Genomics in medicine: Can we help House MD?

L. Glazar[*], A. Kosic [*], M. Milovanovic[*], J. Stancl[*] and D.Vucenovic[**]
[*] Summer School of Science, Požega, Croatia
[**] Faculty of Science, Department of biology, Zagreb, Croatia

After the Human genome project, genomics began to rapidly develop. Nowadays, scientists try to apply genomics in so many different ways. The project we worked on, as a part of the Summer School of Science 2015th in Gymnasium Požega. The project was about applying genomics in medicine. In the beginning we only had Data (patient's RNA seq and reference genome (NA12878)). After processing in UNIX terminal (Quality check, Alignment, Calling Variants) and visualizing in Integrative Genomics Viewer (IGV) we could see variations (deletions, insertions, inversions, tandem duplications, translocations and single-nucleotide polymorphisms (SNPs)) in our patient's genome. Thru them we could predict and say which diseases our patient has or might get. In this project we learned, for us, a lot of new things about genomics and proved that genomics and computing biology can be useful and used in medicine in diagnostics and predicting the percentage of diseases you could get.

I. Introduction

Genome, the research object of genomics, is the genetic material of an organism and consists of the complete information for protein regulation and production. In the end of the last century scientists begun to research it extensively. Human genome project enabled development of new generation sequencers and it determined composition of the whole human genome. Having that information created the new era of understanding nature and relations in it. From then we have more opportunities for research in biology, genomics, medicine, pharmacy, and so on.

In our project we're going to compare genome of our patient with reference human genome and see differences there: deletions, insertions, inversions, tandem duplications, translocations and single-nucleotide polymorphism (SNPs). Then we will try to find out what diseases our patient has.

II. Materials and methods

A. Data:

Because of limited time and computing power, we have decided to focus only on a small part of genome, chromosome 20 which represents only about 2% of genome. Reference genome we used is NA12878, Illumina's platinum genome which was sequenced with high precision and high coverage. We obtained RNA sequences from our imaginary patient. The reason why we used RNA sequences is because we decided to look only at actively transcribing part of genome.

B. Methods:

Obtaining read data
We were given RNA sequencing data from our imaginary patient in FASTQ format (format that along with sequence tells about the quality of sequencing each base in molecule) and reference genome (NA12878) in FASTA format. All operations we used were launched from UNIX terminal.

Quality Control of the reads
Before any further analysis we had to check for the quality of our reads and remove bad ones. Quality control usually involves several steps:

1. Obtaining summary quality statistics for the reads and reviewing diagnostic graphs
2. Filtering out genetic contaminants (primers, vectors, adaptors)
3. Trimming or filtering low-quality reads
4. Recalculating quality statistics and review diagnostic plots on filtered data

For this analysis we used fastqc program which outputs HTML file with all the details about quality of sequences. (Figure 1). After we obtained the results we trimmed the reads which very low quality in the end. For the trimming process we used cutadapt program which was trimming all the reads that had 3' quality scores lower than our specified threshold. By doing this we removed low quality reads so that only good quality reads are used for our project. After trimming we did the quality check again. (Figure 2) Also, out RNA sequences had some

overrepresented motifs (adaptors which are used in sequencing process), again for this process we used cutadapt. (Figure 3 and 4)

Read alignment

In order to find differences between our patient and healthy population, we have to align (determine original position in the genome) all the reads to the reference genome.

Alignment process consists of choosing an appropriate reference genome to map our reads against, and performing the read alignment using one of several alignment tools such as NovoAlign (which uses Burrows-Wheeler algorithm). For our example we will be using bwa-mem which has become one of the standard tools for aligning Illumina reads >75bp to large genomes.

Alignment results we visualized in IGV (figure 5) where we took from IGV's database information about which part of the chromosome is coding for genes and regulatory parts.

Marking duplicate reads

Duplicate reads originate mostly from library preparation methods (unless sequenced very deeply) and bias subsequent variant calling. We will be using samblaster for this step. In order to be called a 'duplicate' reads need to match on the sequence name, strand, and where the 5' end of the read would end up on the genomic sequence if the read is fully aligned.

Calling Variants

So far we have our sequence data, we cleaned up the read and aligned them to the genome. Which is to say we are now finally ready to find sequence variants, i.e., regions where the sequenced sample differs from the human reference genome. For this task we will use FreeBayes program which uses a Bayesian approach to identify SNPs, InDels and more complex events as long as they are shorter than individual reads. It is haplotype based, that is, it calls variants based on the reads aligned to a given genomic region, not on a genomic position. In brief, it looks at read alignments from an individual (or a group of individuals) to find the most likely combination of genotypes at each reference position and produces a variant call file (VCF). To annotate variation we found in our patient we used the data from various databases such as dbSNP, OMIM, ENCODE and HapMap. When our data was annotated we now had the information about the level of impact of our variants to human body sorted in three groups LOW, MEDIUM and HIGH. Initial plan was to look at all three groups but because of technical problems we could only look deeper into the HIGH impact group. We run the IDs of annotated data from that group on the already mentioned databases and found more information about the place on the chromosome where they are influencing the functions that are normally present there.

III. RESULTS

Quality check results are shown on Figure 1 and on Figure 2. On the Figure 1 we can see quality of our data before processing and it is obvious that quality of the reads drops on the end of the sequence. On the Figure 2 quality of the reads after processing is shown and there we can't see drop in read quality in the end of sequences.

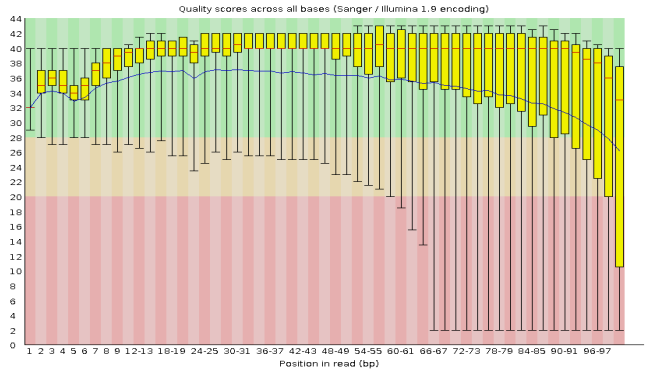

Figure 1 Quality scores before processing data

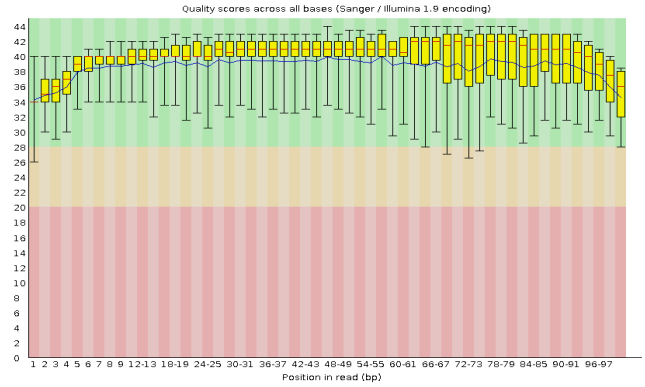

Figure 2 Quality scores after processing data

Figure 3 shows frequency of bases in sequences before trimming. Presence of overepresented motifs near the end of the sequence can be seen because of divergence of base frequencies near the end. Figure 4 shows frequency of bases in reads after removal of overrepresented motifs.

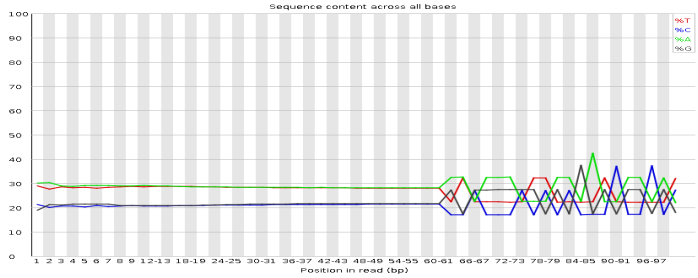

Figure 3 Frequency of the bases before removing overrepresented sequences

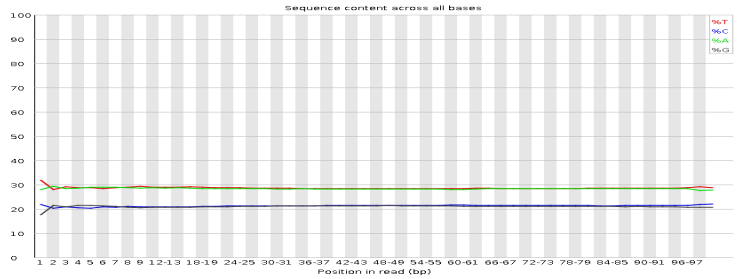

Figure 4 Frequency of the bases after removing overrepresented sequences

Results of the alignment can be seen on the Figure 5. On the top of the figure we can see chromosome 20 and our location. Each line in the middle part of figure represents one RNA read and its original location. On the bottom we can see RefSeq genes, annotated genes from human genome.

Figure 5 Visualisation of alignment in IGV

Variants found in IGV can be seen on Figure 6. In the first line we can see variations often present in the population and on the bottom line are shown variations in our patient. Lines in different colors represent different allels.

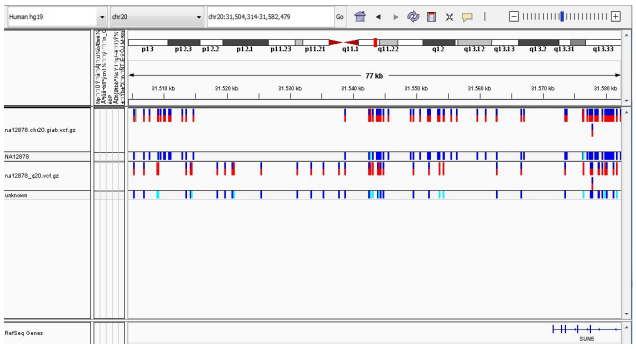

Figure 6 Found variants in IGV

Annotated variations summarised by type can be seen on the Table 1. While the Table 2 shows place where variations occurred.

Table 1 Types of variations found

Type	Total
SNP	52,041
MNP	1,042
INS	2,965
DEL	3,374
MIXED	273
INTERVAL	0
Total	59,695

Table 2 Location of found variations

Type (alphabetical order)	Count	Percent
DOWNSTREAM	5,971	6.114%
EXON	834	0.854%
INTERGENIC	33,889	34.7%
INTRON	50,018	51.214%
SPLICE_SITE_ACCEPTOR	6	0.006%
SPLICE_SITE_DONOR	2	0.002%
SPLICE_SITE_REGION	91	0.093%
UPSTREAM	5,888	6.029%
UTR_3_PRIME	843	0.863%
UTR_5_PRIME	122	0.125%

IV. Discussion

After we got the data (RNA sequences form our patient and reference) and did our first quality check we saw that some parts of sequences were of lower quality and that most of them were in the end of the sequence. The second generation of sequencers, from witch our data is generated, are giving good quality bases in the beginning of sequencing the reads but with the time the percentage of confidence is getting lower and lower. Because of that, the ending part of our sequences are of lower quality. As we can see on the Figure 3, the lines are very irregular. The reason why the line isn't straight are adaptors which scientists use as help in the sequencing. After the trimming we checked our data once again and saw that quality is better than before. This is because the computer removed the overrepresented sequences. After we visualized our results in IGV we could have seen red and blue boxes and lines. In Figure 5 they represent the variations (deletions, insertions, inversions, tandem duplications, translocations and single-nucleotide polymorphism (SNPs)) in our patient's genome compared with reference genome. The most common variations are SNPs – single-nucleotide polymorphisms, the differences in only one nucleotide. SNPs are the main reason why we look different and are similar to our parents. They usually don't cause diseases, but sometimes they can. Found variations are, in the most of the cases, in intergenic zone. That means they don't affect genes directly, but may affect their regulation. We aren't 100 percent sure about that because the intergenic zones are still an object of researches in genetics and genomics. Later, we saw some parts of the chromosome which can, indeed, affect on disease development and protein synthesis.

V. Conlusion

So, to conclude, genomics is very important and useful, but only if it is combined with medicine. If we want to really predict the disease our patient could develop we need to know where he lives, what is his life style and things like that. For that we need medicine. We hope that in not that far future we will be able to use genomics for diagnosing most of the patients.

VI. References

[1] E. Lander, Initial impact of the sequencing of the human genome, Nature, Feb 2011, vol 470, p. 187

[2] E. R. Mardis, A decade's perspective on DNA sequencing technology, Nature, Feb 2011, vol 470, p. 198

[3] M. Gaber et al, Computational methods for transcriptome annotation and quantification using RNA-seq, Nature methods, vol. 8, no. 6, June 2011, p. 469

[4] C. Trapnel and SL Salzberg, How to map billions of short reads onto genomes, Nature Biotehnology, vol. 27, no. 5, May 2009, p. 455

[5] F. Ozsolak and PM Milos, RNA sequencing: advances, challenges and opportunities, Nature Reviews Genetics, vol. 12, Feb 2011, p. 87