



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

Exome Results & Raw Data Summary

Generated on: October 12, 2012

Congratulations on being a part of 23andMe's Exome Pilot! Earlier this year, we provided the raw sequencing data and an initial report of your processed results. Since then, we've been working on improving our analysis and generating a final report to summarize your exome. Here are some important points about your final report:

- Your final data comes in the form of two files: 1) the variant call file ([VCF](#)) that contains information about the positions where you differ from the human reference genome (ie. variants), 2) a [BED](#) file containing the genomic regions where we could confidently assess your genotype including positions where you match the reference genome. Both of these files are viewable using a text editor.
- The final VCF file provided is improved over the initial one. In this version, we identified variants based on the data of all people in the exome pilot, and updated variant quality estimates based on known variation. This allows us to better identify and filter your variants, please see the [appendix](#) for more details.

Your exome at a glance:

- [Your exome in numbers](#)
- [Characterizing your variants](#)
- [How rare are your variants?](#)
- [Comparing your variants](#)
- [Filtering your variants](#)
- [Exome carrier status report](#)
- [See selected variants](#)
- [Appendix](#)

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

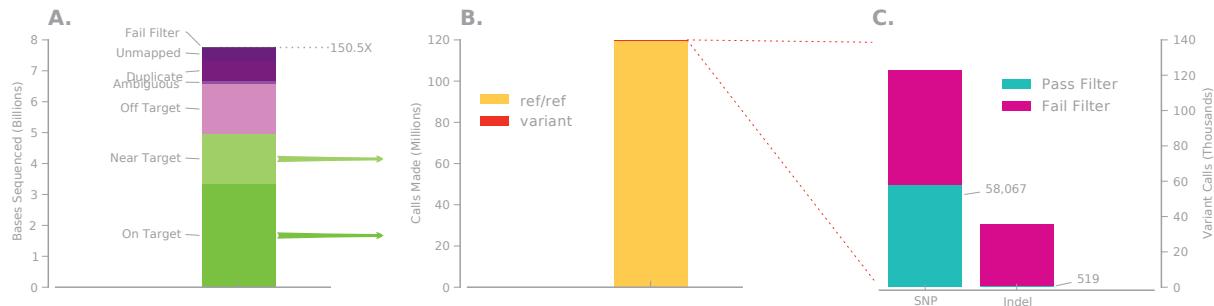


Figure 1: Getting from raw reads to called variants. A) The number of bases obtained by sequencing your exome. The top line indicates total coverage. B) Total number of called bases in your exome. The vast majority are the same as the reference genome. C) An expansion of the small sliver of variants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

Welcome to your exome, the 50 million DNA bases of your genome encoding all your proteins. This data begins as a collection of raw reads which are then aligned against the reference genome (Figure 1A). We analyze the regions where multiple reads overlap to detect where your DNA sequence differs from the reference. In most positions, you will match the reference sequence exactly (Figure 1B), but the small number of variants where you differ are collected into a final VCF file (Figure 1C). The figures in this report are based on the variants that pass all filters.

There are many approaches to this process. We implemented the Broad Institute's "Best Practice" protocol for exome sequence analysis (see [Appendix](#)). You can read a detailed description of it [here](#).

Characterizing your variants

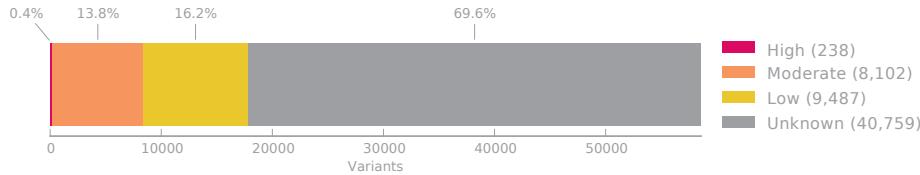


Figure 2: Predicting impact of variants on gene function. An overview of your variants and their predicted impact on gene function.

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, a 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:

All Variant falls either in an intron, UTR, non-coding transcript or up-/downstream of a gene. These variants are less likely to impact the amino-acid sequence of the protein, however may affect other elements of gene expression.

How rare are your variants?

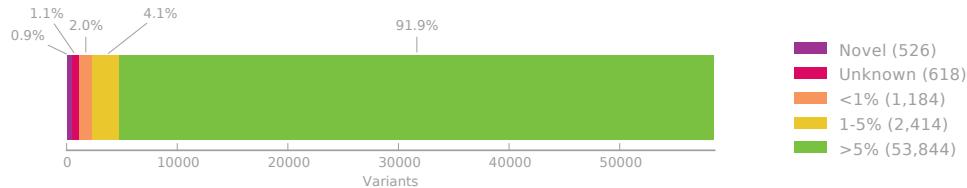


Figure 3: Variant frequencies. The allele frequencies of the variants in your exome. Unknown: allele is present in a public database but no frequency data was available.

One of the advantages of exome sequencing is that we can detect sequence variants that are unique to you! We compared your variants to dbSNP (build 135) and the variants detected by the 1000 Genomes Project (release: 08-26-2011) to divide your variants into the following categories:

- **novel** variant has not been observed in either database
- **unknown** variant has been observed in dbSNP but not the 1000 Genomes dataset and therefore no allele frequency is available
- **rare** variant with an allele frequency <1%
- **somewhat rare** variant with a frequency 1-5%
- **common** frequency of the variant is greater than 5%

Comparing your variants

Now that we have data for everybody in the exome pilot we can see how you compare to the other participants. In the following series of figures we divide your variants into different categories and plot the number of variants in each category as bar chart. We then overlay a [Box Plot](#) showing a summary of the equivalent distribution for all exomes in the pilot.

There are many different ways that we could compare the data, here are the ones that we found to be the most informative:

Impact

Figure 4 breaks down your variants by their predicted impact on gene function.

Effect

Figure 5 takes your high-impact variants and further classifies them according to their predicted effect on the gene product.

Location

Figure 6 looks at the location of your variants relative to the coding sequence.

Frequency

Figure 7 looks at the allele frequencies of your variants.

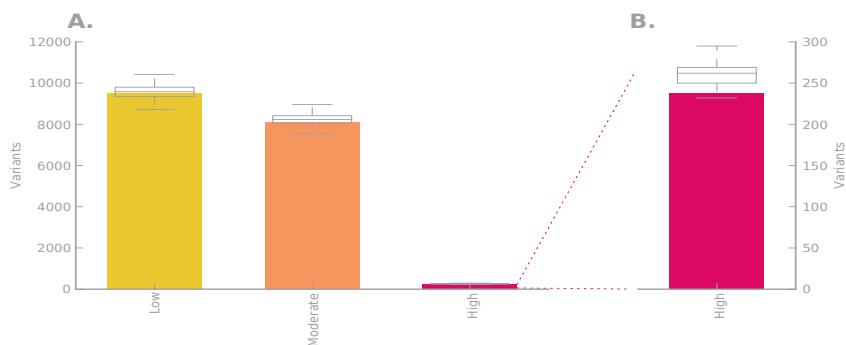


Figure 4: A comparison of the predicted impact of your variants. A) A breakdown of your variants into Low, Medium and High predicted impact (those with Unknown impact not shown). B) Zoom-in of variants predicted to have high impact.

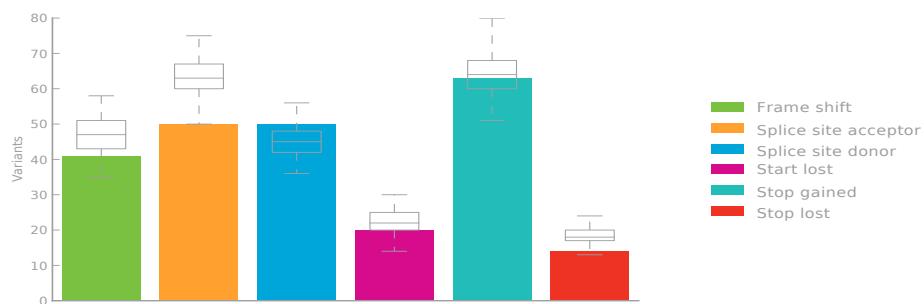


Figure 5: A comparison of the predicted effect of your high-impact variants. Your high-impact variants classified according to their predicted effect on the gene product.

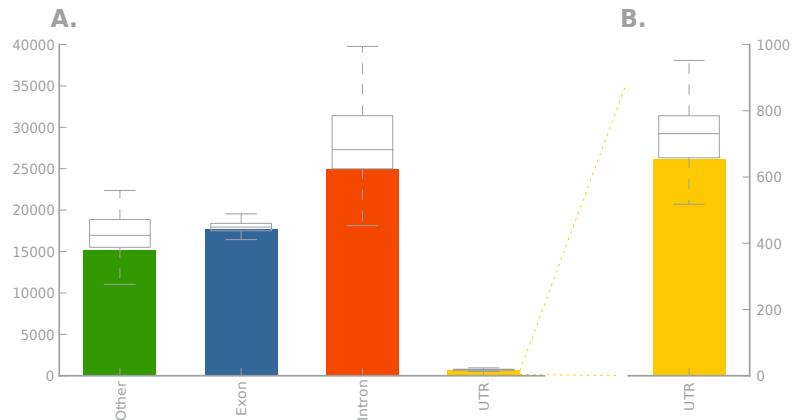


Figure 6: A comparison of the location of your variants relative to the coding sequence. A) Your variants are classified according to whether they overlap the coding portion of a transcript (Exon), the non-coding portion of a transcript (UTR) or an intron. Variants that are either upstream or downstream of a gene or in non-coding transcripts are classified as 'Other'. B) Zoom-in of variants located in the UTRs.

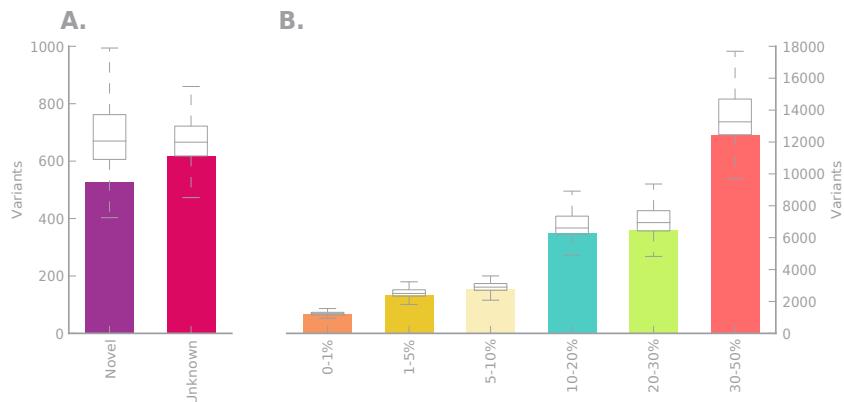


Figure 7: A comparison of the allele frequencies of your variants. A) The number of variants in your exome that are not present in one of the public databases (Novel) and those with no allele frequency in the 1000 Genomes Project (Unknown). B) The remainder of your variants with an allele frequency < 50% categorized by frequency.

Filtering your variants

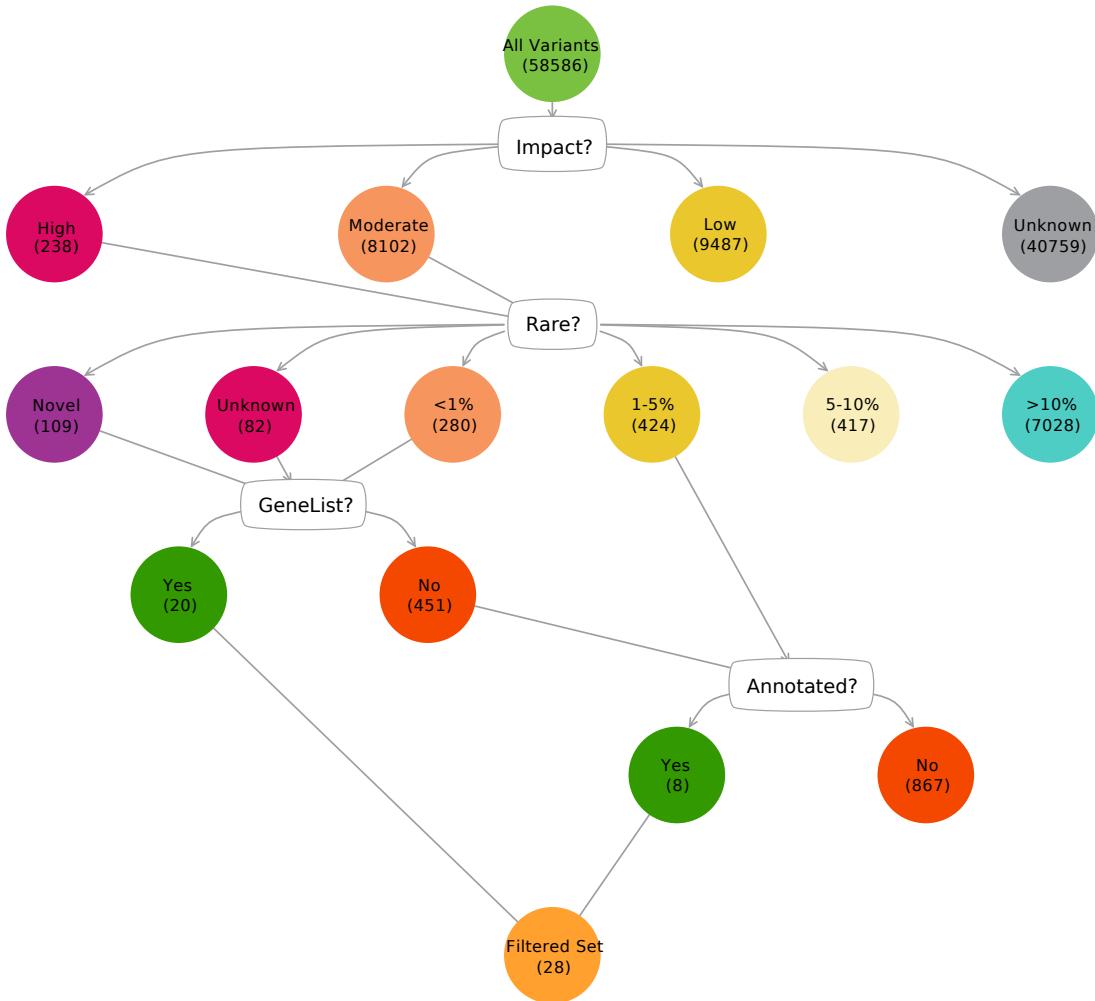


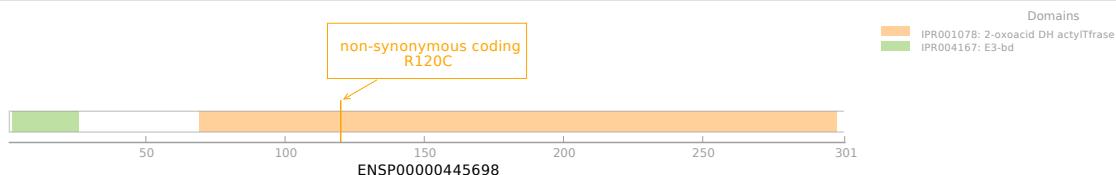
Figure 8: Variant filtering decision tree. A graphical representation of the filtering process that was used to generate your short list of variants of interest.

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 variants with potential functional effects (such as contributing to disease or other phenotypes of interest) we used four consecutive filters, depicted in the figure above: (1) impact 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; (4) annotated in dbSNP as either pathogenic or probably pathogenic.

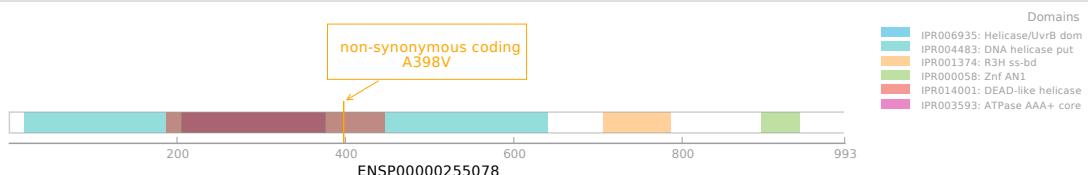
We hope you find this initial list of variants interesting and that it will help you in your journey through your exome. This short list of variants only scratches the surface of what your genome contains and is just the beginning of where your data can take you. Have fun!

List of selected variants

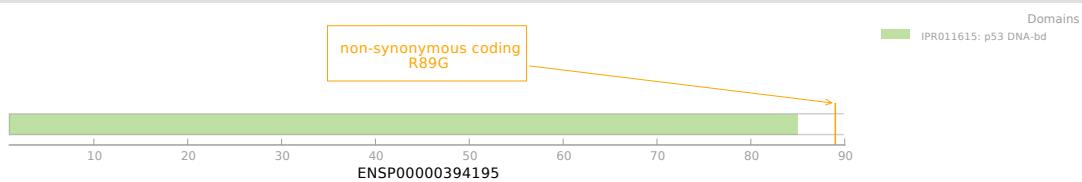
Variant 1:	Gene: DBT Your genotype: G/A Location: chr1:100680411	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0005	dbSNP: rs185492864
Quality:	Genotype quality: 99.00	Coverage depth: 62
Details:	Gene description: dihydrolipoamide branched chain transacylase E2 Transcript: ENST00000543138 EntrezId: 1629 UniProt: NA	AA change: R120C EnsemblId: ENSG00000137992 OMIM: 248610



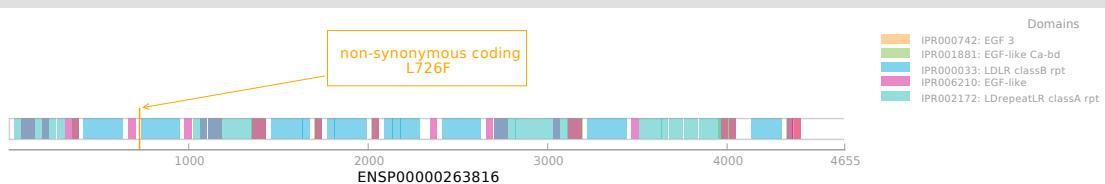
Variant 2:	Gene: IGHMBP2 Your genotype: C/T Location: chr11:68696783	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0005	dbSNP: rs35193202
Quality:	Genotype quality: 99.00	Coverage depth: 87
Details:	Gene description: immunoglobulin mu binding protein 2 Transcript: ENST00000255078 EntrezId: 3508 UniProt: P38935	AA change: A398V EnsemblId: ENSG00000132740 OMIM: 600502



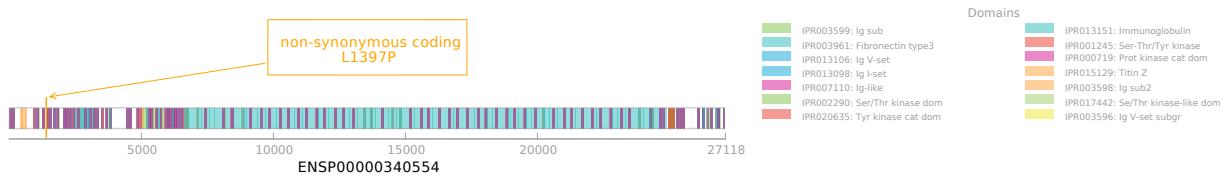
Variant 3:	Gene: TP53 Your genotype: T/C Location: chr17:7578146	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0005	dbSNP: rs34949160
Quality:	Genotype quality: 99.00	Coverage depth: 17
Details:	Gene description: tumor protein p53 Transcript: ENST00000414315 EntrezId: 7157 UniProt: NA	AA change: R89G EnsemblId: ENSG00000141510 OMIM: 191170



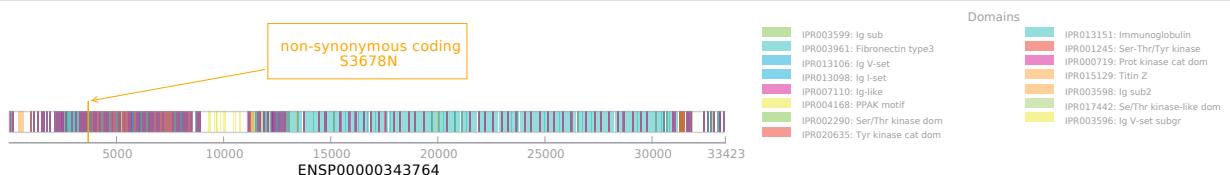
Variant 4:	Gene: LRP2 Your genotype: C/A Location: chr2:170127556	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0009	dbSNP: rs144451000
Quality:	Genotype quality: 99.00	Coverage depth: 66
Details:	Gene description: low density lipoprotein receptor-related protein 2 Transcript: ENST00000263816 EntrezId: 4036 UniProt: P98164	AA change: L726F EnsemblId: ENSG00000081479 OMIM: 600073



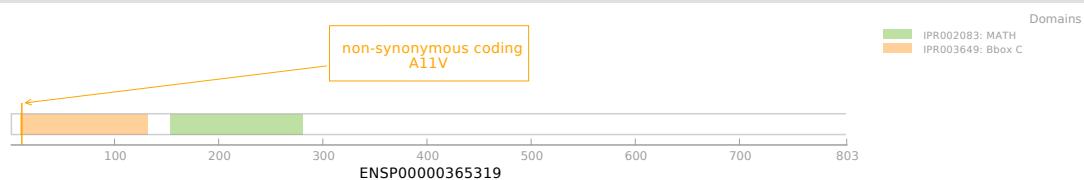
Variant 5:	Gene: TTN Your genotype: A/G Location: chr2:179642583	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0018	dbSNP: rs142317580
Quality:	Genotype quality: 99.00	Coverage depth: 62
Details:	Gene description: titin Transcript: ENST00000342175 EntrezId: 7273 UniProt: NA	AA change: L1397P EnsemblId: ENSG00000155657 OMIM: 188840



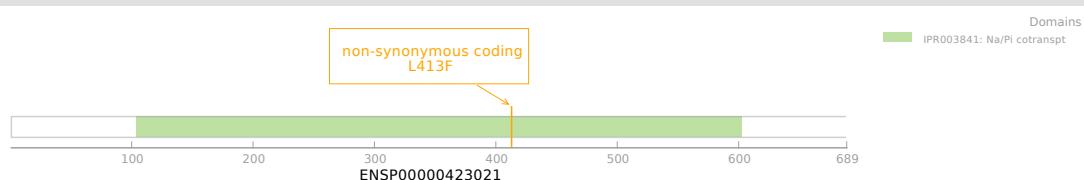
Variant 6:	Gene: TTN Your genotype: C/T Location: chr2:179600408	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0019	dbSNP: rs184740744
Quality:	Genotype quality: 99.00	Coverage depth: 126
Details:	Gene description: titin Transcript: ENST00000342992 EntrezId: 7273 UniProt: NA	AA change: S3678N EnsemblId: ENSG00000155657 OMIM: 188840



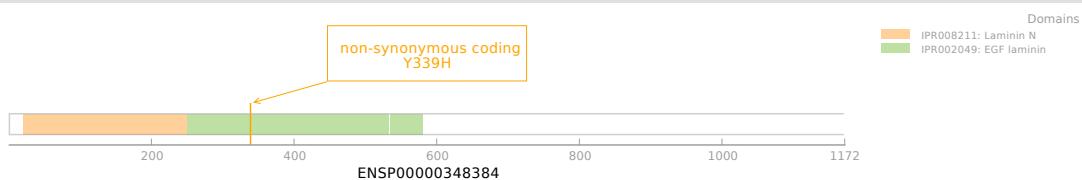
Variant 7:	Gene: TRIM37 Your genotype: G/A Location: chr17:57158552	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0027	dbSNP: rs61758100
Quality:	Genotype quality: 99.00	Coverage depth: 85
Details:	Gene description: tripartite motif containing 37 Transcript: ENST00000376149 AA change: A11V EntrezId: 4591 EnsemblId: ENSG00000108395 UniProt: NA OMIM: 605073	



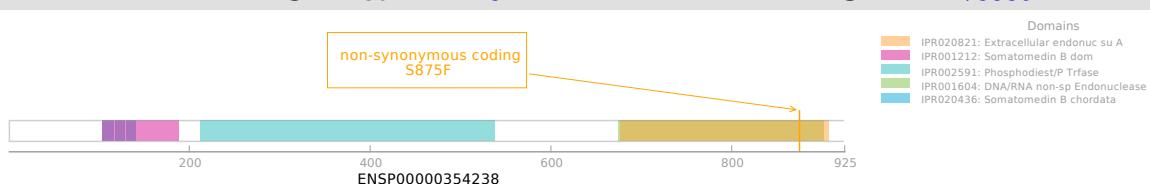
Variant 8:	Gene: SLC34A2 Your genotype: G/C Location: chr4:25675943	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0032	dbSNP: rs76404281
Quality:	Genotype quality: 99.00	Coverage depth: 67
Details:	Gene description: solute carrier family 34 (sodium phosphate), member 2 Transcript: ENST00000503434 AA change: L413F EntrezId: 10568 EnsemblId: ENSG00000157765 UniProt: NA OMIM: 604217	



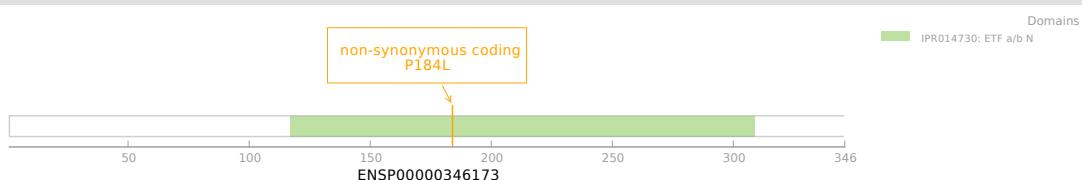
Variant 9:	Gene: LAMB3 Your genotype: A/G Location: chr1:209803199	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0045	dbSNP: rs52814161
Quality:	Genotype quality: 99.00	Coverage depth: 78
Details:	Gene description: laminin, beta 3 Transcript: ENST00000356082 EntrezId: 3914 UniProt: Q13751	AA change: Y339H EnsemblId: ENSG00000196878 OMIM: 150310



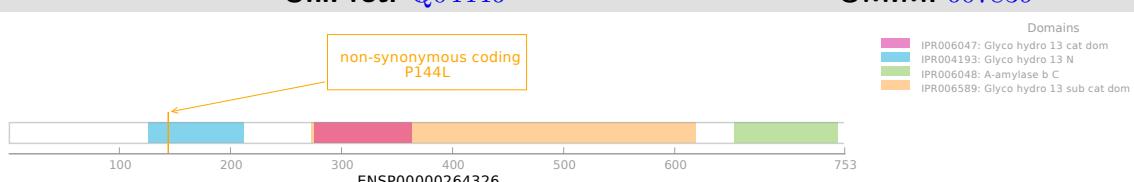
Variant 10:	Gene: ENPP1 Your genotype: C/T Location: chr6:132211497	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs140729669
Quality:	Genotype quality: 99.00	Coverage depth: 80
Details:	Gene description: ectonucleotide pyrophosphatase/phosphodiesterase 1 Transcript: ENST00000360971 EntrezId: 5167 UniProt: P22413	AA change: S875F EnsemblId: ENSG00000197594 OMIM: 173335



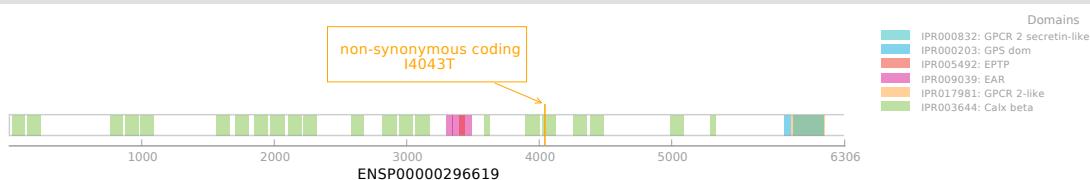
Variant 11:	Gene: ETFB Your genotype: G/A Location: chr19:51856483	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs139519507
Quality:	Genotype quality: 99.00	Coverage depth: 87
Details:	Gene description: electron-transfer-flavoprotein, beta polypeptide Transcript: ENST00000354232 AA change: P184L EntrezId: 2109 EnsemblId: ENSG00000105379 UniProt: NA OMIM: 130410	



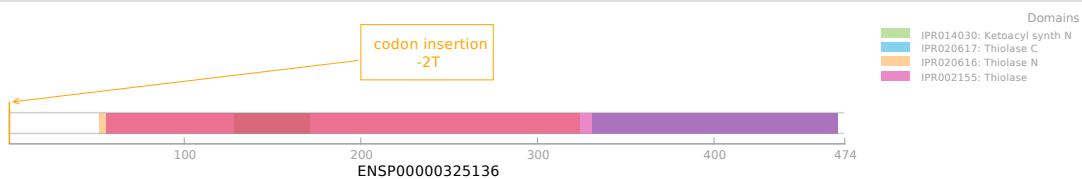
Variant 12:	Gene: GBE1 Your genotype: G/A Location: chr3:81754630	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: NA
Quality:	Genotype quality: 99.00	Coverage depth: 105
Details:	Gene description: glucan (1,4-alpha-), branching enzyme 1 Transcript: ENST00000264326 AA change: P144L EntrezId: 2632 EnsemblId: ENSG00000114480 UniProt: Q04446 OMIM: 607839	



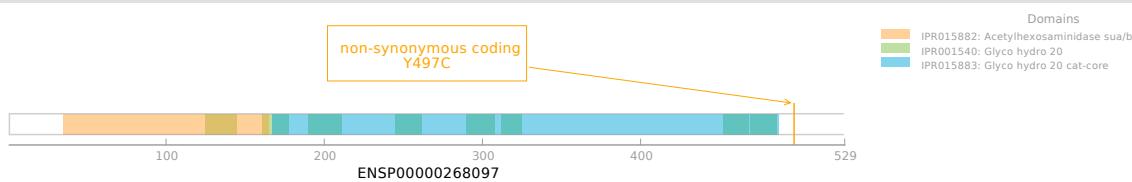
Variant 13:	Gene: GPR98 Your genotype: T/C Location: chr5:90059129	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: NA
Quality:	Genotype quality: 99.00	Coverage depth: 105
Details:	Gene description: G protein-coupled receptor 98 Transcript: ENST00000296619 EntrezId: 84059 UniProt: NA	AA change: I4043T EnsemblId: ENSG00000164199 OMIM: 602851



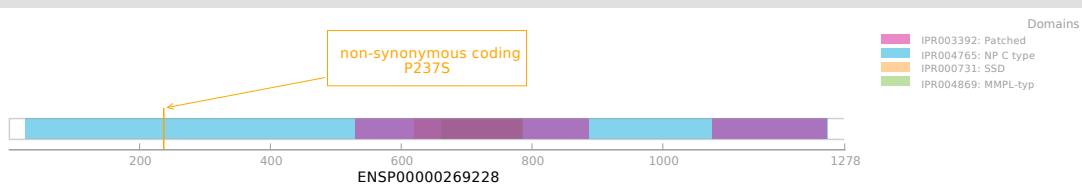
Variant 14:	Gene: HADHB Your genotype: GACT/GACT Location: chr2:26477125	
Effect:	CODON INSERTION	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs3839049
Quality:	Genotype quality: 99.00	Coverage depth: 97
Details:	Gene description: hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit Transcript: ENST00000317799 EntrezId: 3032 UniProt: P55084	AA change: -2T EnsemblId: ENSG00000138029 OMIM: 143450



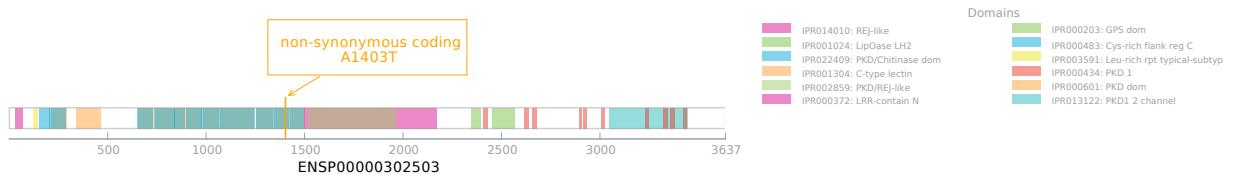
Variant 15:	Gene: HEXA Your genotype: T/C Location: chr15:72637823	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs147502219
Quality:	Genotype quality: 99.00	Coverage depth: 30
Details:	Gene description: hexosaminidase A (alpha polypeptide) Transcript: ENST00000268097 AA change: Y497C EntrezId: 3073 EnsemblId: ENSG00000213614 UniProt: P06865 OMIM: 606869	



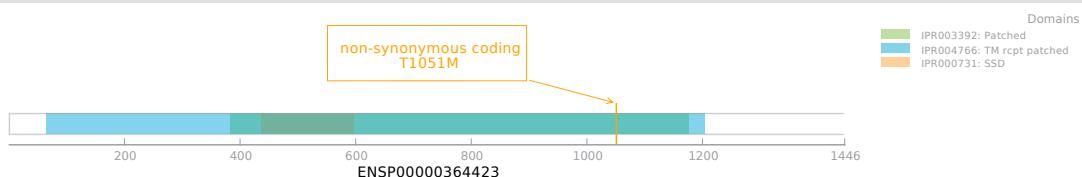
Variant 16:	Gene: NPC1 Your genotype: G/A Location: chr18:21140367	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0066	dbSNP: rs80358251
Quality:	Genotype quality: 99.00	Coverage depth: 29
Details:	Gene description: Niemann-Pick disease, type C1 Transcript: ENST00000269228 AA change: P237S EntrezId: 4864 EnsemblId: ENSG00000141458 UniProt: O15118 OMIM: 607623	



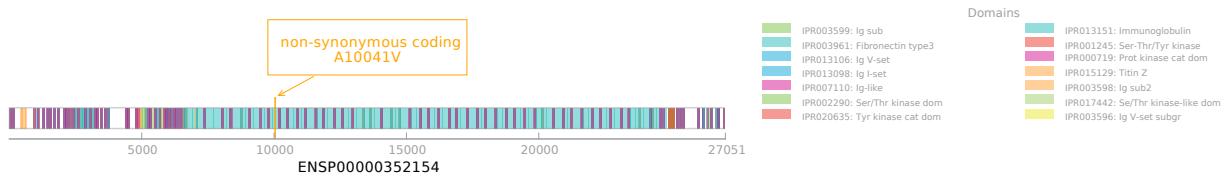
Variant 17:	Gene: PKD1 Your genotype: C/T Location: chr16:2159557	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs144137200
Quality:	Genotype quality: 99.00	Coverage depth: 14
Details:	Gene description: polycystic kidney disease 1 (autosomal dominant) Transcript: ENST00000306101 EntrezId: 5310 UniProt: NA	AA change: A1403T EnsemblId: ENSG00000008710 OMIM: 601313



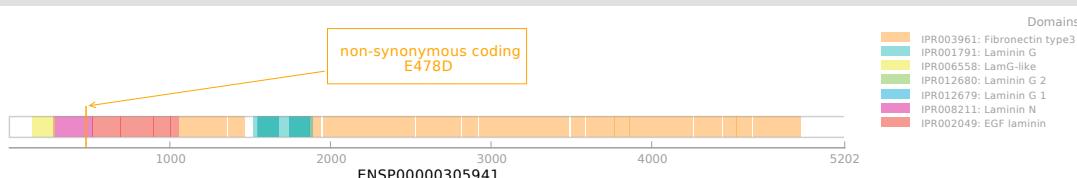
Variant 18:	Gene: PTCH1 Your genotype: G/A Location: chr9:98220308	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: rs138911275
Quality:	Genotype quality: 99.00	Coverage depth: 39
Details:	Gene description: patched 1 Transcript: ENST00000375274 EntrezId: 5727 UniProt: Q13635	AA change: T1051M EnsemblId: ENSG00000185920 OMIM: 601309



Variant 19:	Gene: TTN Your genotype: G/A Location: chr2:179463495	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: NA	dbSNP: NA
Quality:	Genotype quality: 99.00	Coverage depth: 246
Details:	Gene description: titin Transcript: ENST00000359218 EntrezId: 7273 UniProt: NA	AA change: A10041V EnsemblId: ENSG00000155657 OMIM: 188840



Variant 20:	Gene: USH2A Your genotype: C/G Location: chr1:216496932	
Effect:	NON-SYNONYMOUS CODING	Type: MODERATE
Frequency:	1KGenomes: 0.0073	dbSNP: rs35730265
Quality:	Genotype quality: 99.00	Coverage depth: 72
Details:	Gene description: Usher syndrome 2A (autosomal recessive, mild) Transcript: ENST00000307340 EntrezId: 7399 UniProt: O75445	AA change: E478D EnsemblId: ENSG00000042781 OMIM: 608400



Variant 21:	Gene: CASP10 Your genotype: G/A Location: chr2:202074098	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.0284	dbSNP: rs13010627
Quality:	Genotype quality: 99.00	Coverage depth: 82
Details:	Gene description: caspase 10, apoptosis-related cysteine peptidase Transcript: ENST00000313728 EntrezId: 843 UniProt: NA	
		EnsemblId: ENSG00000003400
		OMIM: 603909

Variant 22:	Gene: PPARA Your genotype: C/G Location: chr22:46614274	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.025	dbSNP: rs1800206
Quality:	Genotype quality: 99.00	Coverage depth: 96
Details:	Gene description: peroxisome proliferator-activated receptor alpha Transcript: ENST00000262735 EntrezId: 5465 UniProt: Q07869	
		EnsemblId: ENSG00000186951
		OMIM: 170998

Variant 23:	Gene: MC1R Your genotype: C/T Location: chr16:89986117	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.0322	dbSNP: rs1805007
Quality:	Genotype quality: 99.00	Coverage depth: 66
Details:	Gene description: melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor) Transcript: ENST00000555147 EntrezId: 4157 UniProt: NA	
		EnsemblId: ENSG00000258839
		OMIM: 155555

Variant 24:	Gene: KRT85 Your genotype: C/T Location: chr12:52760957	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.0266	dbSNP: rs61630004
Quality:	Genotype quality: 99.00	Coverage depth: 61
Details:	Gene description: keratin 85 Transcript: ENST00000257901 EntrezId: 3891 UniProt: P78386	EnsemblId: ENSG00000135443 OMIM: 602767

Variant 25:	Gene: CPN1 Your genotype: C/T Location: chr10:101829514	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.032	dbSNP: rs61751507
Quality:	Genotype quality: 99.00	Coverage depth: 63
Details:	Gene description: carboxypeptidase N, polypeptide 1 Transcript: ENST00000370418 EntrezId: 1369 UniProt: P15169	EnsemblId: ENSG00000120054 OMIM: 603103

Variant 26:	Gene: TPMT Your genotype: C/T Location: chr6:18139228	
dbSNP:	Annotation Present	
Frequency:	1KGenomes: 0.0174	dbSNP: rs1800460
Quality:	Genotype quality: 99.00	Coverage depth: 102
Details:	Gene description: thiopurine S-methyltransferase Transcript: ENST00000309983 EntrezId: 7172 UniProt: P51580	EnsemblId: ENSG00000137364 OMIM: 187680

Variant 27:	Gene: TPMT Your genotype: T/C Location: chr6:18130918
dbSNP:	Annotation Present
Frequency:	1KGenomes: 0.0462
Quality:	Genotype quality: 99.00
Details:	Gene description: thiopurine S-methyltransferase Transcript: ENST00000309983 EntrezId: 7172 UniProt: P51580
	dbSNP: rs1142345
	Coverage depth: 65
	EnsemblId: ENSG00000137364
	OMIM: 187680

Variant 28:	Gene: SCN9A Your genotype: T/C Location: chr2:167168083
dbSNP:	Annotation Present
Frequency:	1KGenomes: NA
Quality:	Genotype quality: 99.00
Details:	Gene description: sodium channel, voltage-gated, type IX, alpha subunit Transcript: ENST00000303354 EntrezId: 6335 UniProt: NA
	dbSNP: rs121908920
	Coverage depth: 166
	EnsemblId: ENSG00000169432
	OMIM: 603415

Appendix

To create the final draft of your exome we added some additional steps from the Broad Institute's "[Best Practice](#)" protocol aimed at increasing both the sensitivity and specificity of the variant calls returned to you. In the description that follows, steps 1–5 are unchanged from your first report:

1. We took your raw reads and aligned them against the reference genome (these are the alignments available in the BAM file of the first 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. At this point the protocol begins to differ from that used to generate the first draft of your exome. We now generate allele calls for all exome pilot participants simultaneously. By integrating data from multiple individuals we can more accurately detect i) variants in low coverage regions and ii) signatures of technical artifacts that might lead to incorrect variant calls. In addition we generate a BED file of all confidently called positions in the genome, which can be used in conjunction with the VCF file to determine where you are likely to be homozygous for the allele represented in the reference genome.
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 a technique developed at the Broad Institute known as [Variant Quality Score Recalibration](#) (VQSR). On top of this we applied the following cutoffs: i) $GQ \geq 30$, ii) $DP \geq 10$, iii) variant not on one of the sex-chromosomes. Variants that pass all filters are marked in your VCF file with a PASS, those that fail a filter are marked with the filters that they failed.
8. We then use [snpEff](#) to predict the functional impact of each variant on each gene that it may affect. Note that due to the existence of alternative transcription start/end points and alternative splicing a variant can have different effects on different products of the same gene. To simplify analysis we used GATK to select the highest-impact effect for each variant (see [here](#) for details).
9. We also annotate each variant with its allele frequency in the 1000 Genome's Project if available.