# Exome Results & Raw Data Summary

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

**Generated on: June 20, 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

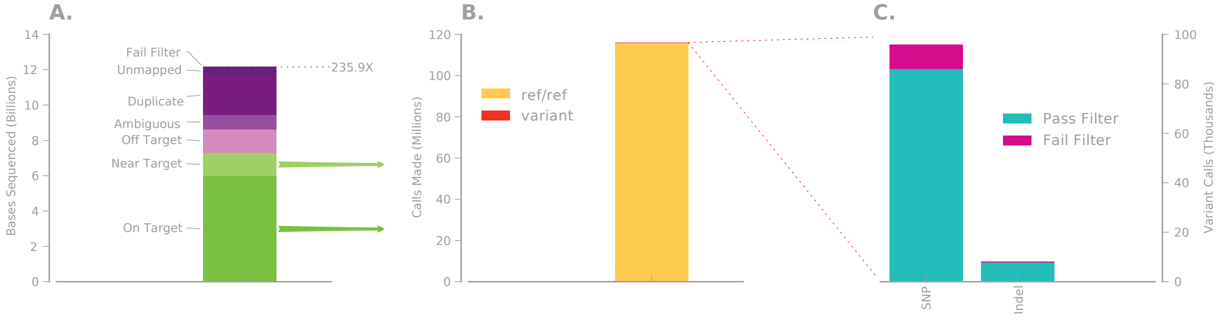


**Figure 1: Getting from raw reads to called variants.** A) The number of bases obtained by sequencing your exome. The top line indicates total coverage. B) Total number of called bases in your exome. The vast majority are the same as the reference genome. C) An expansion of the small sliver of variants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

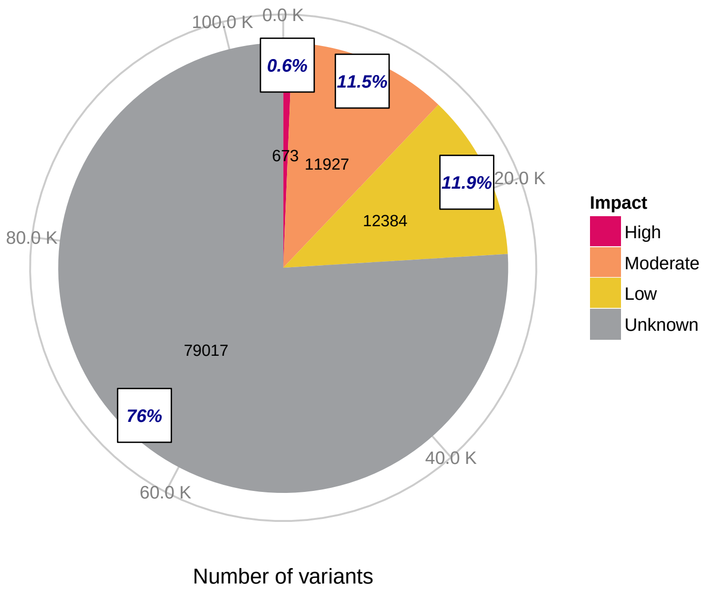


**Figure 2: Predicting impact of variants on gene function.** An overview of your variants and their predicted impact on gene function.

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?

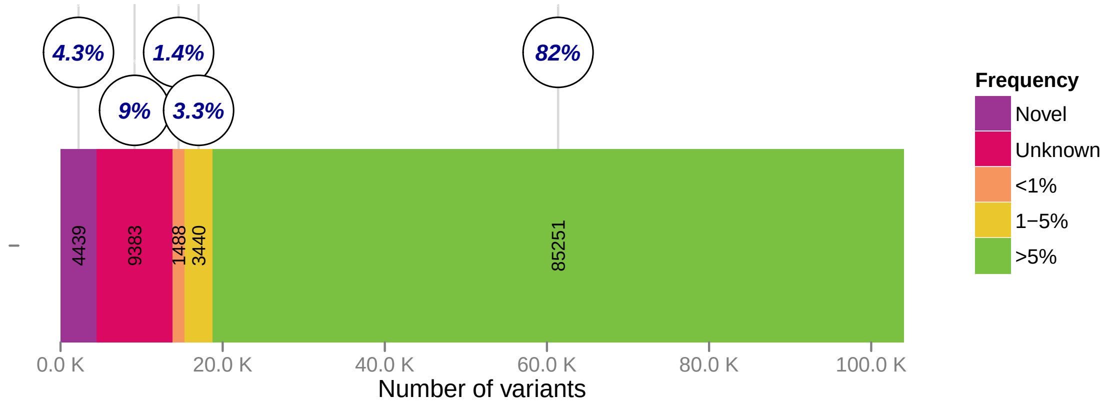


**Figure 3: Variant frequencies.** The allele frequencies of the variants in your exome. Unknown: allele is present in a public database but no frequency data was available.

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

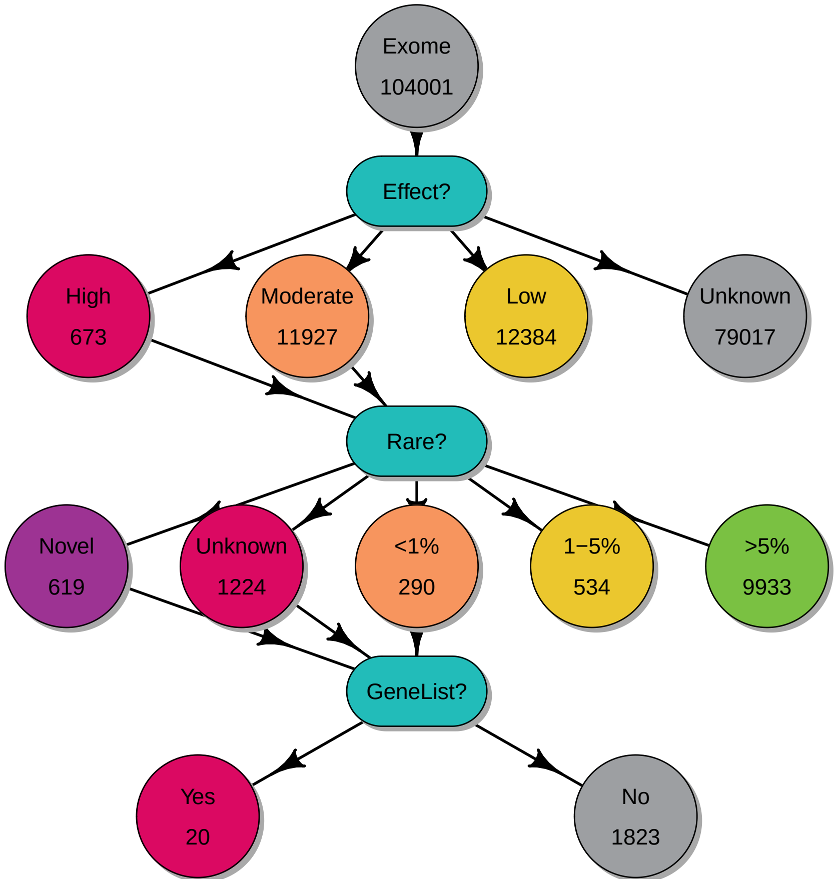


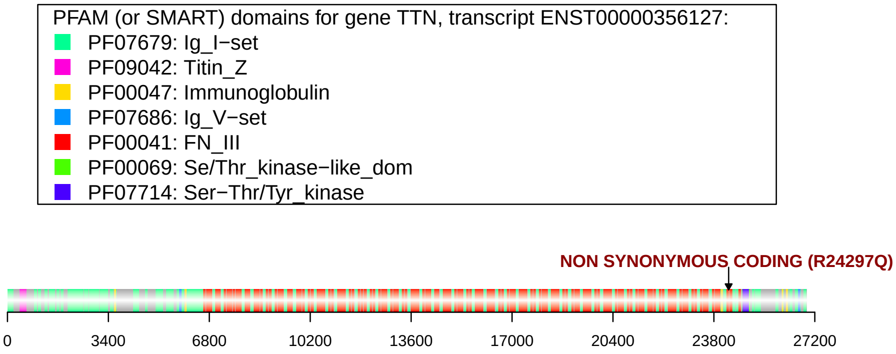
**Figure 4: Variant filtering decision tree.** A graphical representation of the filtering process that was used to generate your short list of variants of interest.

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).

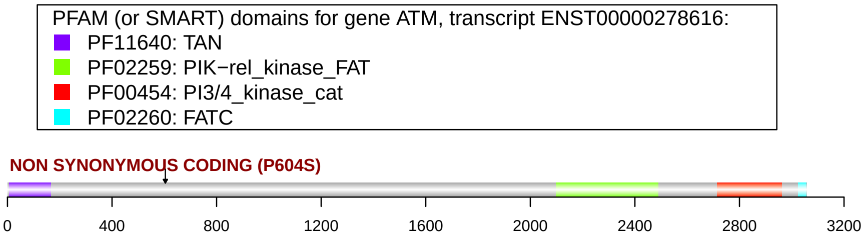
We hope you find this initial list of variants interesting and that it will help you in your journey through your exome. This short list of variants only scratches the surface of what your genome contains and is just the beginning of where your data can take you. Have fun!

# List of selected variants

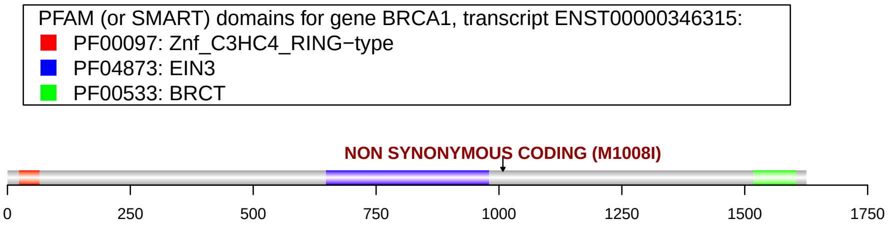
| | | |
|---|---|---|
| Variant 1: | **Gene:** TTN **Your genotype:** C/T | **Location:** chr2:179401742 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00820 | **dbSNP:** rs55742743 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 52 |
| Details: | **Gene description:** titin | |
| | **Transcript:** ENST00000356127 | **AA change:** R24297Q |
| | **EntrezId:** 7273 | **EnsemblId:** ENSG00000155657 |
| | **UniProt:** Q8WZ42 | **OMIM:** 188840 |



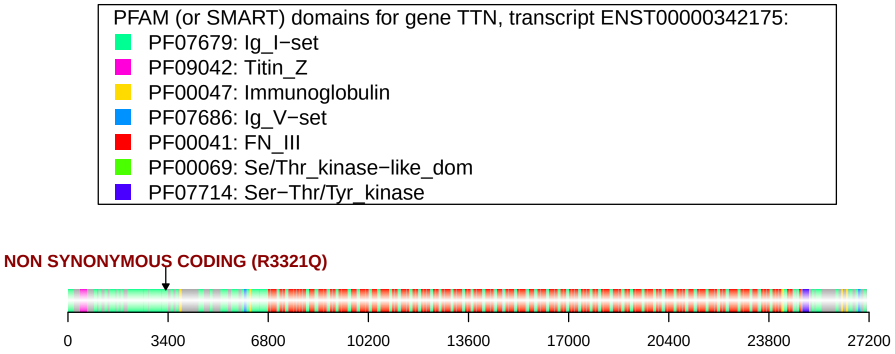
| | | |
|---|---|---|
| Variant 2: | **Gene:** ATM **Your genotype:** C/T | **Location:** chr11:108123551 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00300 | **dbSNP:** rs2227922 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 12 |
| Details: | **Gene description:** ataxia telangiectasia mutated | |
| | **Transcript:** ENST00000278616 | **AA change:** P604S |
| | **EntrezId:** 472 | **EnsemblId:** ENSG00000149311 |
| | **UniProt:** Q13315 | **OMIM:** 607585 |



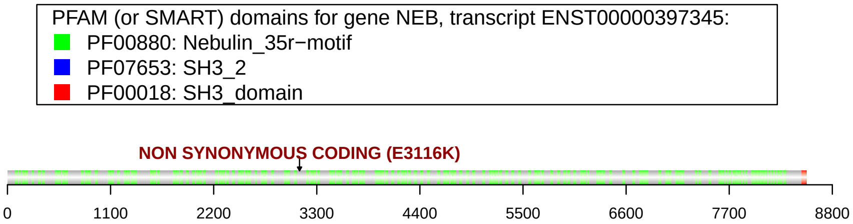
| | | |
|---|---|---|
| **Variant 3:** | **Gene:** BRCA1 **Your genotype:** C/T **Location:** chr17:41244524 | |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00130 | **dbSNP:** rs1800704 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 112 |
| **Details:** | **Gene description:** breast cancer 1, early onset<br>**Transcript:** ENST00000346315<br>**EntrezId:** 672<br>**UniProt:** P38398 | <br>**AA change:** M1008I<br>**EnsemblId:** ENSG00000012048<br>**OMIM:** 113705 |



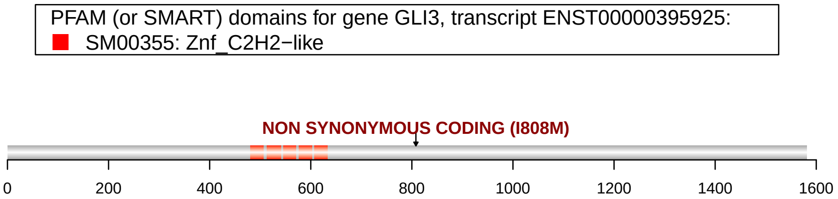
| | | |
|---|---|---|
| **Variant 4:** | **Gene:** TTN **Your genotype:** C/T **Location:** chr2:179628918 | |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00500 | **dbSNP:** rs34819099 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 91 |
| **Details:** | **Gene description:** titin<br>**Transcript:** ENST00000342175<br>**EntrezId:** 7273<br>**UniProt:** Q8WZ42 | <br>**AA change:** R3321Q<br>**EnsemblId:** ENSG00000155657<br>**OMIM:** 188840 |



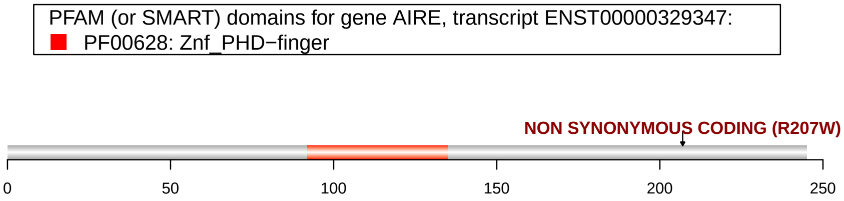
|  |  |  |
|---|---|---|
| **Variant 5:** | **Gene:** NEB **Your genotype:** C/T **Location:** chr2:152490236 |  |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00270 | **dbSNP:** NA |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 250 |
| **Details:** | **Gene description:** nebulin |  |
|  | **Transcript:** ENST00000397345 | **AA change:** E3116K |
|  | **EntrezId:** 4703 | **EnsemblId:** ENSG00000183091 |
|  | **UniProt:** P20929 | **OMIM:** 161650 |


PFAM (or SMART) domains for gene NEB, transcript ENST00000397345:
🟩 PF00880: Nebulin_35r−motif
🟦 PF07653: SH3_2
🟥 PF00018: SH3_domain
NON SYNONYMOUS CODING (E3116K)
0
1100
2200
3300
4400
5500
6600
7700
8800

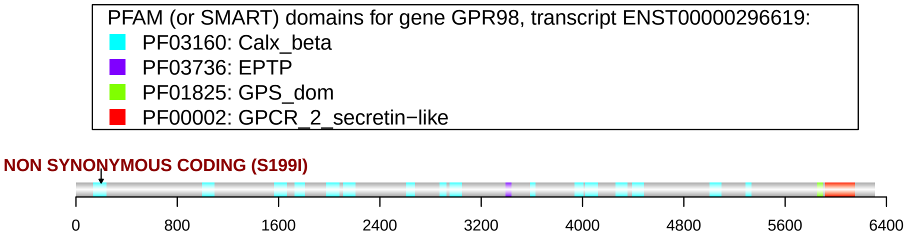
|  |  |  |
|---|---|---|
| **Variant 6:** | **Gene:** GLI3 **Your genotype:** T/C **Location:** chr7:42007201 |  |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00100 | **dbSNP:** rs62622373 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 192 |
| **Details:** | **Gene description:** GLI family zinc finger 3 |  |
|  | **Transcript:** ENST00000395925 | **AA change:** I808M |
|  | **EntrezId:** 2737 | **EnsemblId:** ENSG00000106571 |
|  | **UniProt:** P10071 | **OMIM:** 165240 |


PFAM (or SMART) domains for gene GLI3, transcript ENST00000395925:
🟥 SM00355: Znf_C2H2−like
NON SYNONYMOUS CODING (I808M)
0
200
400
600
800
1000
1200
1400
1600

|  |  |  |
|---|---|---|
| **Variant 7:** | **Gene:** AIRE **Your genotype:** C/T | **Location:** chr21:45713715 |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 5e-04 | **dbSNP:** rs72650677 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 113 |
| **Details:** | **Gene description:** autoimmune regulator<br>**Transcript:** ENST00000329347<br>**EntrezId:** 326<br>**UniProt:** O43918 | <br>**AA change:** R207W<br>**EnsemblId:** ENSG00000160224<br>**OMIM:** 607358 |

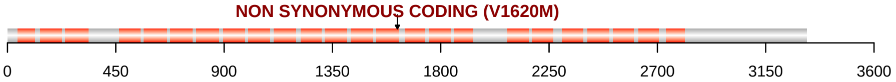


|  |  |  |
|---|---|---|
| **Variant 8:** | **Gene:** GPR98 **Your genotype:** G/T | **Location:** chr5:89920984 |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 5e-04 | **dbSNP:** rs61745496 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 191 |
| **Details:** | **Gene description:** G protein-coupled receptor 98<br>**Transcript:** ENST00000296619<br>**EntrezId:** 84059<br>**UniProt:** Q8WXG9 | <br>**AA change:** S199I<br>**EnsemblId:** ENSG00000164199<br>**OMIM:** 602851 |



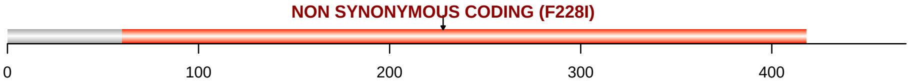
| | | |
|---|---|---|
| **Variant 9:** | **Gene:** CDH23 **Your genotype:** G/A | **Location:** chr10:73537449 |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00560 | **dbSNP:** rs41281330 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 152 |
| **Details:** | **Gene description:** cadherin-related 23 | |
| | **Transcript:** ENST00000398855 | **AA change:** V1620M |
| | **EntrezId:** 64072 | **EnsemblId:** ENSG00000107736 |
| | **UniProt:** Q9H251 | **OMIM:** 605516 |

PFAM (or SMART) domains for gene CDH23, transcript ENST00000398855:
🟥 PF00028: Cadherin

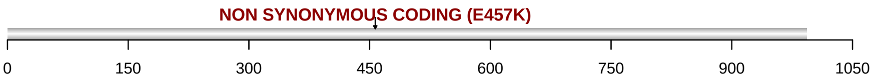


| | | |
|---|---|---|
| **Variant 10:** | **Gene:** WNT10A **Your genotype:** T/A | **Location:** chr2:219755011 |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00970 | **dbSNP:** rs121908120 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 42 |
| **Details:** | **Gene description:** wingless-type MMTV integration site family, member 10A | |
| | **Transcript:** ENST00000258411 | **AA change:** F228I |
| | **EntrezId:** 80326 | **EnsemblId:** ENSG00000135925 |
| | **UniProt:** Q9GZT5 | **OMIM:** 606268 |

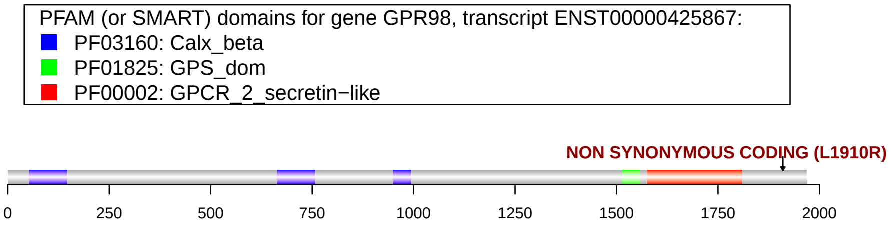
PFAM (or SMART) domains for gene WNT10A, transcript ENST00000258411:
🟥 PF00110: Wnt



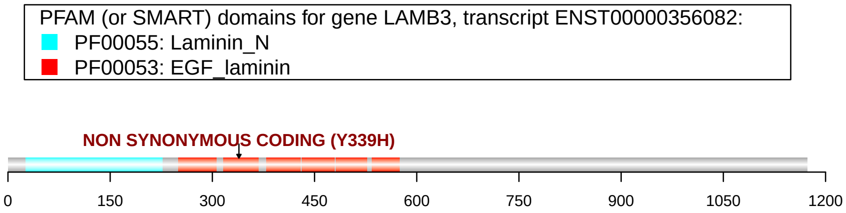
| | | |
|---|---|---|
| Variant 11: | **Gene:** EVC **Your genotype:** G/A **Location:** chr4:5755565 | |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00140 | **dbSNP:** rs141859946 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 137 |
| Details: | **Gene description:** Ellis van Creveld syndrome | |
| | **Transcript:** ENST00000264956 | **AA change:** E457K |
| | **EntrezId:** 2121 | **EnsemblId:** ENSG00000072840 |
| | **UniProt:** P57679 | **OMIM:** 604831 |



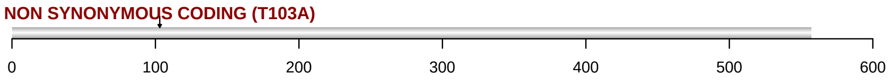
| | | |
|---|---|---|
| Variant 12: | **Gene:** GPR98 **Your genotype:** T/G **Location:** chr5:90449159 | |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00280 | **dbSNP:** rs41311625 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 57 |
| Details: | **Gene description:** G protein-coupled receptor 98 | |
| | **Transcript:** ENST00000425867 | **AA change:** L1910R |
| | **EntrezId:** 84059 | **EnsemblId:** ENSG00000164199 |
| | **UniProt:** Q8WXG9 | **OMIM:** 602851 |



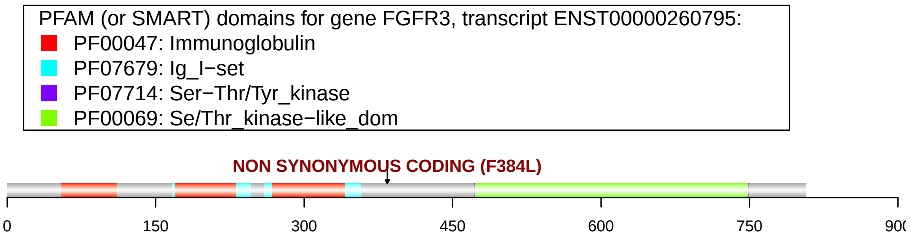
| | | |
|---|---|---|
| Variant 13: | **Gene:** LAMB3 **Your genotype:** A/G | **Location:** chr1:209803199 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00450 | **dbSNP:** rs52814161 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 129 |
| Details: | **Gene description:** laminin, beta 3 | |
| | **Transcript:** ENST00000356082 | **AA change:** Y339H |
| | **EntrezId:** 3914 | **EnsemblId:** ENSG00000196878 |
| | **UniProt:** Q13751 | **OMIM:** 150310 |


PFAM (or SMART) domains for gene LAMB3, transcript ENST00000356082:
🟦 PF00055: Laminin_N
🟥 PF00053: EGF_laminin
NON SYNONYMOUS CODING (Y339H)
0
150
300
450
600
750
900
1050
1200

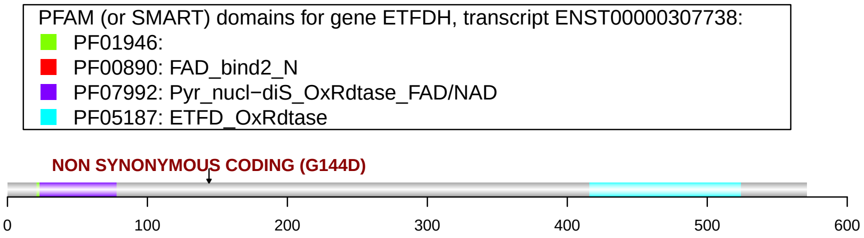
| | | |
|---|---|---|
| Variant 14: | **Gene:** SEPN1 **Your genotype:** A/G | **Location:** chr1:26131638 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00940 | **dbSNP:** rs35019869 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 207 |
| Details: | **Gene description:** selenoprotein N, 1 | |
| | **Transcript:** ENST00000354177 | **AA change:** T103A |
| | **EntrezId:** 57190 | **EnsemblId:** ENSG00000162430 |
| | **UniProt:** Q9NZV5 | **OMIM:** 606210 |



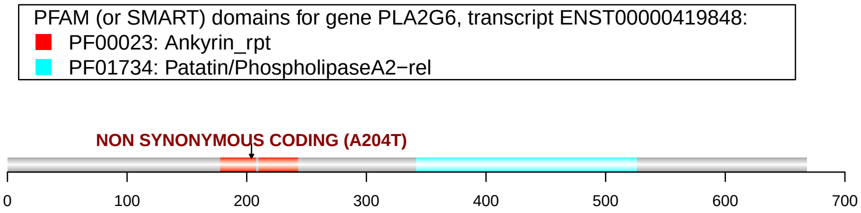
| | | |
|---|---|---|
| **Variant 15:** | **Gene:** FGFR3 **Your genotype:** T/C **Location:** chr4:1806131 | |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 0.00190 | **dbSNP:** rs17881656 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 250 |
| **Details:** | **Gene description:** fibroblast growth factor receptor 3 | |
| | **Transcript:** ENST00000260795 | **AA change:** F384L |
| | **EntrezId:** 2261 | **EnsemblId:** ENSG00000068078 |
| | **UniProt:** P22607 | **OMIM:** 134934 |


PFAM (or SMART) domains for gene FGFR3, transcript ENST00000260795:
🟥 PF00047: Immunoglobulin
🟦 PF07679: Ig_I−set
🟪 PF07714: Ser−Thr/Tyr_kinase
🟩 PF00069: Se/Thr_kinase−like_dom
NON SYNONYMOUS CODING (F384L)
0
150
300
450
600
750
90

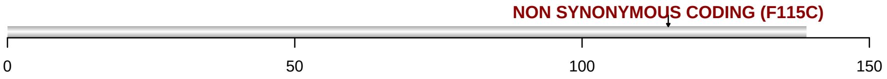
| | | |
|---|---|---|
| **Variant 16:** | **Gene:** ETFDH **Your genotype:** G/A **Location:** chr4:159606337 | |
| **Effect:** | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| **Frequency:** | **1KGenomes:** 5e-04 | **dbSNP:** rs147219158 |
| **Quality:** | **Genotype quality:** 99 | **Coverage depth:** 18 |
| **Details:** | **Gene description:** electron-transferring-flavoprotein dehydrogenase | |
| | **Transcript:** ENST00000307738 | **AA change:** G144D |
| | **EntrezId:** 2110 | **EnsemblId:** ENSG00000171503 |
| | **UniProt:** Q16134 | **OMIM:** 231675 |


PFAM (or SMART) domains for gene ETFDH, transcript ENST00000307738:
🟩 PF01946:
🟥 PF00890: FAD_bind2_N
🟪 PF07992: Pyr_nucl−diS_OxRdtase_FAD/NAD
🟦 PF05187: ETFD_OxRdtase
NON SYNONYMOUS CODING (G144D)
0
100
200
300
400
500
600

| | | |
|---|---|---|
| Variant 17: | **Gene:** PLA2G6 **Your genotype:** C/T | **Location:** chr22:38528888 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 0.00590 | **dbSNP:** rs11570680 |
| Quality: | **Genotype quality:** 42.48 | **Coverage depth:** 12 |
| Details: | **Gene description:** phospholipase A2, group VI (cytosolic, calcium-independent) | |
| | **Transcript:** ENST00000419848 | **AA change:** A204T |
| | **EntrezId:** 8398 | **EnsemblId:** ENSG00000184381 |
| | **UniProt:** O60733 | **OMIM:** 603604 |

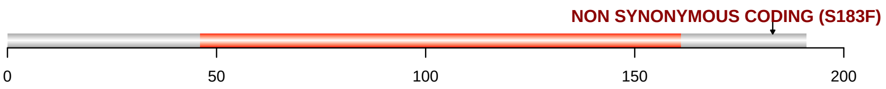


| | | |
|---|---|---|
| Variant 18: | **Gene:** ATM **Your genotype:** T/G | **Location:** chr11:108160480 |
| Effect: | **Impact:** NON SYNONYMOUS CODING | **Type:** MODERATE |
| Frequency: | **1KGenomes:** 5e-04 | **dbSNP:** rs138327406 |
| Quality: | **Genotype quality:** 99 | **Coverage depth:** 22 |
| Details: | **Gene description:** ataxia telangiectasia mutated | |
| | **Transcript:** ENST00000389511 | **AA change:** F115C |
| | **EntrezId:** 472 | **EnsemblId:** ENSG00000149311 |
| | **UniProt:** Q13315 | **OMIM:** 607585 |



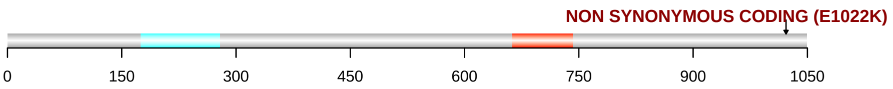
| | | |
|---|---|---|
| Variant 19: | Gene: HSPG2 Your genotype: G/A Location: chr1:22216398 | |
| Effect: | Impact: NON SYNONYMOUS CODING | Type: MODERATE |
| Frequency: | 1KGenomes: 0.00180 | dbSNP: NA |
| Quality: | Genotype quality: 99 | Coverage depth: 62 |
| Details: | Gene description: heparan sulfate proteoglycan 2 | |
| | Transcript: ENST00000439717 | AA change: S183F |
| | EntrezId: 3339 | EnsemblId: ENSG00000142798 |
| | UniProt: P98160 | OMIM: 142461 |

PFAM (or SMART) domains for gene HSPG2, transcript ENST00000439717:
🟥 SM00200: SEA



| | | |
|---|---|---|
| Variant 20: | Gene: ANK2 Your genotype: G/A Location: chr4:114294537 | |
| Effect: | Impact: NON SYNONYMOUS CODING | Type: MODERATE |
| Frequency: | 1KGenomes: 0.00140 | dbSNP: rs45454496 |
| Quality: | Genotype quality: 99 | Coverage depth: 250 |
| Details: | Gene description: ankyrin 2, neuronal | |
| | Transcript: ENST00000509550 | AA change: E1022K |
| | EntrezId: 287 | EnsemblId: ENSG00000145362 |
| | UniProt: Q01484 | OMIM: 106410 |

PFAM (or SMART) domains for gene ANK2, transcript ENST00000509550:
🟦 PF00791: ZU5
🟥 PF00531: Death



# 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.