NGS Analysis of Sequence 27 and Its Implications

Jin Jun Li

Introduction

The advent of personal genomics allowed scientists and doctors to discover connections between human diseases and genetic expressions. By sequencing and analyzing the DNA sequences of individuals, scientists were not only able to better understand how genes affect human health, but also be able to use it to improve medical care and extend people's lives. For example, personal genomics had been used to reveal various genetic diseases such as sickle cell anemia before the disease becomes fatal. And as advances in technology make genetic sequencing more available to consumers, personal genomics is becoming increasingly influential to our daily lives.

Next-generation sequencing (NGS for short) was a major key for revolutionizing personal genomics. Before NGS was introduced, scientists used the Sanger method to sequence genes. Although this technique was highly accurate, it was not only very expensive but also extremely slow (taking almost a decade for the Human Genome Project), making the Sanger method unready for the common consumer. But after the NGS technique was introduced, scientists were able to sequence entire human genomes under a single week. In addition, an NGS platform can make 150 million reads for about 1000 Euros, compared to about 1 Euro for a single read in Sanger's method.[15] In addition, the cost of sequencing continues to trend downwards as scientists develop more cost-effective ways of sequencing genes. This revolutionized the field of personal genomics, as sequencing become a viable medical solution to not only search for diseases but also help find appropriate medicine to combat the diseases.

NGS was not the only innovation that made personal genomics more accessible to the public; exome technology also increased the efficiency of genetic sequencing. In the DNA of an organism, genes fall under two categories: exons and introns. However, the introns in DNA are not expressed, meaning that we can find the most important sections of genetic expression by only searching for and sequencing the exons. Because the human genome is about 98-99% introns, capturing only the exons dramatically increase the speed of personal genomics. In the exome sequencing process, the DNA sequence is first fragmented into pieces and hybridized to a microarray that contains oligonucleotides (that match the exons of interest). The non-hybridized DNA is removed, and NGS is performed on the remain genes.[2]

In this paper, we will use similar techniques to sequence and analyze the genes of an sequence 27 from the 1000 Genomes Project. We first performed alignment, which refers to arranging the sequences of DNA or protein to prepare for genotyping, which is the process of determining the differences of sequence 27 with respect to the reference human genome. In this

process, we gather single nucleotide polymorphisms (SNPs) and insertion-deletion events (indels). Next, we annotate the variants, which refers to labeling each significant variant with appropriate information. And lastly, we picked and discussed a certain disease that a certain variant may indicate.

Methods

We first started with two files of pair-end 75 bp sequence reads of sequence 27, and we used the Burrows-Wheeler alignment (BWA) algorithm to align our sequence to the reference human genome. We use the distances between each pair read to increase the efficiency of the alignment algorithm. To perform this alignment, we used "bwa aln" program to align the reads from each paired-end file to the human reference genome. For each file, we set the read trimming parameter (the flag -q) to 5 and the number of threads parameter (the flag -t) to 28. Then we used the "bwa sampe" and "samtools" programs to determine the best alignment position for each pair of paired-end reads based on the genomic locations of the possible alignment positions from each individual read. For the "bwa sampe" parameter, we used the -P flag, indicating that we loaded the entire FM-index into memory to reduce disk operations. Finally, we sort the aligned sequences according to the genomic coordinates of the human genome to which the sequences are aligned to facilitate genotyping and variant calling. We used "samtools sort" with the parameters (-@) set to 28 (which set the number of sorting and compression threads.) and (-m) set to 1500M (the maximum required memory per thread).

Next, we performed variant calling, which refers to finding differences in SNPs and indels in sequence 27 to the reference human genome. To do this, we calculated the probability of observing a given combination of nucleotide reads at a particular site under every possible genotype and weigh our calculated likelihoods by known allele frequencies of previously described SNPs. Then we filter the results by QC parameters to ensure that variants in well-aligned regions are retained for further analysis. In the first step, we used the samtools mpileup program to create a raw set of genotype calls. Then we filter the genotypes according to the read depth. Furthermore, we filter the variants by the Phred Score of 225 to make sure we only annotate variations that are high quality and not very likely to be annotated incorrectly. And lastly, we filtered out all variants that are not exonic, leaving us with 960 variants that are worth annotating.

Now that we have the variants, we performed annotation to label the genomic type using ANNOVAR. The databases that we chose for annotation are "refGene" (to annotation refGene of the affected genes), "avsnp150" (for the dbSNP rs ID in the NCBI database), "clinvar_20170905" (to indicate whether the genetic locus near the given SNP is associated with any diseases in prior genome-wide association studies), and "dbnsfp33a" (to apply annotations

from the dbNSFP database to provide various metrics indicating how deleterious a particular nucleotide mutation is likely to be).

And lastly, we picked 20 variants out of the 960 exons with Phred Score of over 225 that are known to be linked to certain diseases using the clinvar_20170905 database.

Results

After performing the analysis, we identified 176392 variant, of which 2799 are high quality (with a Phred Score of over 225). Of those that are high quality, 960 were exonic, 2653 were SNPs, and 146 were indels. Of the high-quality exons, we also identified 512 synonymous variants, 375 non-synonymous variants, 3 frameshift deletions and insertions, and 4 premature stop codons. Furthermore, we found that only rs6050 and rs8012 is associated with diseases in GWAS (Genome-wide association study).

Location of Variant	**RS ID**	**Type of Variant**	**Implications**
Chromosome 1, Position 156876441	rs6334	synonymous SNV	Hereditary insensitivity to pain with Anhidrosis
Chromosome 2, Position 214767531	rs2070094	nonsynonymous SNV	Hereditary cancer predisposing syndrome, neoplasm of the breast
Chromosome 4, Position 154586438	rs6050	nonsynonymous SNV	Venous thromboembolism, Afibrinogenemia, congenital, Familial visceral amyloidosis, Ostertag type
Chromosome 6, Position 159692840	rs4880	nonsynonymous SNV	Superoxide dismutase 2 polymorphism, Microvascular complications of diabetes 6, cyclophosphamide response efficacy
Chromosome 7, Position 144401899	rs727714	synonymous SNV	Premature ovarian failure
Chromosome 8, Position 10607375	rs56382513	synonymous SNV	Occult macular dystrophy
Chromosome 8,	rs4538	synonymous	Hyperaldosteronism, familial, type I,

Position 142913286		SNV	Corticosterone methyloxidase type 2 deficiency, Corticosterone methyloxidase type 1 deficiency
Chromosome 9, Position 113391620	rs1139488	synonymous SNV	Porphobilinogen synthase deficiency
Chromosome 10, Position 16876973	rs1801241	synonymous SNV	Megaloblastic anemia
Chromosome 10, Position 121479598	rs1047057	synonymous SNV	Jackson-Weiss syndrome, Isolated coronal synostosis, Saethre-Chotzen syndrome, Craniosynostosis, Crouzon syndrome, Pfeiffer syndrome, Levy-Hollister syndrome, Acrocephalosyndactyly type I, Cutis Gyrata syndrome of Beare and Stevenson
Chromosome 11, Position 124919859	rs3802904	nonsynonymous SNV	Megalencephalic leukoencephalopathy with subcortical cysts
Chromosome 12, Position 47844974	rs731236	synonymous SNV	Vitamin D-Dependent Rickets
Chromosome 15, Position 65201874	rs2073711	nonsynonymous SNV	Lumbar Disc Disease
Chromosome 16, Position 8750532	rs2229157	synonymous SNV	Gamma-aminobutyric acid transaminase deficiency
Chromosome 17, Position 61456507	rs3744448	nonsynonymous SNV	Ischiopatellar dysplasia
Chromosome 18, Position 75287404	rs55679337	nonsynonymous SNV	Aural atresia, congenital
Chromosome 19, Position 12899706	rs8012	nonsynonymous SNV	Glutaric aciduria, type 1, Glutaric acidemia
Chromosome 20, Position	rs3746682	synonymous SNV	Hypogonadism with anosmia

5302610			
Chromosome 22, Position 20859128	rs1061064	synonymous SNV	Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome
Chromosome 22, Position 50625611	rs743616	nonsynonymous SNV	Metachromatic leukodystrophy

On Chromosome 2 at Position 214767531 we identified a nonsynonymous SNP in gene BARD1 that changed the amino acid from C to T. This mutation is associated with hereditary cancer predisposing syndrome and neoplasm of the breast. For hereditary cancer predisposing syndrome, the patient is likely to develop tumors in a young age and are more likely to develop cancer.[13] Neoplasm of the breast indicates that there might be tumors developing in the breast, but there are likely to be benign. In addition, tissues involving milk production (like ductal and lobular tissues) are likely to be affected.[18] This mutation has an RS ID of rs2070094 and it has been evaluated at on November 18, 2014.[10] In addition, Asians and Sub-Saharan Africans have high allele frequencies for this mutation.

On Chromosome 7 at Position 144401899 we identified a synonymous SNP in gene NOBOX that changed the amino acid from G to A. This mutation is associated with Premature ovarian failure, and some symptoms include loss of normal functions of the ovaries before 40 years of age. People with this disease don't produce normal amounts of the hormone estrogen or release eggs regularly. And some additional symptoms include irregular or skipped periods (amenorrhea), night sweats, vaginal dryness, and decreased sexual desire.[7] In addition, infertility is a common result.This mutation has an RS ID of rs727714, and has been discovered on June 14, 2016.[12] In addition, Asians have high allele frequencies for this mutation.

On Chromosome 15 at Position 65201874 we identified a nonsynonymous SNP in gene CILP that changed the amino acid from A to G. This mutation is associated with Lumbar disc disease (LDD). Some symptoms of this disease include drying out of the spongy interior matrix of an intervertebral disc in the spine and low back pain and unilateral leg pain.[1] Some additional symptoms include loss of muscle strength and loss of touch sensation. This mutation has been discovered before in June 1, 2005, and it has a RS ID of rs2073711.[11] In addition, Sub-Saharan Africans and Asians have high allele frequencies for this mutation.

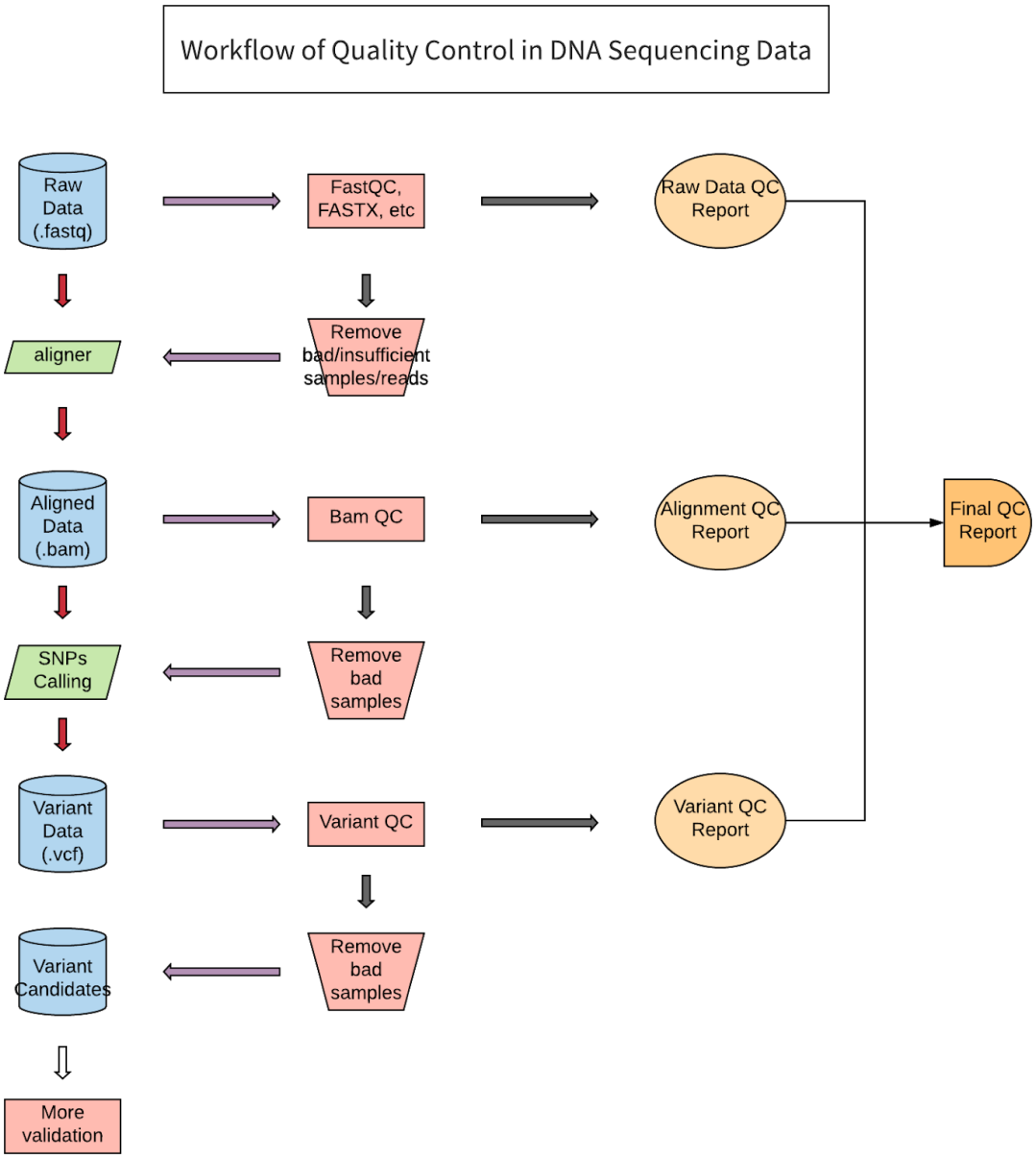
In a more thought out research paper, it would be useful for improving quality control throughout the data processing pipeline (raw data, alignment and variant calling).

Starting with quality control of the raw data, the most important thing to check for raw sequencing data quality are the base quality, the nucleotide distribution, and GC content distribution. We can use the FastQC package developed by the Babraham Institute bioinformatics group, which have parameters that deal with these quality control parameters.[19]

When using alignment algorithms, the base quality usually drops as the number of cycles increase. One way to address this problem is to include a trimming parameter to remove bases within a certain distance from the read ends. In addition, the nucleotide distribution is closely associated with the quality of the raw data, and we expect that the distribution of (A T C G) to remain stable across reads.[14] We can include parameters that filter out distributions that have high deviations. Furthermore, we can also filter out regions that have irregular GC contents, as those regions provide evidence of contamination.[19]

For alignment quality control, we can use more powerful exome sequencing capture kits, like Illumina TrueSeq, Agilent SureSelect, and NimbleGen SeqCap EZ.[19] Furthermore, we can also use tools like Picard and QPLOT for performing alignment quality control. In these tools, we can use parameters that ensure high capture efficiency, median depth, the percentage of the genome covered by the sequencing at that depth, and mapping quality.[3]

And lastly, we can include quality control of variant calling. For evaluating the SNP quality at a per sample level, we can use quality control parameters like transition/transversion (Ti/Tv) ratio (the number of transition SNPs divided by the number of transversion SNPs) and heterozygosity to non-reference homozygosity ratio. We can also use GATK, a DNA sequencing processing pipeline, to implement variant quality score recalibration based on SNP chip data to improve SNP quality called from exome sequencing.[3] For evaluating SNP quality at a per SNP level, we can include the parameters of depth (higher depth gives more statistical confidence to the SNP call), base quality, and mapping quality filters (to prevent bad reads from contributing). We can also use the Kolmogorov–Smirnov test to compare the SNP density difference, as high SNP frequency in a short region is an indication of false positives.[14]



Discussions

One of the more interesting variants is the nonsynonymous SNP in Chromosome 15, Position 65201874. This variant causes a mutation in gene CILP, changing the amino acid from Isoleucine (a hydrophobic amino acid) to Threonine (a hydrophilic amino acid).[16] This gene has a variety of functions, including encoding the cartilage intermediate layer protein, suppressing sulfated proteoglycan synthesis, and inhibiting ligand-induced IGF1R autophosphorylation. It is also known that CILP may inhibit TGF-β, a regulating growth factor that is important for maintaining extracellular matrix (ECM) proteins in intervertebral discs and is crucial for

metabolism of the intervertebral disc. Poor regulation of TGF-β signaling can cause various human connective tissue disorder, like LDD.[1]

In the paper by Shoji Seki, et al, researchers found that gene CILP encodes the cartilage intermediate layer protein, which affects LDD susceptibility.[16] This protein affects the intervertebral disc, which is composed of the annulus fibrosus (the outer layer) and the nucleus pulposus (the interior structure). The authors of the paper also found that CILP regulates TGF-β signaling, and the observed variant plays a crucial role in the cause of LDD, as individuals with that mutation are more likely to be affected by disc degeneration. The author of the paper drew this conclusion by examining candidate genes that cause LDD and finding that there was a statistical significance for an association between the SNP in CILP and LDD. To further strengthen the association, the author examined CILP expression in various human tissues and cells using real-time PCR and found that expression of CILP mRNA in intervertebral disc tissue from individuals increases as disc degeneration increases. The author then used linkage disequilibrium (LD) mapping to find the specific region where the variation in CILP causes LDD. The author was also able to rule out the possibility that the mutation may simply be a marker SNP with the true disease-associated alleles existing elsewhere in CILP by analyzing the haplotype structure of CILP.

Although there we found that there may be a strong possibility that the individual with sequence 27 may be linked to LDD, there are many caveats to consider. One thing to consider is that NGS is not completely flawless, as it may make mistakes on reads, thus reducing the accuracy of its sequencing. Furthermore, although BWA is one the most efficient algorithms, it can still produce errors and its performance is known to degrade for long reads, especially when the sequencing error rate is high.[6] In addition, even though there are statistical connections between genes and certain diseases, this does not necessarily imply that such connections exist. For certain individuals, a certain gene mutation may not end up affecting the individual at all, as genetic expression differs by a case by case basis.

Although exome sequencing is a powerful tool for understanding the human genome and providing a tool for revealing diseases, it has many drawbacks. In many instances, exome sequencing can have different interpretations, based on inconsistent replication results, differences in allele frequencies across populations, and probabilities of family segregation.[17] In addition, exome sequencing only targets known annotated exons, meaning much information may be lost during the process, as it may miss some genes that can encode for diseases. Furthermore, exome sequencing has shown to produce significant amounts of both false positive and false negative results for patients; associations between gene mutation and diseases can be ambiguous, and at some times, simply guesswork. Some scientists have a hard time replicating the findings and associations of previous works (for example, multiple reports for HABP2 rs7080536 and FNMTC produced both positive associations and no associations from reputable geneticists).[5] To add on, errors may have been introduced during NGS method (including during polymerase chain reaction (PCR) amplification).

There are many caveats of personal genomics in general. Consumer genetic testing has been known to cause privacy issues, as many companies require the individual to give consent for the company to use their genetic data. It has also called into question the intentions about how companies are using the data, as biology companies have been known to share data with third parties.[9] Furthermore, there are ethical concerns that geneticists are playing god by sequencing genes, as it challenges traditional values of religion and nature. In addition, there is also discussion on the implications of the future of DNA sequencing, as it may lead to artificial humans (who have their genes modified). And lastly, there are many legal problems to deal with, including who takes responsibility for damages that genetic sequencing may cause (as DNA sequencing has produced false results).[8]

But even with all these drawbacks, DNA sequencing may still be incredibly useful for patients. Knowing mutations may help warn individuals of possible diseases they might have in the future, which provides a huge heads up that can allow the individual to change their lifestyle to avoid the disease. In addition, DNA sequencing allows scientists to create drugs that specifically work for certain individuals, which may be incredibly helpful for diseases that have no known cure.[8]

Conclusions

In this paper, we used the BWA algorithm to align two files of sequence reads. We then performed variant calling and found 176392 variants. We used a Phred Score of 225 to filter out many of the variants, and we considered only the exons, resulting in 960 mutations that are worth analyzing. After performing annotations, we discussed the mutation in gene CILP and its role in Lumbar Disk Disease. Finally, we discussed the problems of personal genomics.

References

1) Alpesh A. Patel, William Ryan Spiker, et al. 2011. Evidence for an Inherited Predisposition to Lumbar Disc Disease. *J Bone Joint Surg Am.* 93(3): 225–229.
2) Amanda Warr, Christelle Robert, et al. 2015. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda)*. 5(8): 1543–1550.
3) Christoph Endrullat, Jörn Glökler, et al. 2016. Standardization and quality management in next-generation sequencing. Applied & Translational Genomics. 10:2-9.
4) Emilia Niemiec, Heidi Carmen Howard. 2016. Ethical issues in consumer genome sequencing: Use of consumers' samples and data. *Appl Transl Genom.* 8: 23–30.

5) Glenn S. Gerhard, Darrin V. Bann, et al. 2017. Pitfalls of exome sequencing: a case study of the attribution of HABP2 rs7080536 in familial non-medullary thyroid cancer. *Nature.* 8.
6) Heng Li, Richard Durbin. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 25(14): 1754–1760.
7) Katarzyna Jankowska. 2017. Premature ovarian failure. *Prz Menopauzalny*. 16(2): 51–56.
8) Kato K., Minari J. 2013. Ethical issues in personal genome research. *Brain Nerve.* 65(3):267-72.
9) Michael Snyder, Jiang Du, and Mark Gerstein. 2010. Personal genome sequencing: current approaches and challenges. *Genes & Dev.* 24: 423-431.
10) NCBI. "Reference SNP (RefSNP) Cluster Report: rs2070094. ** With Likely Benign Allele **." *Current Neurology and Neuroscience Reports., U.S. National Library of Medicine.*
11) NCBI. "Reference SNP (RefSNP) Cluster Report: rs2073711. With Other Allele **." *Current Neurology and Neuroscience Reports., U.S. National Library of Medicine.*
12) NCBI. "Reference SNP (RefSNP) Cluster Report: rs727714. ** With Benign Allele **." *Current Neurology and Neuroscience Reports., U.S. National Library of Medicine.*
13) Nils Rahner, Verena Steinke. 2008. Hereditary Cancer Syndromes. *Dtsch Arztebl Int.* 105(41): 706–714.
14) Qian Zhou, Xiaoquan Su, et al. 2013. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *Plos one.*
15) Sam Behjati, Patrick S Tarpey. 2013. What is next generation sequencing? *BMJ Journals.* 98(6): 236–238.
16) Seki S., Kawaguchi Y., et al. 2005. A functional SNP in CILP, encoding cartilage intermediate layer protein, is associated with susceptibility to lumbar disc disease. *Nat Genet.* 37(6):607-12.
17) Someswa Kesh, Wullianallur Raghupathi. 2004. Critical Issues in Bioinformatics and Computing. *Perspect Health Inf Manag.* 1: 9.
18) Xavier de Sousa, Pedro Santos Ferreira, et al. 2018. Neoplasm of uncertain behaviour of the breast—a retrospective study in a breast unit. *Ecancermedicalscience*. 12: 839.
19) Yan Guo, Fei Ye, et al. 2013. Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics.* 15(6): 879–889.