-1244 408-415 QQQQQQQQ/QQQQQQ CAGCAACAGCAGCAGCAACAGCAG/CAGCAACAGCAGCAGCAG rs374381483 - -1 EntrezGene TGCTGTTGCTGCTGCTGTTGCTG TGCTGTTGCTGCTGCTGTTGCTG - - - - 4.147e-06 0 0 0 0 0 0 0 3.26e-05 - - - -

. 3:42750471-42750471 A 3\_prime\_UTR\_variant MODIFIER CCDC13 152206 Transcript NM\_144719.3 protein\_coding 16/16 - NM\_144719.3:c.\*1G>T - 2233 - - - - rs751529017 - -1 EntrezGene C C OK - - - 4.14e-06 0 0 0 0 0 9.141e-06 0 0 - - - -

. 3:14106332-14106332 C non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.1111G>C - 1111 - - - - rs113667859 - 1 EntrezGene G G OK - - - 0.001484 0.0003399 0.001851 0.0006944 0.001539 0.00463 0.001487 0 0.0009111 - - - -

. 3:14106063-14106063 C non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.842T>C - 842 - - - - rs746073387 - 1 EntrezGene T T OK - - - 0.0002058 0.0005774 0 0 0.0009524 0 0.0001272 0 0.0005245 - - - -

. 3:14106033-14106033 C non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.812T>C - 812 - - - - rs752066220 - 1 EntrezGene T T OK - - - 0.005759 0.1525 0.004527 0.0009163 0.0006464 0.0005509 0.0009847 0.001733 0.0003531 - - - -

. 3:14106003-14106003 C non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.782T>C - 782 - - - - rs752648109 - 1 EntrezGene T T OK - - - 6.545e-05 0 0 0 0 0 0.0001728 0 0 - - - -

. 3:14105998-14105998 A non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.777T>A - 777 - - - - rs758834646 - 1 EntrezGene T T OK - - - 8.389e-06 0 0 0 0 0 2.181e-05 0 0 - - - -

. 3:14105982-14105982 AG non\_coding\_transcript\_exon\_variant MODIFIER TPRXL 348825 Transcript NR\_002223.3 pseudogene 3/3 - NR\_002223.3:n.761\_762insAG - 761-762 - - - - rs767319826 - 1 EntrezGene - - OK - - - 0.0005816 0 0.0002104 0.0003746 0 0.001103 0.0009834 0 0.0004862 - - - -

. 2:230456456-230456456 A missense\_variant MODERATE DNER 92737 Transcript NM\_139072.3 protein\_coding 2/13 - NM\_139072.3:c.425C>T NP\_620711.3:p.Thr142Ile 572 425 142 T/I ACT/ATT rs147533391, COSM3047999 - -1 EntrezGene G G -

0.01

0.029

0.0098 0.002575 0.03657 0.001936 0 0 0 8.979e-06 0.001094 6.498e-05 - 0, 1 0, 1 -

. 2:227661395-227661418 TGCTGCTGCTGCTGCTGCTG inframe\_deletion MODERATE IRS1 3667 Transcript NM\_005544.2 protein\_coding 1/2 - NM\_005544.2:c.2057\_2059del NP\_005535.1:p.Ser686del 2089-2111 2037-2059 679-687 PSSSSSSSN/PSSSSSSN CCCAGCAGCAGCAGCAGCAGCAGCAAC/CCCAGCAGCAGCAGCAGCAGCAAC rs138975702, TMP\_ESP\_2\_227661396\_227661401 - -1 EntrezGene TGCTGCTGCTGCTGCTGCTGCTG TGCTGCTGCTGCTGCTGCTGCTG OK - - 0.0140 0.005295 0.02573 0.003733 0.006066 0.0003718 0.002401 0.005216 0.006576 0.001385 likely\_benign - 1, 0 -

. 2:185803766-185803766 C 3\_prime\_UTR\_variant MODIFIER ZNF804A 91752 Transcript NM\_194250.1 protein\_coding 4/4 - NM\_194250.1:c.\*13T>C - 4237 - - - - rs142431760 - 1 EntrezGene T T OK - - 0.0010 0.0002063 0.003082 3.017e-05 0 0 0 1.826e-05 0 0 - - - -

. 2:166854631-166854631 C missense\_variant MODERATE SCN1A 6323 Transcript NM\_001202435.1 protein\_coding 25/28 - NM\_001202435.1:c.4393A>G NP\_001189364.1:p.Ile1465Val 4620 4393 1465 I/V ATT/GTT rs138231868 - -1 EntrezGene T T -

0.22

0.124

0.0018 0.0002661 0.003888 0.0001503 0 0 0 9.041e-06 0 0 benign, likely\_benign - 1 25741868

. 2:145156548-145156548 C missense\_variant MODERATE ZEB2 9839 Transcript NM\_014795.3 protein\_coding 8/10 - NM\_014795.3:c.2206A>G NP\_055610.1:p.Met736Val 2728 2206 736 M/V ATG/GTG rs139191491 - -1 EntrezGene T T -

0.3

0

0.0006 0.0002194 0.001961 0.0005063 0 0 0 5.378e-05 0.0001824 0 - - - -

. 2:131521770-131521770 T missense\_variant MODERATE AMER3 205147 Transcript NM\_152698.2 protein\_coding 2/2 - NM\_152698.2:c.2125C>T NP\_689911.2:p.Arg709Cys 2315 2125 709 R/C CGC/TGC rs759025770, COSM3938502 - 1 EntrezGene C C -

0.03

0

- 1.244e-05 0 0 0 0 0 1.829e-05 0 3.351e-05 - 0, 1 0, 1 -

. 2:115200423-115200423 T splice\_region\_variant, intron\_variant LOW DPP10 57628 Transcript NM\_020868.3 protein\_coding - 1/25 NM\_020868.3:c.60+8C>T - - - - - - rs7560788 - 1 EntrezGene C C - - - 0.0292 0.006873 0.09386 0.003917 0.003759 0 0 0.0004691 0.004223 0.0002602 - - - -

. 2:27258525-27258525 T missense\_variant MODERATE TMEM214 54867 Transcript XM\_005264381.1 protein\_coding 4/18 - XM\_005264381.1:c.566C>T XP\_005264438.1:p.Ala189Val 648 566 189 A/V GCA/GTA rs202006590 - 1 EntrezGene C C - - - - 8.122e-06 0 0 0 0 0 1.79e-05 0 0 - - - -

. 1:223536702-223536724 TGCTGCTGCTGCTGCTGCTGCTGCT inframe\_insertion MODERATE SUSD4 55061 Transcript XM\_005273169.1 protein\_coding 3/10 - XM\_005273169.1:c.65\_66insGCA XP\_005273226.1:p.Gln22dup 242-263 44-65 15-22 EQQQQQQQ/EQQQQQQQQ GAGCAGCAGCAGCAGCAGCAGCAA/GAGCAGCAGCAGCAGCAGCAGCAGCAA rs371162328, TMP\_ESP\_1\_223536703\_223536705 - -1 EntrezGene TGCTGCTGCTGCTGCTGCTGCT TGCTGCTGCTGCTGCTGCTGCT - - - - 0.005693 0.07158 0.005776 0 0.005109 0 0.0004356 0.005424 0.0003665 - - - -

. 1:209605636-209605676 AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA inframe\_insertion MODERATE MIR205HG 642587 Transcript NM\_001104548.1 protein\_coding 4/4 - NM\_001104548.1:c.291\_292insGCAGCA NP\_001098018.1:p.Ala96\_Ala97dup 635-674 252-291 84-97 VAAAAAAAAAAAAA/VAAAAAAAAAAAAAAA GTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA/GTAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA rs565985624 - 1 EntrezGene AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCA - - - - 0.00176 0.02243 0.002401 0.0007367 0.0003574 0 0.0001348 0.0009124 0.0004261 - - - 27885248

. 1:53547797-53547802 GA inframe\_deletion MODERATE PODN 127435 Transcript NM\_153703.4 protein\_coding 10/11 - NM\_153703.4:c.1953\_1955del NP\_714914.2:p.Glu659del 2119-2123 1951-1955 651-652 EE/E GAAGAG/GAG rs371150672 - 1 EntrezGene GAAGA GAAGA OK - - 0.0008 0.001123 0.003448 0.00381 0 0.001413 0.0001353 0.0001263 0.001302 0.00173 - - - -

. 1:20669084-20669084 G missense\_variant MODERATE VWA5B1 127731 Transcript NM\_001039500.2 protein\_coding 15/22 - NM\_001039500.2:c.2293C>G NP\_001034589.2:p.Arg765Gly 2489 2293 765 R/G CGA/GGA rs74056519 - 1 EntrezGene C C -

0.13

0.046

0.0184 0.004745 0.05934 0.00326 0 0 0 0.000393 0.00285 0.0002434 - - - -

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Compared with the genelist.txt file in the lab3 data folder, I found the following genes:

DPP10

ZEB2

SCN1A

ZNF804A

DPP6

FOLR1

AVPR1A

CREBBP

SLC9A6

RPL10

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

I select five variants at random and searched in OMIM to see if the genes are listed there. Then I checked in ClinVar to see if the variants are listed.

It doesn't show that DPP10 has a relationship with autism in OMIM, but there are articles in ncbi showing there are relationship between DPP10 and autism spectrum disorder - implications of a copy number variation involving DPP10.

Results: The genes I choose to keep working on deep research are DPP10, ZEB2, SCN1A, CREBBP, AND RPL10.

-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

DPP10:

According to OMIM, the DPP10 doesn't show to have relationships with autism. In ClinVar, it shows a relationship connected the gene to autism.

\* 608209

DIPEPTIDYL PEPTIDASE X; DPP10

Alternative titles; symbols

DIPEPTIDYL PEPTIDASE IV-RELATED PROTEIN 3; DPRP3

KIAA1492

HGNC Approved Gene Symbol: DPP10

Cytogenetic location: 2q14.1 Genomic coordinates (GRCh38): 2:114,442,357-115,845,751 (from NCBI)

TEXT

▼ Cloning and Expression

By sequencing clones obtained from a size-fractionated fetal brain cDNA library, Nagase et al. (2000) cloned DPP10, which they designated KIAA1492. The deduced 711-amino acid protein shares 52% identity with DPP6 (126141). RT-PCR ELISA detected low expression in whole adult brain and all individual brain regions examined except cerebellum, which did not express DPP10. Little to no expression was detected in all other tissues examined, including fetal brain.

By searching an EST database for sequences similar to DPP4 (102720), followed by 5-prime and 3-prime RACE and PCR of a hypothalamus cDNA library, Qi et al. (2003) cloned DPP10. The deduced 796-amino acid protein contains a transmembrane domain, 10 N-glycosylation sites, and several conserved amino acids found in the 6 domains characteristic of members of the peptidase, lipase, esterase, epoxide hydrolase, or serine hydrolase (PLEES) superfamily. However, DPP10 lacks the active-site serine, which is substituted with a glycine residue. Database analysis suggested the presence of a second DPP10 transcript. DPP10 shares 48% and 51% amino acid identity with the short and long DPP6 isoforms, respectively. Northern blot analysis of several human tissues detected a major 3.5-kb DPP10 transcript expressed only in brain and pancreas. Transcripts of 5.0 and 7.5 kb were also detected in brain. ESTs representing DPP10 mRNA were abundant in tissues derived from multiple sclerosis (126200) lesions, retinoblastoma (180200), hypothalamus, hippocampus, and whole brain, with few transcripts found in uterus, colon, and various tumors. Analysis of mouse ESTs indicated that mouse Dpp10 was expressed in several brain regions and retina. DPP10 was recovered in the membrane fraction of transfected cells.

▼ Gene Function

Qi et al. (2003) confirmed that DPP10 does not possess dipeptidyl peptidase activity due to the lack of a critical active-site serine.

▼ Gene Structure

Qi et al. (2003) determined that the DPP10 gene contains at least 23 exons.

▼ Mapping

By genomic sequence analysis, Qi et al. (2003) mapped the DPP10 gene to chromosome 2q12.3-q14.2.

Allen et al. (2003) pointed out that 4 separate reports had linked asthma and related phenotypes to an ill-defined interval between 2q14 and 2q32, and that 2 mouse genome screens linked bronchial hyperresponsiveness to the region homologous to 2q14. They found and replicated association between asthma and the D2S308 microsatellite marker, 800 kb distal to the IL1 (IL1A; 147760) cluster on 2q14. Sequencing the surrounding region, they constructed a comprehensive, high-density, single-nucleotide polymorphism (SNP) linkage disequilibrium map. They found SNP association limited to the initial exons of the DPP10 gene. DPP10 encodes a homolog of dipeptidylpeptidases that cleave terminal dipeptides from cytokines and chemokines, and presents a potential new target for asthma therapy.

By searching DPP10 from ncbi.nlm.nih.gov Clinvar section, I found the article

Use of clinical chromosomal microarray in Chinese patients with autism spectrum disorder-implications of a copy number variation involving DPP10.

Mol Autism. 2017 Jun 26;8:31. doi: 10.1186/s13229-017-0136-x. eCollection 2017.

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

SCN1A:

According to OMIM, the SCN1A gene shows to have a relationships with Autism. In ClinVar, it doesn't show a relationship connected the gene to autism.

\* 182389

SODIUM CHANNEL, NEURONAL TYPE I, ALPHA SUBUNIT; SCN1A

Alternative titles; symbols

SODIUM CHANNEL, BRAIN TYPE I, ALPHA SUBUNIT; NAC1

NAV1.1

Description

The vertebrate sodium channel is a voltage-gated ion channel essential for the generation and propagation of action potentials, chiefly in nerve and muscle. Voltage-sensitive sodium channels are heteromeric complexes consisting of a large central pore-forming glycosylated alpha subunit and 2 smaller auxiliary beta subunits. Functional studies have indicated that the transmembrane alpha subunit of the brain sodium channels is sufficient for expression of functional sodium channels

Animal Model:

Han et al. (2012) reported that mice with Scn1a haploinsufficiency exhibit hyperactivity, stereotyped behaviors, social interaction deficits, and impaired context-dependent spatial memory. Olfactory sensitivity is retained, but novel food odors and social odors are aversive to Scn1a +/- mice. GABAergic neurotransmission is specifically impaired by this mutation, and selective deletion of Na(v)1.1 channels in forebrain interneurons is sufficient to cause these behavioral and cognitive impairments. Remarkably, treatment with low-dose clonazepam, a positive allosteric modulator of GABA(A) receptors, completely rescued the abnormal social behaviors and deficits in fear memory in the mouse model of Dravet syndrome (607208), demonstrating that they are caused by impaired GABAergic neurotransmission and not by neuronal damage from recurrent seizures. Han et al. (2012) concluded that their results demonstrated a critical role for Na(v)1.1 channels in neuropsychiatric functions and provided a potential therapeutic strategy for cognitive deficit and autism spectrum behaviors in Dravet syndrome.

Keywords: autism spectrum behaviors

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

ZEB2:

It doesn't show there would be a relationship between ZEB2 gene and autism in OMIM and ClinVar.

\* 605802

ZINC FINGER E BOX-BINDING HOMEOBOX 2; ZEB2

Alternative titles; symbols

ZINC FINGER HOMEOBOX 1B; ZFHX1B

SMAD-INTERACTING PROTEIN 1; SMADIP1

SIP1

KIAA0569

TEXT

▼ Description

The ZEB2 gene is a member of the ZEB1 (189909)/Drosophila Zfh1 family of 2-handed zinc finger/homeodomain proteins and functions as a DNA-binding transcriptional repressor that interacts with activated SMADs (see 601595), the transducers of TGF-beta (190180) signaling, and interacts with the nucleosome remodeling and histone deacetylation (NURD) complex (Verstappen et al., 2008).

ClinVar results:

Gene ID: 9839, updated on 19-Mar-2019

ZEB2 zinc finger E-box binding homeobox 2 [ Homo sapiens (human) ]

Official Symbol:ZEB2provided by HGNC

Official Full Name: zinc finger E-box binding homeobox 2

Primary source: HGNC:HGNC:14881

See related: Ensembl:ENSG00000169554 MIM:605802

Gene type: protein coding

RefSeq status: REVIEWED

Organism: Homo sapiens

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as: SIP1; SIP-1; ZFHX1B; HSPC082; SMADIP1

Summary: The protein encoded by this gene is a member of the Zfh1 family of 2-handed zinc finger/homeodomain proteins. It is located in the nucleus and functions as a DNA-binding transcriptional repressor that interacts with activated SMADs. Mutations in this gene are associated with Hirschsprung disease/Mowat-Wilson syndrome. Alternatively spliced transcript variants have been found for this gene.[provided by RefSeq, Jan 2010]

Expression: Ubiquitous expression in brain (RPKM 13.4), appendix (RPKM 10.2) and 23 other tissues

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

CREBBP:

It doesn't show a relationship between CREBBP and autism in both the OMIM and the ClinVar.

\* 600140

CREB-BINDING PROTEIN; CREBBP

Alternative titles; symbols

CBP

Other entities represented in this entry:

CBP/MOZ FUSION GENE, INCLUDED

TEXT

▼ Cloning and Expression

When cellular levels of cAMP increase, a cascade of events leads to the induction of genes that contain cis-regulatory elements called cAMP-response elements (CREs). Elevated cAMP levels cause stimulation and nuclear translocation of protein kinase A (PKA; see 176911), which activates the transcription factor CREB (CRE-binding protein; see 123810) by phosphorylating it at a single residue, serine-133 (Gonzalez and Montminy, 1989).

Chrivia et al. (1993) reported the discovery of a nuclear transcriptional coactivator protein, CREB-binding protein (CBP), that binds specifically to the PKA-phosphorylated form of the CREB protein. CBP is a large protein with a molecular mass of about 250 kD which contains a bromodomain, i.e., a conserved structural unit important for protein-protein interactions. In Drosophila and yeast, this domain is found in coactivator proteins involved in signal-dependent, but not basal, transcription (Nordheim, 1994).

To isolate the gene responsible for Rubinstein-Taybi syndrome (RSTS1; 180849), which is associated with breakpoints in and microdeletions of chromosome 16p13.3, Petrij et al. (1995) used FISH to situate all RSTS breakpoints in an area of 150 kb, thus defining a candidate region. Complementary DNAs from this area showed very high sequence homology with mouse CBP. Further studies indicated that the human CBP gene spans at least the 150-kb genomic area containing the RSTS breakpoints. Giles et al. (1997) reported the cloning and sequencing of human CREBBP, which encodes a deduced 2,442-amino acid protein with a molecular mass of 265 kD with 95% homology to the mouse protein.

In ClinVar, I can only find:

CREBBP CREB binding protein [ Homo sapiens (human) ]

Gene ID: 1387, updated on 5-Mar-2019

Official Symbol: CREBBPprovided

Official Full Name: CREB binding proteinprovided

Primary source: HGNC:HGNC:2348

See related: Ensembl:ENSG00000005339 MIM:600140

Gene type: protein coding

RefSeq status: REVIEWED

Organism: Homo sapiens

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as: CBP; RSTS; KAT3A; MKHK1; RSTS1

Summary: This gene is ubiquitously expressed and is involved in the transcriptional coactivation of many different transcription factors. First isolated as a nuclear protein that binds to cAMP-response element binding protein (CREB), this gene is now known to play critical roles in embryonic development, growth control, and homeostasis by coupling chromatin remodeling to transcription factor recognition. The protein encoded by this gene has intrinsic histone acetyltransferase activity and also acts as a scaffold to stabilize additional protein interactions with the transcription complex. This protein acetylates both histone and non-histone proteins. This protein shares regions of very high sequence similarity with protein p300 in its bromodomain, cysteine-histidine-rich regions, and histone acetyltransferase domain. Mutations in this gene cause Rubinstein-Taybi syndrome (RTS). Chromosomal translocations involving this gene have been associated with acute myeloid leukemia. Alternative splicing results in multiple transcript variants encoding different isoforms. [provided by RefSeq, Feb 2009]

Expression: Ubiquitous expression in testis (RPKM 12.8), bone marrow (RPKM 12.2) and 25 other tissues

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

RPL10:

It shows that there are relationships between RPL10 and autism in both the OMIM and the ClinVar

\* 312173

RIBOSOMAL PROTEIN L10; RPL10

Alternative titles; symbols

QM GENE

HGNC Approved Gene Symbol: RPL10

Cytogenetic location: Xq28 Genomic coordinates (GRCh38): X:154,398,064-154,402,338 (from NCBI)

▼ Description

Ribosomal protein L10 (RPL10) is a highly conserved component of the large ribosome subunit (60S) that plays a crucial role in protein synthesis (summary by Gong et al., 2009).

▼ Molecular Genetics

Susceptibility to X-linked Autism 5

Klauck et al. (2006) identified 2 missense mutations (L206M, 312173.0001 and H213Q, 312173.0002) in conserved residues at the end of exon 7 of the RPL10 gene in patients with autism (AUTSX5; 300847). Chiocchetti et al. (2011) identified an additional family with the H213Q mutation.

X-linked Syndromic Mental Retardation 35

In 3 members of a family with X-linked syndromic mental retardation-35 (MRXS35; 300998), Brooks et al. (2014) identified a hemizygous missense mutation in the RPL10 gene (K78E; 312173.0003). The mutation, which was found by sequencing of an X-linked gene panel and confirmed by Sanger sequencing, segregated with the disorder in the family. Carrier females showed fully skewed X inactivation of the mutation-bearing X chromosomes. Injection of the mutation failed to rescue the microcephaly phenotype of zebrafish rpl10-null morphants (see ANIMAL MODEL), suggesting that K78E is a functionally null allele.

In 4 male members of a family with MRXS35, Thevenon et al. (2015) identified a hemizygous missense mutation in the RPL10 gene (G161S; 312173.0004). The mutation, which was found by exome sequencing, segregated with the disorder in the family. Carrier females showed fully skewed X inactivation of the mutation-bearing X chromosome. Functional studies of the variant and studies of patient cells were not performed.

In 2 male first cousins from Italy with MRXS35, Zanni et al. (2015) identified a hemizygous missense mutation in the RPL10 gene (A64V; 312173.0005). The mutation, which was found by X-chromosome exome sequencing, confirmed by Sanger sequencing, and filtered against public databases, segregated with the disorder in the family. Carrier females showed fully skewed X inactivation. Studies in yeast showed that the A64V mutant protein was functional and able to restore temperature-sensitive growth and translational defects. Ribosomal profile analysis showed that the A64V mutation was associated with a reduction in large 80S peak, indicating a reduction in translation initiation, with an increase in polysomes, indicating an increase in translationally active ribosomes. Of note, the patients had spondyloepiphyseal dysplasia.

Somatic Mutation in T-cell ALL

De Keersmaecker et al. (2013) identified mutations affecting the ribosomal proteins RPL5 (603634) and RPL10 (312173) in 12 of 122 (9.8%) pediatric T-ALLs, with recurrent alterations in RPL10 of arg98, an invariant residue from yeast to human. Yeast and lymphoid cells expressing the RPL10 arg98 to ser mutant showed a ribosome biogenesis defect.

In ClinVar, I got the following result:

RPL10 ribosomal protein L10 [ Homo sapiens (human) ]

Gene ID: 6134, updated on 3-Mar-2019

Official Symbol: RPL10provided

Official Full Name: ribosomal protein L10provided

Primary source: HGNC:HGNC:10298

See related: Ensembl:ENSG00000147403 MIM:312173

Gene type: protein coding

RefSeq status: REVIEWED

Organism: Homo sapiens

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as: QM; L10; NOV; AUTSX5; DXS648; MRXS35; DXS648E

Summary: This gene encodes a ribosomal protein that is a component of the 60S ribosome subunit. The related protein in chicken can bind to c-Jun and can repress c-Jun-mediated transcriptional activation. Some studies have detected an association between variation in this gene and autism spectrum disorders, though others do not detect this relationship. There are multiple pseudogenes of this gene dispersed throughout the genome. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Jan 2015]

Expression: Ubiquitous expression in ovary (RPKM 795.4), lymph node (RPKM 396.3) and 25 other tissues

----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------