Exome Results & Raw Data Summary

1390 Shorebird Way
Mountain View, CA 94043
www.23andme.com

Generated on: July 4, 2012

Congratulations! Your exome has been sequenced and your data is ready for you to download. We have also included this overview of your data to get you started on your exome exploration. Here are a few important points about your exome data:

- Two types of files are available for download: 1) the aligned sequencing reads in BAM format, 2) a file containing variant calls (VCF file).
- The raw data VCF file is a preliminary draft of your exome. Our ability to call variants, especially indels, is greatly improved with each additional exome added to our database. Moreover we will build upon this protocol to include additional steps such as custom treatment of the sex chromosomes. To this end we will update your VCF file at the end of the pilot. We will contact you when this data is available.

Your exome at a glance:

The Exome Service is a pilot project, and this report contains preliminary data only. 23andMe does not represent that all of this information is accurate. **In this report we have used 1000 Genome Project data to report frequencies of variants to determine how common or rare a particular variant is.** We have also only provided information about a subset of the many gene-disrupting variants present in the human genome, in a chosen set of genes. Sequencing was performed such that the total number of bases read was at least 80X the size of the exome. As described in the Exome Terms of Use, 23andMe will not be providing the reports and explanations that 23andMe typically provides to customers with respect to their genotyping results for this data. 23andMe Services are for research, informational, and educational use only. We do not provide medical advice. Please keep in mind that genetic information you share with others could be used against your interests.

Your exome in numbers

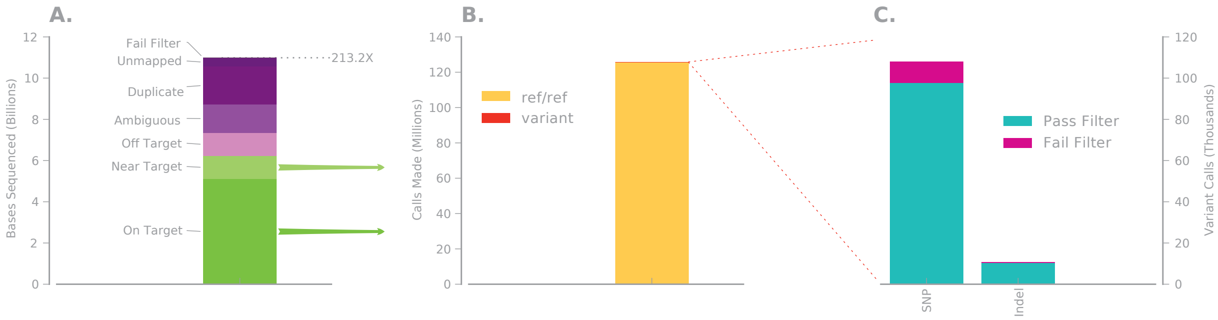


Figure 1: Getting from raw reads to called variants. A) The number of bases obtained by sequencing your exome. The top line indicates total coverage. B) Total number of called bases in your exome. The vast majority are the same as the reference genome. C) An expansion of the small sliver of variants depicted in B. These are the variants present in your VCF file.

Welcome to your exome. Your exome is the 50 million DNA bases of your genome containing the information necessary to encode all your proteins. Your exome data consists of two parts, the raw data (both aligned and unaligned Illumina reads, fig1A) and a draft of the variants present in your exome (fig1C). While this draft is provisional and we will be improving upon it, we wanted to allow you to dig in to your exome as soon as possible so you can tell us what you think is important and should be included.

To create the first draft of your exome we implemented the Broad Institute's "Best Practice" protocol for exome sequencing analysis. You can read a detailed description of it here (for brief summary see Appendix).

Characterizing your variants

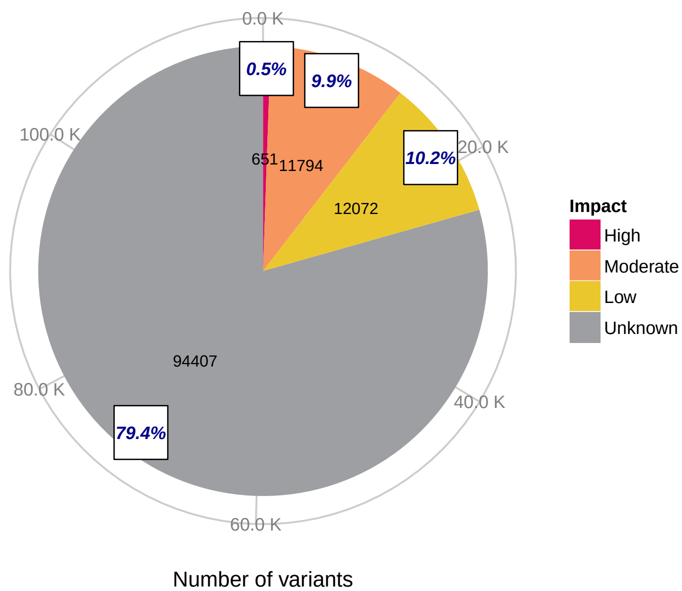


Figure 2: Predicting impact of variants on gene function. An overview of your variants and their predicted impact on gene function.

The variants in your VCF file are the positions in your genome that differ from the reference genome. Most of these variants are likely to be functionally neutral and unlikely to cause any severe disorders. Pinpointing genuine disease mutations is still challenging and we used a number of software tools to identify those that may be functionally important. We estimated the impact a variant has on gene function based on the severity of its effect on the gene product:

High impact:

Frame shift Insertion or deletion of bases, not multiple of 3.

Splice site Variant at the 'splicing site' may disrupt the consensus splicing site sequence.

Stop gain Premature termination of peptides, which would disable protein function.

Start loss Loss of the start codon.

Stop loss Loss of the stop codon.

Moderate impact:

Nonsynonymous substitution Non-conservative change altering an amino acid in a protein.

Codon insertion or deletion Insertion or deletion of bases, multiple of 3.

Low impact:

Synonymous substitution Variant that does not alter the amino acid sequence due to codon degeneracy.

Start gain Variant resulting in the gain of a start codon.

Synonymous stop Variant changing one stop codon into another.

Unknown impact: Variants unlikely to affect gene products.

How rare are your variants?

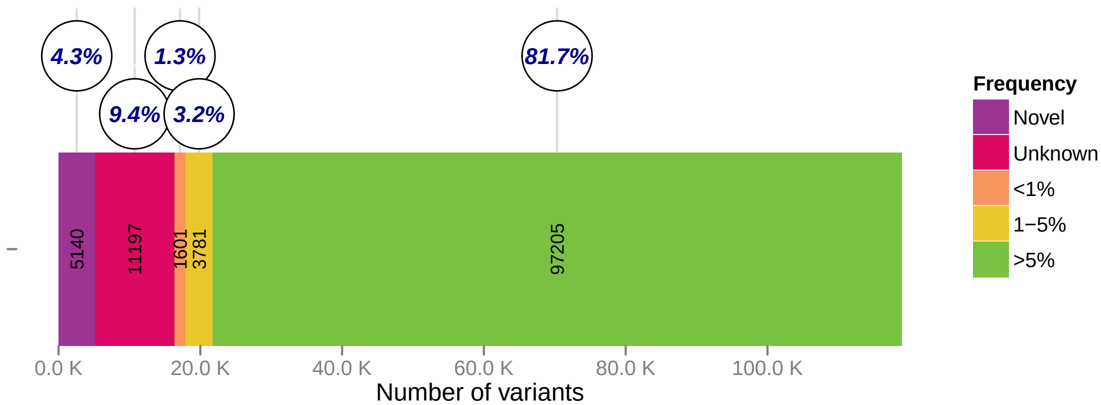


Figure 3: Variant frequencies. The allele frequencies of the variants in your exome. Unknown: allele is present in a public database but no frequency data was available.

One of the advantages of exome sequencing is that we can detect sequence variants that are unique to you! By comparing your variants to all those that have been discovered so far, we can divide your variants into the following categories:

- **novel** variant hasn't been observed in current public sequence databases
- **unknown** variant has been observed in public databases but allelic frequency has not been calculated and therefore is not available
- **rare** variant with allelic frequency $<1\%$
- **somewhat rare** variant with frequency 1-5%
- **common** frequency of the variant is greater than 5%

One of the most comprehensive human variation public datasets is maintained by the 1000 Genomes Project. We use 1000 Genomes Project data (project release: 08-26-2011) to report frequencies of alleles found in your exome, including reporting if it is absent from the public database ($i.e.$ a novel variant).

Filtering your variants

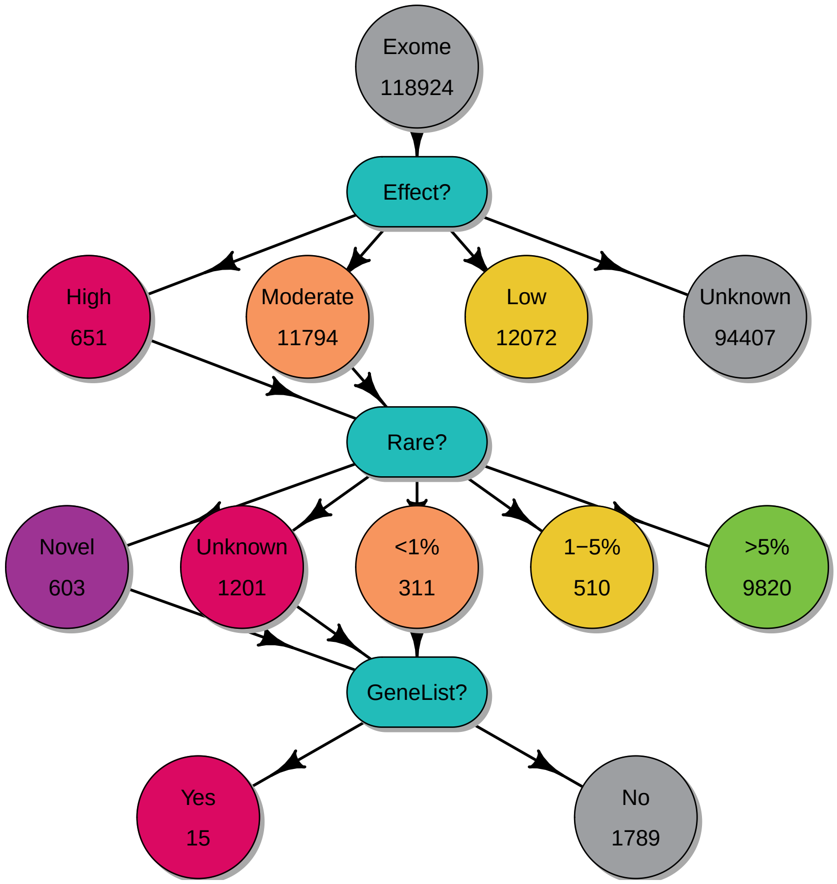


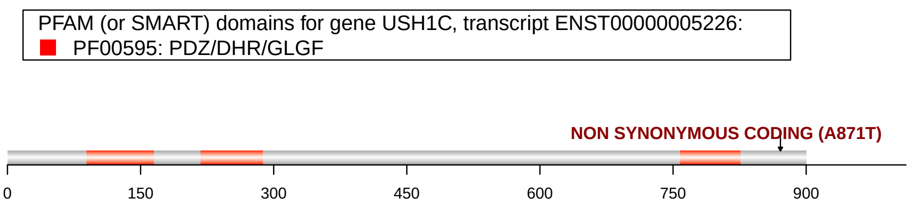
Figure 4: Variant filtering decision tree. A graphical representation of the filtering process that was used to generate your short list of variants of interest.

Most sequence variants in your exome are likely to be neutral and do not cause any severe disorders. A filtering process is often undertaken to prioritize variants discovered through sequencing. To identify potentially interesting and relevant variants with potential functional effects (contributing to disease and other phenotypes of interest) we used three consecutive filters, depicted in the figure above: (1) effect of the variant on the gene product; (2) allele frequency of the variant; (3) location of the variant in one of 592 genes involved in Mendelian disorders (at this point we also exclude indels and variants on the sex chromosomes).

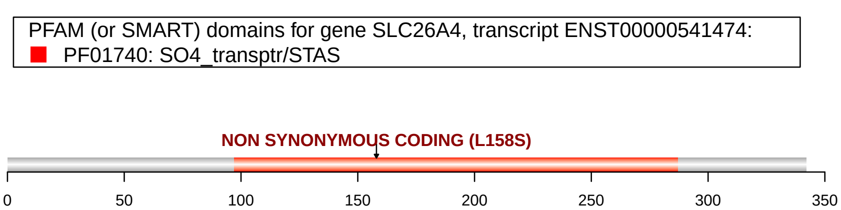
We hope you find this initial list of variants interesting and that it will help you in your journey through your exome. This short list of variants only scratches the surface of what your genome contains and is just the beginning of where your data can take you. Have fun!

List of selected variants

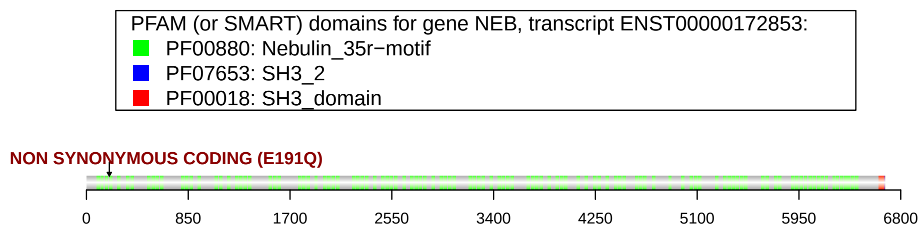
Variant 1:	**Gene:** USH1C **Your genotype:** C/T	**Location:** chr11:17517160
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00320	**dbSNP:** rs56165709
Quality:	**Genotype quality:** 99	**Coverage depth:** 136
Details:	**Gene description:** Usher syndrome 1C (autosomal recessive, severe)	
	Transcript: ENST00000005226	**AA change:** A871T
	EntrezId: 10083	**EnsemblId:** ENSG00000006611
	UniProt: Q9Y6N9	**OMIM:** 605242



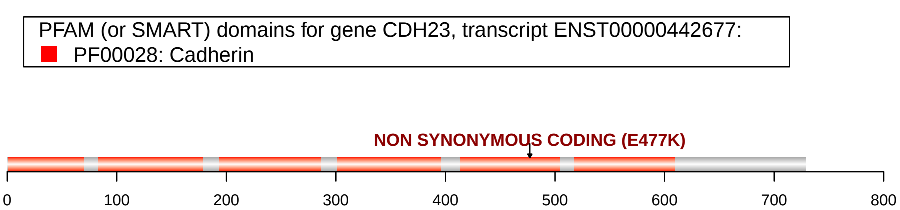
Variant 2:	**Gene:** SLC26A4 **Your genotype:** T/C	**Location:** chr7:107341628
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00410	**dbSNP:** rs55638457
Quality:	**Genotype quality:** 99	**Coverage depth:** 170
Details:	**Gene description:** solute carrier family 26, member 4	
	Transcript: ENST00000541474	**AA change:** L158S
	EntrezId: 5172	**EnsemblId:** ENSG00000091137
	UniProt: O43511	**OMIM:** 605646



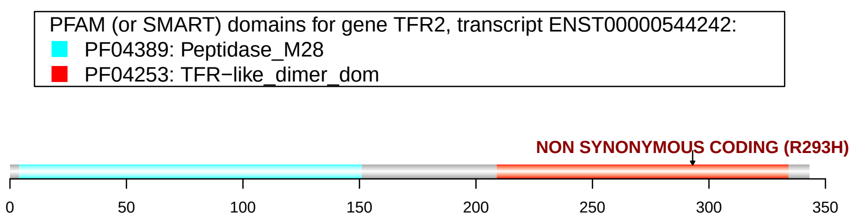
Variant 3:	Gene: NEB Your genotype: C/G Location: chr2:152580815	
Effect:	Impact: NON SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.00870	dbSNP: rs35686968
Quality:	Genotype quality: 99	Coverage depth: 105
Details:	Gene description: nebulin	
	Transcript: ENST00000172853	AA change: E191Q
	EntrezId: 4703	EnsemblId: ENSG00000183091
	UniProt: P20929	OMIM: 161650



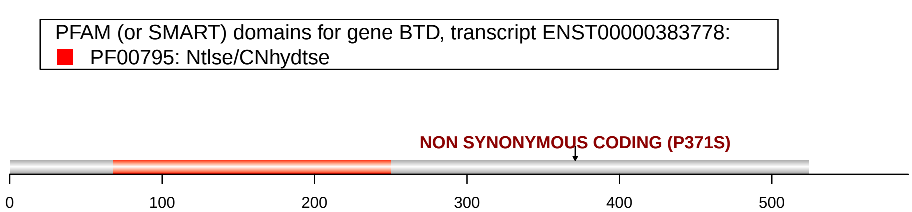
Variant 4:	Gene: CDH23 Your genotype: G/A Location: chr10:73464812	
Effect:	Impact: NON SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.00320	dbSNP: rs111033458
Quality:	Genotype quality: 99	Coverage depth: 77
Details:	Gene description: cadherin-related 23	
	Transcript: ENST00000442677	AA change: E477K
	EntrezId: 64072	EnsemblId: ENSG00000107736
	UniProt: Q9H251	OMIM: 605516



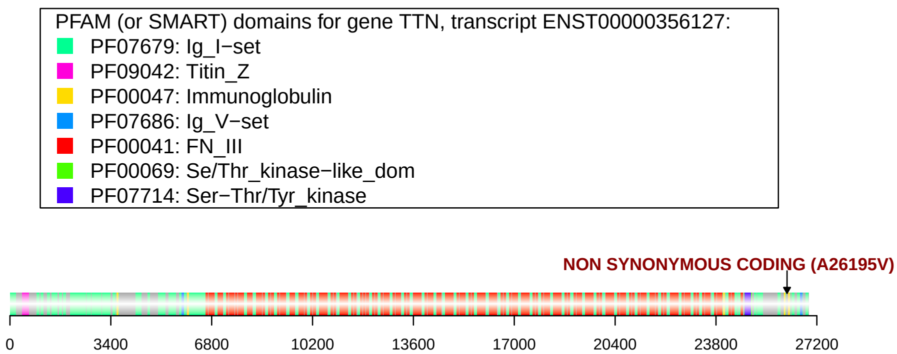
Variant 5:	**Gene:** TFR2 **Your genotype:** C/T **Location:** chr7:100218631	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00990	**dbSNP:** rs41295942
Quality:	**Genotype quality:** 99	**Coverage depth:** 19
Details:	**Gene description:** transferrin receptor 2	
	Transcript: ENST00000544242	**AA change:** R293H
	EntrezId: 7036	**EnsemblId:** ENSG00000106327
	UniProt: Q9UP52	**OMIM:** 604720



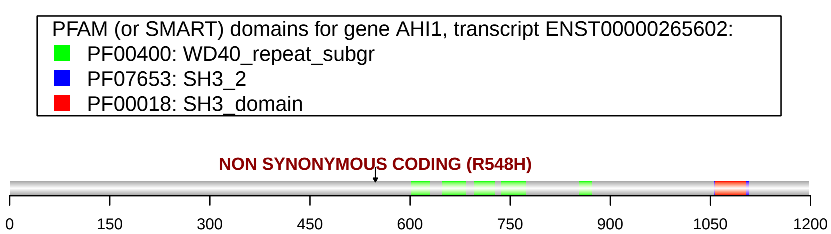
Variant 6:	**Gene:** BTD **Your genotype:** C/T **Location:** chr3:15686534	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00870	**dbSNP:** rs35034250
Quality:	**Genotype quality:** 99	**Coverage depth:** 173
Details:	**Gene description:** biotinidase	
	Transcript: ENST00000383778	**AA change:** P371S
	EntrezId: 686	**EnsemblId:** ENSG00000169814
	UniProt: P43251	**OMIM:** 609019



Variant 7:	**Gene:** TTN **Your genotype:** G/A **Location:** chr2:179395554	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00530	**dbSNP:** rs66961115
Quality:	**Genotype quality:** 99	**Coverage depth:** 195
Details:	**Gene description:** titin	
	Transcript: ENST00000356127	**AA change:** A26195V
	EntrezId: 7273	**EnsemblId:** ENSG00000155657
	UniProt: Q8WZ42	**OMIM:** 188840



Variant 8:	**Gene:** AHI1 **Your genotype:** C/T **Location:** chr6:135768282	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00930	**dbSNP:** rs35433555
Quality:	**Genotype quality:** 99	**Coverage depth:** 177
Details:	**Gene description:** Abelson helper integration site 1	
	Transcript: ENST00000265602	**AA change:** R548H
	EntrezId: 54806	**EnsemblId:** ENSG00000135541
	UniProt: Q8N157	**OMIM:** 608894



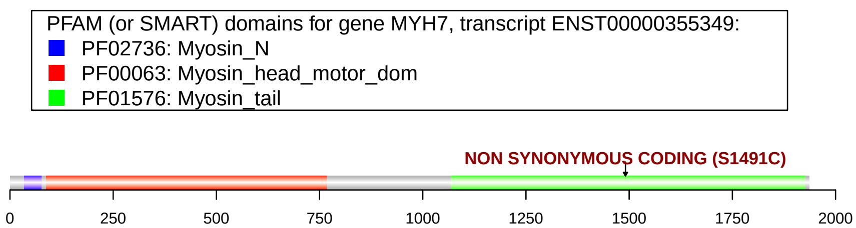
Variant 9: **Gene:** MYH7 **Your genotype:** G/C **Location:** chr14:23886409

Effect: **Impact:** NON SYNONYMOUS CODING **Type:** MODERATE

Frequency: **1KGenomes:** 0.00760 **dbSNP:** rs3729823

Quality: **Genotype quality:** 99 **Coverage depth:** 153

Details: **Gene description:** myosin, heavy chain 7, cardiac muscle, beta
Transcript: ENST00000355349 **AA change:** S1491C
EntrezId: 4625 **EnsemblId:** ENSG00000092054
UniProt: P12883 **OMIM:** 160760



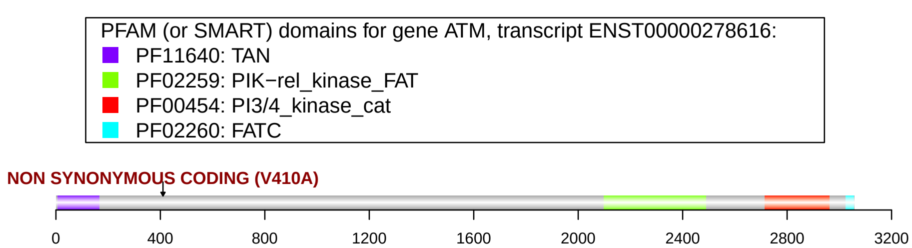
Variant 10: **Gene:** ATM **Your genotype:** T/C **Location:** chr11:108119823

Effect: **Impact:** NON SYNONYMOUS CODING **Type:** MODERATE

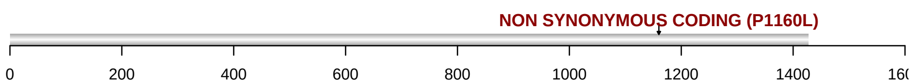
Frequency: **1KGenomes:** 5e-04 **dbSNP:** rs56128736

Quality: **Genotype quality:** 99 **Coverage depth:** 100

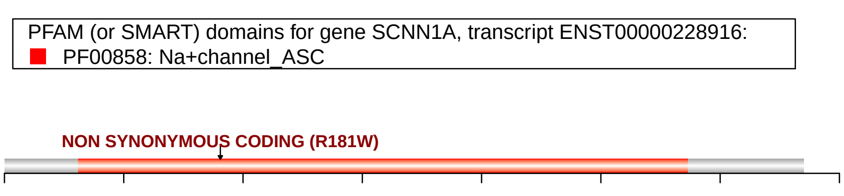
Details: **Gene description:** ataxia telangiectasia mutated
Transcript: ENST00000278616 **AA change:** V410A
EntrezId: 472 **EnsemblId:** ENSG00000149311
UniProt: Q13315 **OMIM:** 607585



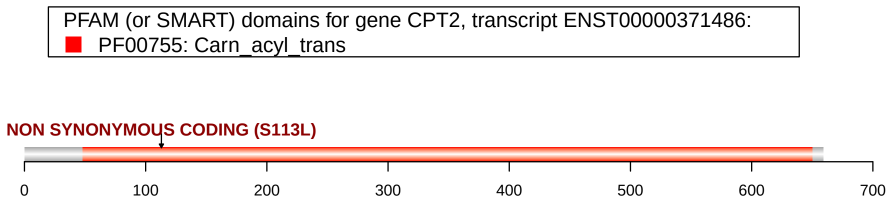
Variant 11:	**Gene:** NPHP4 **Your genotype:** G/A **Location:** chr1:5927169	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00720	**dbSNP:** rs113445782
Quality:	**Genotype quality:** 99	**Coverage depth:** 19
Details:	**Gene description:** nephronophthisis 4	
	Transcript: ENST00000378156	**AA change:** P1160L
	EntrezId: 261734	**EnsemblId:** ENSG00000131697
	UniProt: O75161	**OMIM:** 607215



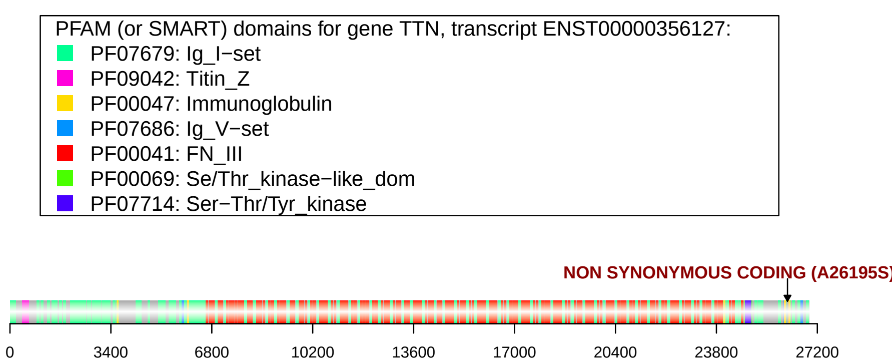
Variant 12:	**Gene:** SCNN1A **Your genotype:** G/A **Location:** chr12:6472752	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00820	**dbSNP:** rs55797039
Quality:	**Genotype quality:** 99	**Coverage depth:** 13
Details:	**Gene description:** sodium channel, non-voltage-gated 1 alpha subunit	
	Transcript: ENST00000228916	**AA change:** R181W
	EntrezId: 6337	**EnsemblId:** ENSG00000111319
	UniProt: P37088	**OMIM:** 600228



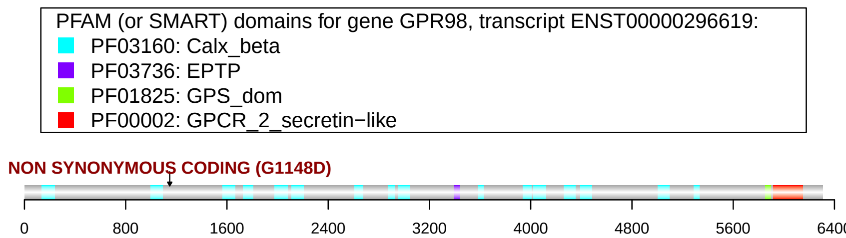
Variant 13:	**Gene:** CPT2 **Your genotype:** C/T **Location:** chr1:53668099	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 9e-04	**dbSNP:** rs74315294
Quality:	**Genotype quality:** 99	**Coverage depth:** 55
Details:	**Gene description:** carnitine palmitoyltransferase 2 **Transcript:** ENST00000371486 **EntrezId:** 1376 **UniProt:** P23786	 **AA change:** S113L **EnsemblId:** ENSG00000157184 **OMIM:** 600650



Variant 14:	**Gene:** TTN **Your genotype:** C/A **Location:** chr2:179395555	
Effect:	**Impact:** NON SYNONYMOUS CODING	**Type:** MODERATE
Frequency:	**1KGenomes:** 0.00530	**dbSNP:** rs67254537
Quality:	**Genotype quality:** 99	**Coverage depth:** 196
Details:	**Gene description:** titin **Transcript:** ENST00000356127 **EntrezId:** 7273 **UniProt:** Q8WZ42	 **AA change:** A26195S **EnsemblId:** ENSG00000155657 **OMIM:** 188840



Variant 15:	Gene: GPR98 Your genotype: G/A Location: chr5:89948189	
Effect:	Impact: NON SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 8e-04	dbSNP: NA
Quality:	Genotype quality: 99	Coverage depth: 174
Details:	Gene description: G protein-coupled receptor 98	
	Transcript: ENST00000296619	AA change: G1148D
	EntrezId: 84059	EnsemblId: ENSG00000164199
	UniProt: Q8WXG9	OMIM: 602851



Appendix

To create the first draft of your exome we implemented the Broad Institute's "Best Practice" protocol for exome sequencing analysis. You can read a detailed description of it here, however a brief summary of it follows:

1. We took your raw reads and aligned them against the reference genome (these are the alignments available in the BAM file of the encrypted download).
2. We used these alignments to identify probable contamination (unaligned reads) and artifacts of sample preparation (PCR duplicates) which are then removed from subsequent steps.
3. From this point on we focus on the reads that align either to one of the exons or within the regions 250 bases up and downstream of it.
4. To improve the quality of the alignments we carry out a more accurate alignment of the reads that overlap known indels or are likely to contain indels themselves.
5. We also recalibrate the base quality scores of the reads to bring them in line with the empirically-determined values.
6. Using these realigned+recalibrated reads we generate allele calls at every position with enough high-quality data and filter out those that are homozygous for the allele present in the reference genome (the vast majority of these are at such a high frequency in the population they're unlikely to be interesting). The remaining SNP and indel calls (variants) are the ones available in the VCF file that you downloaded.
7. As yet no sequencing technology is 100% accurate and the highly duplicated nature of the human genome makes variant calling a challenging task. Consequently, a small proportion of the variant calls in your VCF are likely to be incorrect. To reduce this proportion we applied the filters recommended by the Broad Institute to remove technical artifacts. Variants that pass all filters are marked in your VCF file with a PASS. As the exome pilot progresses and we gather more data we will be able to use more advanced techniques identify potential errors and improve the quality of your exome.