Whole-Genome Sequencing Using Supercomputation to Identify Various Exonic Disease-Causing Mutations

Avni Gupta[1], Ashima Kamra[2]

[1] Wilton High School, Wilton, CT.
[2] Cate School, Carpinteria, CA.

Introduction

Personal Genomics is a newly emerging field within bioinformatics that utilizes an individual's genetic information to provide risk assessment and information regarding familial history and disease inheritance. A vital component of this technology is next generation sequencing, or NGS, a prominent method of analysis defined by its ability to “sequence DNA at unprecedented speed, thereby enabling previously unimaginable scientific achievements and novel biological applications" [1]. NGS’s speed and precision has made whole genome and exome sequencing more feasible than ever before. And with the accompanying decrease in cost and increase in the speed of genome sequencing, personal genomics and sequencing one’s own genome has become increasingly more common [2].

Some companies, like 23andMe and AncestryDNA, readily offer sequencing services, capitalizing on the recent wave of interest in understanding personal history through DNA [3]. But even outside of mainstream culture trends, personal genomics can serve as a valuable component in precision clinical medicine. Providing specific information about an diseased individual’s mutated exomes is an effective starting point for developing specific treatment plans that cater to the individual’s exact needs, rather than reverting to a plan generalized for a diverse population [4].

Personal genomics and NGS most commonly (at least in the public eye) rely on a full genome; however, these techniques can also be applied to only the exome. Whole-exome sequencing (WES) is very similar to other methods of DNA analysis but it differs in that it focuses on the 1-2% of the genome that codes for proteins. As approximately 85% of malignant genetic variants are located in the exomes and most of the comprehensible genetic information is stored in those regions, WES is a much more efficient, direct method of the DNA sequencing than whole-genome sequencing. The only difference in method lies in the addition of target enrichment step that is used to capture the exomes specifically and wash the introns and other non-exonic regions out [5].

This research project relies on whole exome sequencing to process and analyze its resulting data with the help of the University of Chicago’s supercomputer Midway. We will annotate large samples of exomic data from the 1000 Genomes Project, an international effort that resulted in a catalogue of common human genetic variations, in order to identify disease-relevant variants within our samples.

Methods

We performed the initial alignment on the University of Chicago’s supercomputer, Midway. We used the Burrows-Wheeler Aligner (BWA), a software to map sequences against a reference, such as the human reference genome, as well as samtools, a package to manipulate DNA sequence read alignments. Using “bwa aln”, we found the SA coordinates of the input fastq reads and generated two suffix array-index (sai) files [7]. We then used “bwa sampe,” a command that generates alignments, in order to merge the files and find the best alignment for the paired end reads [7]. This produced a binary alignment map (bam) file that we sorted using “samtools sort” [8].

After generating the sorted file, we genotyped it to search for single-nucleotide variants (SNVs) from the human reference genome. Using “samtools index” to index the file [9] and “samtools mpileup” to generate the text output [10], we created a flt.vcf file listing all SNVs,

along with their location and Phred score. Since variants with scores under 50 are generally considered unreliable in their accuracy, we filtered out these low-quality variants and transferred the high-quality variants to a new flt.hq.vcf file.

We then annotated the file using several databases: snp135 for the rs ID of SNPs in the NCBI dbSNP database; gwas catalog and clinvar_20170905 to check if the region around an SNPs has been associated with any diseases or conditions; and tljb2_all, which applies annotations from the dbNSFP database, which compiles different functional prediction metrics in order to determine how dangerous a mutation might be. From the resulting annotation, we filtered out first exonic variants, as these have the most effect on phenotypic expression.

From these exonic, high-quality variants, we then selected several for further analysis. These were selected across the genome, with at least one coming from each chromosome. Selected mutations were associated with some kind of condition or disease, using the dbSNP database to ensure there was enough information on the variant. Since there were far more variants that fit these criteria than could feasibly be discussed, there was a level of random choice involved as well. Once the variants were chosen, we used NCBI's dbSNP, ClinVar, and MedGen databases to further investigate them.

Results

In total, we identified 161016 variants. Of these, 102932 were high-quality (Phred score greater than 50). Within the high-quality variants, 16330 were exonic, 6594 were nonsynonymous, and 7993 were synonymous. There were 95593 single nucleotide polymorphisms and 7338 insertions, deletions, and substitutions (indels). Of the indels, 52 caused frameshift mutations while 132 did not. Of the 65 nonsense mutations, 56 inserted a premature stop codon and nine deleted a stop codon.

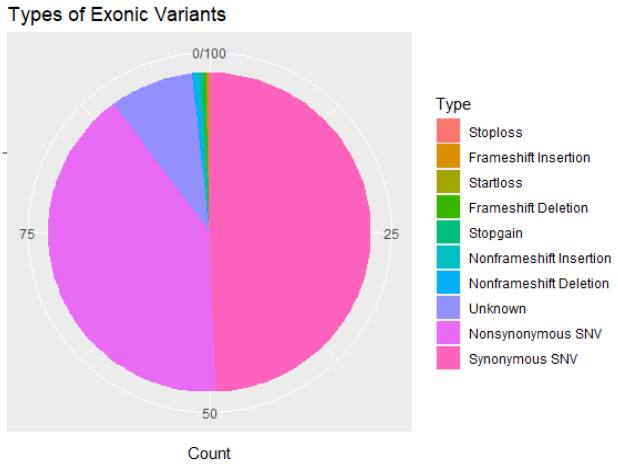

***Fig 1**: Types of mutations located in exonic regions*

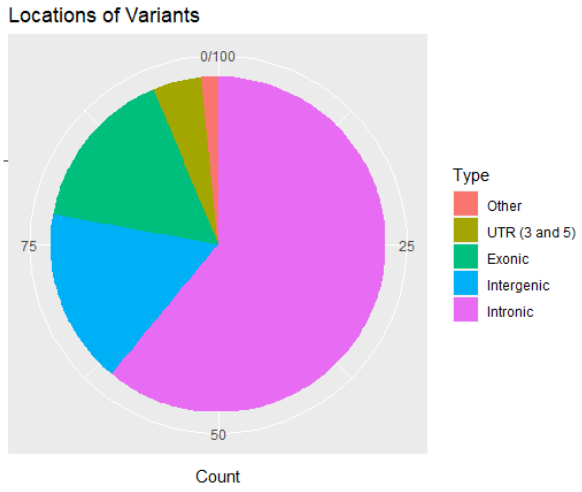

***Fig 2:** Location distribution of mutations in the genome*

From these mutations, we found a nonsynonymous SNV on chromosome 5 at position 256404. This SNV, located in gene SDHA, changed a cytosine nucleotide to a guanine, causing an alanine amino acid to switch to a glycine [11]. This mutation has been previously documented (rsID rs191412461), and is associated with a Mitochondrial Complex II Deficiency, a condition which weakens the capacity of the second mitochondrial complex to reduce FAD to $FADH_2$, and thus to reduce ubiquinone to ubiquinol in the respiratory chain. This condition manifests in a variety of ways, from muscle and cardiovascular deficiencies to full-body involvement [12]. This mutation also can cause Paragangliomas 5, a condition characterized by tumors along the spine from the base of the skull to the pelvis, and pheochromocytomas, which are a kind of paraganglioma confined to the adrenal medulla within the kidney [13]. This mutation is very uncommon, with all demographics expressing the alternative allele with a frequency of under 0.1% [11].

The second mutation that caught our interest was on position 47954148 on chromosome 13. Another nonsynonymous SNV in the SUCLA2 gene changed a G to an A, causing Aspartic Acid to change to Tyrosine [14]. This mutation has been previously documented (rsID rs117412559), and causes Mitochondrial DNA depletion syndrome. This condition generally manifests in infancy or early childhood, with symptoms including psychomotor retardation, hypotonia, muscular atrophy, feeding difficulties, and less frequently, contractures and epilepsy. Individuals with this condition have a median survival age of 20 years [15]. 0.119% of the general population displays this mutation, though expression rates are somewhat higher for other and east Asians (2.9% and 3.6%, respectively) [14].

The third mutation of interest was a frameshift insertion at position 44876705 on chromosome 18. This occurred in gene SETBP1 and inserted the DNA sequence “TCTT” in its location, changing a Threonine amino acid to a Serine (and affecting all following amino acids) [16]. This was the only frameshift mutation associated with an identified syndrome in our experiment (rsID rs3085861) - Schinzel-Giedion Syndrome. This condition involved distinctive facial features, neurological problems, and organ and bone abnormalities. The condition usually manifests at birth and causes severe developmental delays and heart defects. Due to the severity of the affected individuals’ health problems, most children with this syndrome die in infancy. Those that survive have an increased risk of developing neuroepithelial tumors (a kind of brain tumor) [17]. This is a very common mutation, occurring in 53.84% of the general population. Non-south or -east Asians display the highest frequency of 62.5%, while south Asians display the lowest of 40.5% [16].

Our last mutation of interest was a synonymous SNV at position 32798541 on chromosome 20. This mutation (rsID rs6058891), in the DNMT3B gene, changed a T to a C, which does not change the amino acid (Cystonine) [18[. However, this mutation still causes Immunodeficiency-centromeric instability-facial anomalies syndrome 1, a condition characterized by facial dysmorphism and immunoglobulin deficiency. This syndrome also causes chromosomes 1, 9, and 1 to branch after phytohemagglutinin stimulation of lymphocytes [19]. This is an extremely common mutation, manifesting in 47.35% of individuals. 99.8% of east Asians display this mutation, while only 43.41% of Europeans do [18].

Variant Location	Variant Type	Gene	Implications
Chromosome 1, Position 11787392	Nonsynonymous SNV	C1orf167	Neural tube defects (folate sensitive)

Chromosome 2, Position 26232259	Synonymous SNV	HADHA	LCHAD Deficiency, Mitochondrial trifunctional protein deficiency
Chromosome 3, Position 12503330	Nonsynonymous SNV	TSEN2	Pontoneocerebellar Hypoplasia
Chromosome 4, Position 5783715	Nonsynonymous SNV	EVC	Ellis-vanCreveld Syndrome, Curry-Hall Syndrome
Chromosome 5, Position 256404	Nonsynonymous SNV	SDHA	Mitochondrial complex II deficiency, Paragangliomas 5
Chromosome 6, Position 35455885	Synonymous SNV	FANCE	Fanconi Anemia
Chromosome 7, Position 21739660	Synonymous SNV	DNAH11	Primary Ciliary Dyskinesia
Chromosome 8, Position 10610079	Nonsynonymous SNV	RP1L1	Occult macular dystrophy
Chromosome 9, Position 441952	Synonymous SNV	DOCK8	Hyperimmunoglobulin E recurrent infection syndrome, Hyper-IgE syndrome
Chromosome 10, Position 49459059	Nonsynonymous SNV	ERCC6	Cerebro-Oculo-Facio-Skeletal Syndrome, Macular degeneration, Cockayne syndrome
Chromosome 11, Position 18297633	Nonsynonymous SNV	HPS5	Hermansky-Pudlak syndrome
Chromosome 12, Position 40320099	Nonsynonymous SNV	LRRK2	Parkinson’s Disease
Chromosome 13, Position 47954148	Nonsynonymous SNV	SUCLA2	Mitochondrial DNA depletion syndrome
Chromosome 14, Position 64137877	Nonsynonymous SNV	SYNE2	Emery-Dreifuss muscular dystrophy
Chromosome 15, Position 45111868	Nonsynonymous SNV	DUOX2	Congenital hypothyroidism
Chromosome 16, Position 1453878	Synonymous SNV	CLCN7	Osteopetrosis
Chromosome 17, Position 6461433	Synonymous SNV	PITPNM3	Cone-Rod Dystrophy
Chromosome 18, Position 44876705	Frameshift Insertion	SETBP1	Schinzel-Giedion Syndrome
Chromosome 19, Position 6495725	Synonymous SNV	TUBB4A	Leukodystrophy, (hypomyelinating, 6), Dystonia
Chromosome 20, Position 32798541	Synonymous SNV	DNMT3B	Immunodeficiency-centromeric instability-facial anomalies syndrome 1 (ICF1)
Chromosome 21, Position 46357148	Nonsynonymous SNV	PCNT	Microcephalic Osteodysplastic Primordial Dwarfism
Chromosome 22, Position 41152004	Nonsynonymous SNV	EP300	Rubinstein-Taybi Syndrome
Chromosome X, Position 47203135	Nonsynonymous SNV	UBA1	Arthrogryposis multiplex congenita (distal, X-linked)

Table 1*: Disease-causing variants on each chromosome*

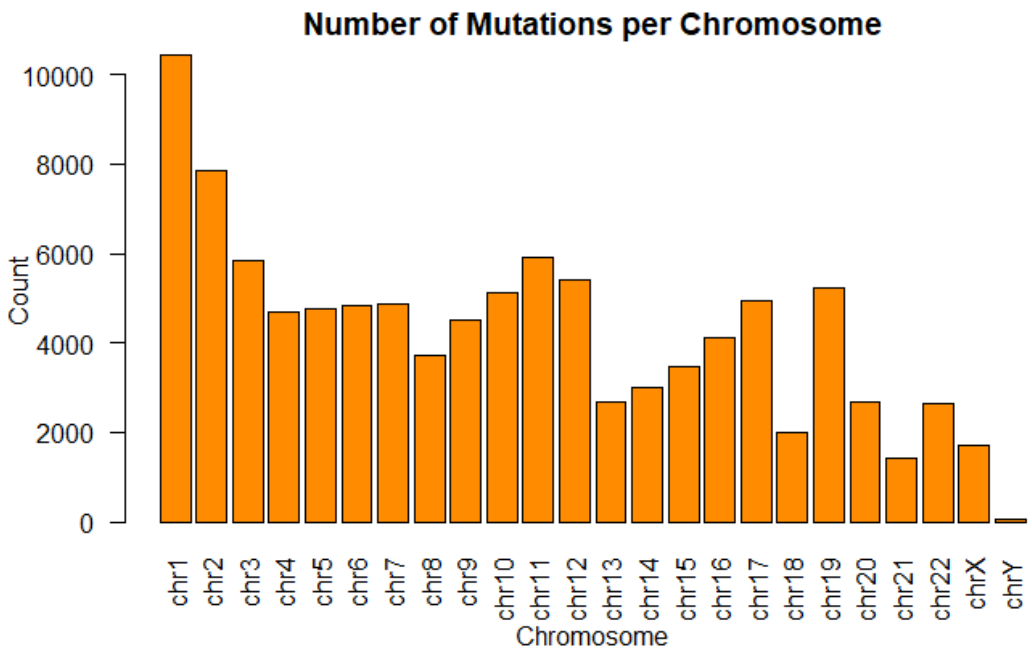

Figure 3: *Number of mutations on each chromosome*

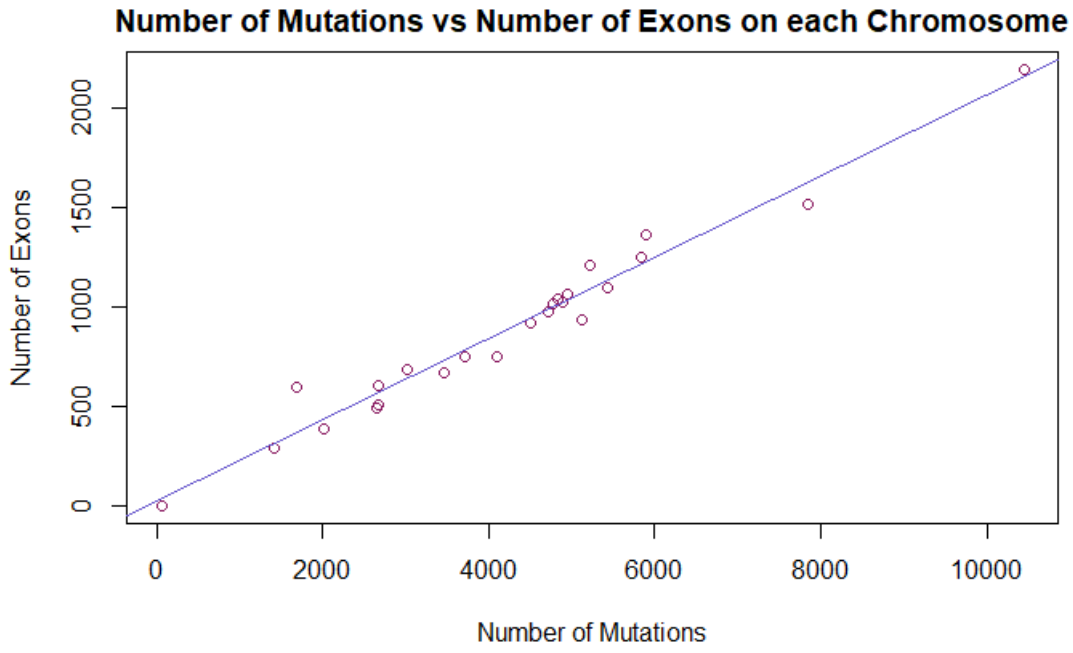

Figure 4: *Relationship between the number of mutations and the number of exons on each chromosome*

The relationship between mutation and exon quantity on each chromosome is extremely proportional. With an r-squared value of 0.9665, the regression line is a very good fit, showing that the relationship between the number of exons and the number of mutations on each chromosome is almost linear - there is a constant rate of mutations per exon.

Discussion

The mutation rsID rs3085861 proves to be more intriguing when further investigated. It is located on the gene SETBP1, a relatively unknown gene that codes for a set binding protein. These proteins attach to specific regions of the DNA in order to amplify or suppress gene expression and are also involved in DNA replication. Its mechanistic role in disease is unknown; however, we can infer from the protein's role in gene expression that the disease may be caused by a combination of over-suppression or over-amplification of a phenotype leading to the distinctive facial features and higher-than-normal prevalence of tumors that characterize the disease.

One 2010 study by researchers in the Netherlands found that the mutations that cause Schinzel-Giedion Syndrome may have resulted in a gain-of-function effect or a dominant negative effect [20]. This is due to mutation clustering in SETBP1 - a phenomenon also seen for other syndromes and genes that are described by a gain-of-function mechanism.

The specific protein associated with the gene is unknown, as mentioned previously, but a study done in 2017 studied the effects of the genetic variants of SETBP1 protein stability and found that the gain-of-function mutations do, indeed, affect a degron, a protein that is important in regulation of protein degradation rate, which in turn affects the stability of the SETBP1 protein structure [21].

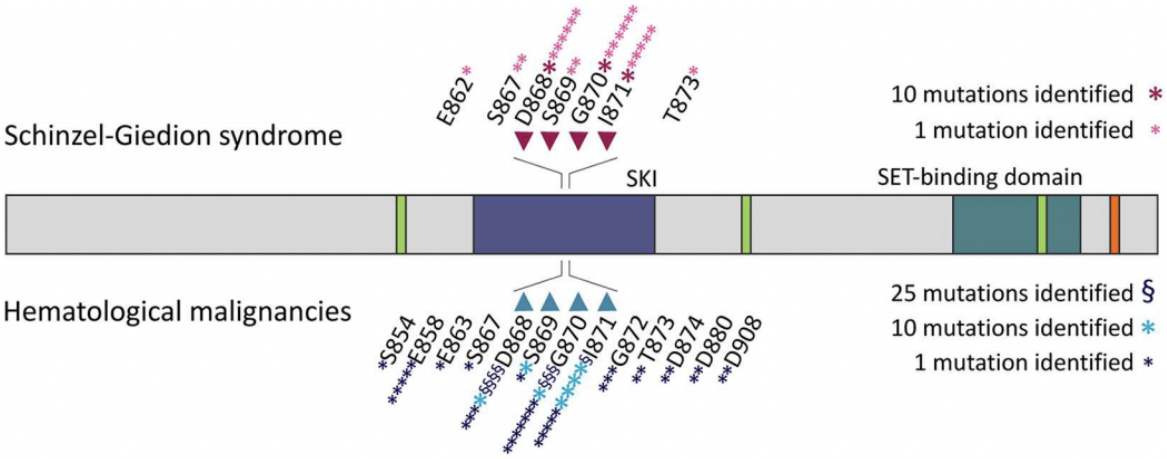

Figure 5: *Representation of the SETBP1 protein, indicating changes found in the disease and in hematologic malignancies. The residues of the degron are the bolded arrows [21].*

The 2010 study mentioned that exome sequencing limited their experimental approach as it was unable to identify any specific structural genomic variation associated with SGS. This caveat is also apparent in the 2017 study, as they were not able to produce any data relating to the specific structural changes that the SETBP1 protein underwent.

This brings up the point that exome sequencing is not ideal for all situations. Although its speed and cost are important factors in deciding whether to use it, exome sequencing may not be ideal if dealing with the genome of an unknown species. WES can only find variants when they are in a part of the genome that we are familiar with. Thus, if a mutation is located in an area not recognized as an exome, even if the mutation has a major impact of a phenotype, it will never be detected [22]. But the genetic information collected by WES helps us greatly to understand the

basis for disease and the evolution of viruses over time. Especially during the COVID-19 pandemic, it was vital to be able to recognize core components of each variant disease within various countries to develop a universal vaccine that can protect individuals against the adapted forms of the coronavirus. We could use this technology to predict various diseases by knowing which mutations are more likely to occur and which we would be probable to see in the near future. As a result, the scientific community would be able to prepare effective treatments before the threat even appears.

Conclusion

To summarize, after utilizing the University of Chicago's supercomputer Midway2, we were able to generate a list of exomes aligned against a human reference sequence. From that list, we determined four exonic variants of interest on chromosomes 5, 13, 18 and 20. One of these mutations was determined to be a gain-of-function mutation vital to the development of Schinzel-Giedion Syndrome. Afterwards, we recognized some of the limitations of exome sequencing to be its inability to analyze unknown genome or exome sequences. However, we also understood the future benefits of this technology as well as its potential to prepare defenses against unencountered diseases or viruses.

References

[1] Zhang J, Chiodini R, Badr A et al. 2011. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics 38 (3).

[2] Wetterstrand, K. 2020, "The Cost of Sequencing a Human Genome." National Human Genome Research Institute.
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

[3] Mountain J. 2013. Personal Genomics. Genomic and Personalized Medicine

[4] Kaur J, Rahat B, Thakur S et al. 2017. Trends in Precision Medicine. Progress and Challenges in Precision Medicine.

[5] Seaby E, Pengelly R, Ennis S. 2016. Exome sequencing explained: a practical guide to its clinical application. Briefings in Functional Genomics 15 (5).

[7] 2013. Manual Reference Pages - bwa (1). http://bio-bwa.sourceforge.net/bwa.shtml

[8] 2021. Samtools-Sort(1) Manual page. http://www.htslib.org/doc/samtools-sort.html

[9] 2021. Samtools-Index(1) Manual page. http://www.htslib.org/doc/samtools-index.html

[10] 2021. Samtools-Mpileup(1) Manual page. http://www.htslib.org/doc/samtools-mpileup.html

[11] 2021. Reference SNP (rs) Report: rs191412461.
https://www.ncbi.nlm.nih.gov/snp/rs191412461

[12] 2021. Mitochondrial Complex Ii Deficiency, Nuclear Type 1; Mc2dn1. https://www.omim.org/entry/252011

[13] 2021. Paragangliomas 5. https://www.ncbi.nlm.nih.gov/medgen/481622

[14] 2021. Reference SNP (rs) Report: rs117412559. https://www.ncbi.nlm.nih.gov/snp/rs117412559

[15] 2021. Mitochondrial DNA depletion syndrome 5 (encephalomyopathic with or without methylmalonic aciduria). https://www.ncbi.nlm.nih.gov/medgen/C2749864

[16] 2021. Reference SNP (rs) Report: rs3085861. https://www.ncbi.nlm.nih.gov/snp/rs3085861

[17] 2020. Schinzel-Giedion syndrome. https://www.ncbi.nlm.nih.gov/medgen/?term=Schinzel-Giedion+Syndrome

[18] 2021. Reference SNP (rs) Report: rs6058891. https://www.ncbi.nlm.nih.gov/snp/rs6058891

[19] 2021. Immunodeficiency-centromeric instability-facial anomalies syndrome 1(ICF1) https://www.ncbi.nlm.nih.gov/medgen/?term=Immunodeficiency-centromeric+instability-facial+anomalies+syndrome+1

[20] Hoischen A, Van Bon B, Gilissen C, et a.l. 2010. De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nature Genetics, 42(6),* https://www.nature.com/articles/ng.581.pdf

[21] Acuna-Hidalgo R, Deriziotis P, Steehouwer M, et al. 2017. Overlapping SETBP1 gain-of-function mutations in Schinzel-Giedion syndrome and hematologic malignancies. *PLoS Genet 13(3)*, https://doi.org/10.1371/journal.pgen.1006683

[22] Biesecker, L. G; Shianna, K. V; & Mullikin, J. C. 2011. Exome sequencing: the expert view. *Genome biology*, *12*(9), 128. https://doi.org/10.1186/gb-2011-12-9-128