ISSN (online): 2349-784X

Module for Genomic Analysis of Rheumatic Arthritis using High throughput Sequencing Technology

Chetan Kumar M
M. Tech. Student (BMSPI)
Department of Electronics & Instrumentation Engineering
R V College of Engineering, Bangalore

Dr. K B Ramesh
Head of Dept. & Associate & Professor
Department of Electronics & Instrumentation Engineering
R V College of Engineering, Bangalore

Dr. Vidya Niranjan
Professor & Associate Dean
Department of Biotechnology
R V College of Engineering, Bangalore

Abstract

Bioinformatics is usually concerned with applying statistical and computational methods to analysis of data determined from sequenced DNA and/or RNA or the simulation of protein-protein interactions. Rheumatic Arthritis (RA) is an immune system ailment which implies that our own body insusceptible framework assaults the tissues. Due to this Rheumatic arthritis causes aggravation and delicate tissue swelling mainly at diarthrodial joints. This will result in significant loss of portability or movability because of the pain and joints destruction. According to NHS (National Health Survey) that rheumatic arthritis affects almost 2-4% of world population and in India it is around 0.5 %, with women has twice chance of developing RA compared to men. This project aims at development of software module for detecting and analysis of Rheumatic Arthritis by genome analysis using high-throughput sequencing. Genome sequencing is a technique in which the query sequence (subject sequence) is compared with entire stretch of human normal sequence (reference sequence) genome, in order to detect and analyze the variants in the query sequence. The process starts by mapping (aligning) the query sequence with reference sequence using burrows wheel aligner tool, then ambiguities and duplicates are removed to increase the accuracy using Picard tool. Variants are detected from the mapped data with help of GATK java toolkit, detecting and annotating these variants will provide essentials details for experts to analyze the condition of the rheumatic arthritis disease. The Rheumatic Arthritis disease affected data is imported from the expert doctor and normal genome data is imported from GenBank to perform whole genome sequencing with reference genome. The proposed methodology is applied to several RA affected and normal sample, and obtained results will provide the information about existence of disease and gene-level analysis of genes that are associated or in relation to the cause for RA disease. Outcome of proposed methodology for a sample of acquired RA disease data resulted in total 897 affected genes, the gene TYK2 has major impact factor score on RA disease is 0.4215 and MAP1A low impact factor score of 0.009e-03. Other genes that are normally found in RA disease subject are HLA-B, HLA-C and TNFRS10B has score 0.4, 0.3 and 0.25 respectively are also appeared in the detected sample. Thereby, whole genome analysis and variants detection will able to detect and analyze the presence of rheumatic arthritis disease.

Keywords: Rheumatic Arthritis (RA), Single Nucleotide Polymorphism (SNP), National Health Survey (NHS), SAM (Sequence Alignment Mapping), Variant Call Format (VCF)

I. INTRODUCTION

Rheumatic Arthritis is an autoimmune disease which means that our own body immune system attacks the tissues. Due to this Rheumatic arthritis causes aggravation and delicate tissue swelling mainly at diarthrodial joints [1][2]. This will result in significant loss of portability or movability because of the agony and joints destruction. The disease is systemic, naturally influencing the synovial joints and peri-articular synovial structures specifically. The disease starts its progression with small bones of arms and feet, and it can appear or spread to any joints of body which involves in locomotion movement. The systemic method for the condition infers that various distinctive organs may get the opportunity to be incorporated as the condition propels. Case of additional articular association can incorporate side effects and impacts, for example, fever, weight reduction, exhaustion or shortcoming, swollen lymph hubs, iron deficiency, knobs, dry eyes and other disorders [3].

The initial cause for the disorder is not known. Side effects can happen in a single instance of pain and excruciating joints which may last a few weeks or months, or as a forceful and damaging condition which advances quickly, and if not taken care, prompts extreme physical handicap. Distinctively, disease typically shows as a respective polyarthritis. This condition will follow a degeneration and pattern of remission as time progress.

In a healthy joint, the tissue covering the joint (called the synovial layer or joint synovium) as showed in Figure 1. It is thin and creates liquid that greases up and feeds joint tissues. In rheumatoid joint pain, the insusceptible framework assaults the synovial film, creating aggravation, torment, swelling and firmness. This causes synovial film to wind up thick and kindled, bringing about undesirable tissue development. In the end, bone decimation and everlasting joint harm can happen, prompting lasting inability.

According to NHS (National Health Survey) that rheumatic arthritis affects almost 2-4% of world population and in India it is around 0.5 %, with women has twice chance of developing RA compared to men. And it is reported that RA can develop at any age period, but the condition is more common in those aged 45 and older [4][5].

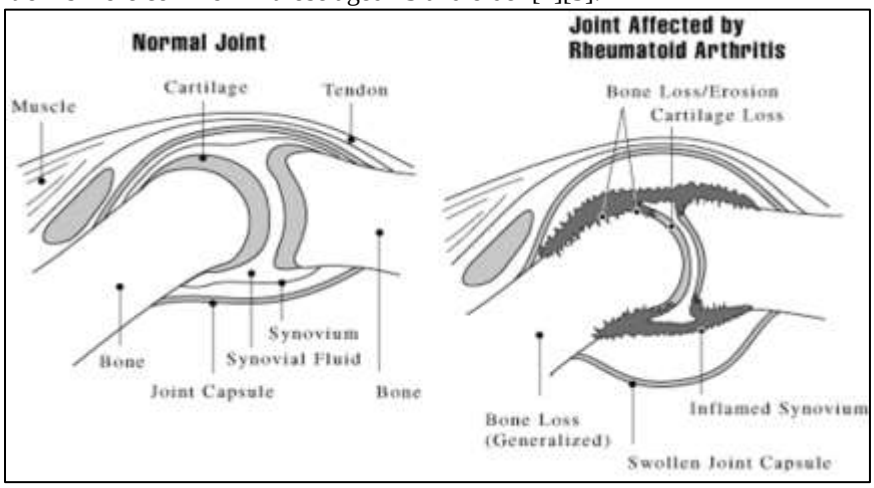

Fig. 1: Effects of Rheumatic Arthritis

II. IMPLEMENTATION

The proposed method starts by aligning the subject sequence with the reference genome sequence in order to create a SAM/BAM file format which is an alignment mapping. Next, marking the duplicates to mitigate biases introduced by data generation steps such as PCR amplification. At that point, local realignment around Indels is performed, since algorithm that are utilized as a part of the underlying mapping step tend to deliver different sorts of artifacts in the locales around indels. At last, recalibrating the base quality scores, for the reason that the variation calling calculations depend intensely on score that are indicating the quality of reads allocated to each bases in every sequence read [6].

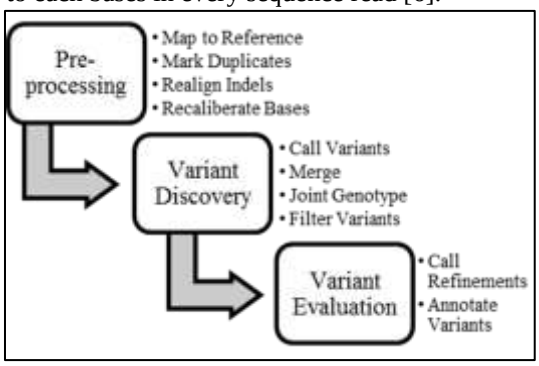

Fig. 2: Implemented methodology

A. Pre-Processing

This first phase is to depicts the pre-handling process that are important to set up the data for further process, beginning the analysis with FASTQ data & finishing with BAM file which is ready for further analysis.

When sequence data is retrieved from available database or sequence provider, the data is regularly in a crude state which is not proper for variant discovery. Regardless of the possibility that a ready BAM file has to be applied with some data analysis in order to maximize the accuracy of the data [6][7].

B. Variant Discovery

Once pre-processing of data is done, variant discovery process is to be take place, i.e. recognize the locales where data shows variation in respect to the reference genome, and find out genotypes for every specimen at that region utilizing GATK tools.

But, some percent of the variation may be caused by sequencing and alignment (mapping) artifacts, hence the best aspect here is to adjust the requirement for sensitivity (i.e. neglecting to recognize genuine variations) versus specificity (i.e. neglecting to

reject antiquities). It is exceptionally hard to accommodate these objective in a solitary step, so it is decomposed the variants discovery into two process: variation calling and variation filtering. The initial process is intended to expand sensitivity, whereas the filtering process plans to convey the level of customized specificity [7].

For DNA, the variant calling step is further subdivided into two separate steps in order to enable scalable and incremental processing of cohorts comprising many individual samples. In the final analysis, this workflow produces results equivalent to traditional joint calling, in which all samples are given simultaneously to the variant caller, but it scales much better and resolves the so-called N+1 problem. Note that this is equally applicable to small cohorts or even single samples.

The most ideal approach to channel the subsequent variant callset is to utilize variation quality score recalibration (VQSR), which uses machine learning to identify annotation profiles of variants that are likely to be real. The drawback of this sophisticated method is that it requires a large callset and highly curated sets of known variants.

C. *Variants Evaluation*

When you have created and separated callset, there are a few alternatives for assessing and refining the variation and genotype calls further.

In the previous segment, some refinement ventures in the genotype calls taking into account population frequencies and family data if accessible, add functional annotation identified with anticipated biological impact, and perform some of the quality assessment by contrasting the callset with known assets.

Functional annotations are predictions of what effect a variant may have on biological function based on its position within coding regions, regulatory regions and so on. These depend on external databases of known variants and functional predictions. Annovar tools are used for performing the functional annotation [8].

III. RESULT AND DISCUSSION

All The proposed methodology is applied and analyzed for rheumatic arthritis data retrieved from NCBI (National Centre for Biological Information) [9] [10] and UCSC genome database [11]. Functional annotation is performed for obtained variants using Annovar web portal. The proposed methodology is used detect the variants or SNPs in the rheumatoid arthritis and conclude the genes that are playing major role in RA disease [12] [13]. The below list of figures shows the functional annotation performed for rheumatic arthritis data. Figure 3 shows the basic information of associated in the computation of variants. Figure 4 and 5 displays the exome and genome variants and its complete information of variants or SNP respectively. Figure 6 provides the bar plot view and their score for the genes associated with rheumatic arthritis. Figure 7 and 8 provides the network view of genes associated with rheumatic arthritis, figure 7 provides the network view of genes that are directly associated with rheumatic arthritis disease. Whereas figure 8 displays the genes which are directly associated with RA.

Consider figure 4 and 5 are screenshots of exome and genome variant call format of analyzed rheumatic arthritis data. These annotated variants file provides information about the genes chromosome, location of variants defined by the start and end position of the variants, change in query sample nucleotide with respect to reference sample nucleotide, function of the genes associated with variants and its name, disease associated with variants with respect to GWAS and other essential details (but, in below figure the variants file fields are restricted).

The proposed methodology helps the pharmaceutical company to analyze about the disease about genes and their drug discovery because it provides clear picture of individual genes associated information and their impact scores associated with Rheumatoid arthritis disease.

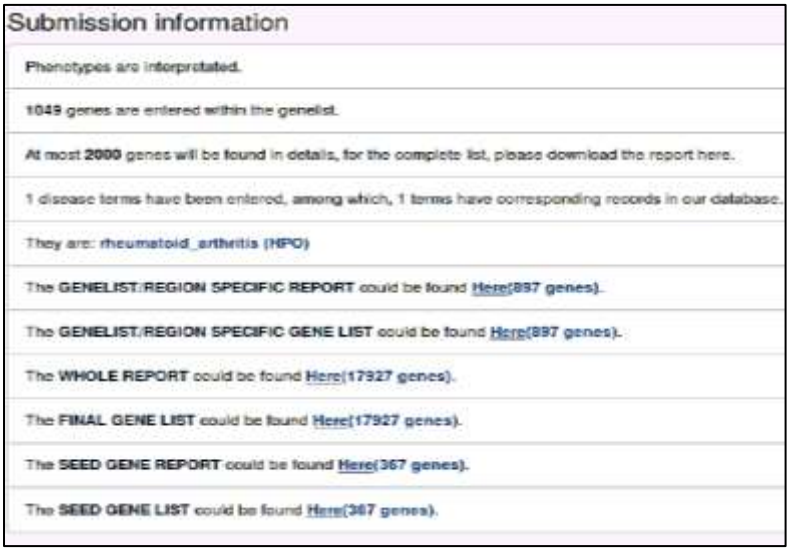

Fig. 3: Summary of genes associated with RA disease

Chr	Start	End	Ref	Alt	Func	Gene
chr1	1014274	1014274	A	G	exonic	ISG15
chr1	1046551	1046551	A	G	exonic	AGRN
chr1	1054900	1054900	C	T	exonic	AGRN
chr1	1223251	1223251	A	G	exonic	SDF4
chr1	1312114	1312114	T	C	exonic	CPSF3L
chr1	1313807	1313807	G	A	exonic	CPSF3L
chr1	1319461	1319461	C	G	exonic	CPSF3L
chr1	1327586	1327586	C	T	exonic	CPTP
chr1	1512376	1512376	G	C	exonic	ATAD3A
chr1	1543953	1543953	A	G	exonic	SSU72
chr1	1616547	1616547	T	C	exonic	MIB2
chr1	1754601	1754601	G	T	exonic	NADK
chr1	1756611	1756611	G	C	exonic	NADK
chr1	3497304	3497304	C	G	exonic	MEGF6
chr1	3499885	3499885	C	A	exonic	MEGF6
chr1	3883678	3883678	A	G	exonic	DFFB
chr1	6098502	6098502	A	G	exonic	KCNAB2
chr1	6554331	6554331	A	C	exonic	NOL9
chr1	6554355	6554355	A	G	exonic	NOL9

Fig. 4: Exomes variants genes list and details

Chr	Start	End	Ref	Alt	Func	Gene
chr1	629816	629816	T	C	upstream	LOC101928626
chr1	629906	629906	C	T	upstream	LOC101928626
chr1	630026	630026	C	T	intergenic	LOC101928626,MIR6723
chr1	631811	631811	C	T	downstream	MIR6723
chr1	631862	631862	G	A	downstream	MIR6723
chr1	633887	633887	G	A	intergenic	MIR6723,OR4F16
chr1	633987	633987	C	T	intergenic	MIR6723,OR4F16
chr1	634112	634112	T	C	intergenic	MIR6723,OR4F16
chr1	919397	919397	A	G	ncRNA_exonic	LOC100130417
chr1	919695	919695	C	G	upstream	LOC100130417
chr1	944296	944296	G	A	UTR3	NOC2L,SAMD11
chr1	944307	944307	T	C	UTR3	NOC2L,SAMD11
chr1	951998	951998	A	G	splicing	NOC2L
chr1	959084	959084	G	C	intronic	NOC2L
chr1	965350	965350	G	A	UTR3	KLHL17
chr1	1014274	1014274	A	G	exonic	ISG15
chr1	1046551	1046551	A	G	exonic	AGRN
chr1	1050593	1050593	T	A	splicing	AGRN
chr1	1054553	1054553	T	G	splicing	AGRN
chr1	1054900	1054900	C	T	exonic	AGRN

Fig. 5: Genome variants genes list and details

IV. CONCLUSION

Detecting and analyzing the Rheumatoid arthritis at their gene level is a difficult task. But the proposed methodology will provide will help to detect and analyze the rheumatoid arthritis by providing complete idea for generating SNP (Single Nucleotide Polymorphism) or variants of the Rheumatoid arthritis sample using available GATK, BWA and other tools. Functional annotation is used to interpret the obtained variants using Annovar, it provides information pertaining to the genes that are directly or indirectly associated with rheumatic arthritis. The results shown will provide clear cut indication about existence of disease and also provide detailed information about genes associated with disease. The proposed methodology is not only restricted to analyze rheumatic arthritis disease, it can also be applied to detect and analyze other disease.

ACKNOWLEDGMENT

I would like to thank Mr. Sanjay Deshpande for guiding us carrying this research and providing support for performing the whole genome analysis.

REFERENCES

[1] Minghui Wang, Xiang Chen, Meizhuo Zhang, Wensheng Zhu, Kelly Cho and Heping Zhang." Detecting significant single-nucleotide polymorphisms in a rheumatoid arthritis study using random forests" BMC Proceedings2009,3(Suppl 7): S69
[2] Yan V Sun, Zhaohui Cai, Kaushal Desai, Rachael Lawrance, et al. "Classification of rheumatoid arthritis status with candidate gene and genome-wide single-nucleotide polymorphisms using random forests" BMC Proceedings2007, 1(Suppl 1): S62.
[3] National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS). Handout on Health: Rheumatoid Arthritis,2014. [Online] Available: http://www.niams.nih.gov/health_info/rheumatic_disease/
[4] Neetu Sachan, Saurabh Chaddha, Kamal Kishore, Mayank Yadav, Phool Chandra. "Evaluation of Awareness in Professional Students of IFTM University About Rheumatoid Arthritis" Indian Journal of Drugs, 2013, 1(2), 38-41.
[5] K.B. Ramesh, Prabhu Shankar. K. S, B. P. Mallikarjunaswamy, E.T. Puttaiah. "Power Spectrum Sequence Analysis of Rheumatic Arthritis (RA Disease Using DSP Technique)." International Journal of Research in Engineering and Technology, 2013, Volume: 02 Issue: 09.
[6] Krithika Bhuvneshwar, Dinanath Sulakhe, Robinder Gauba, et al, "A case study for cloud based high throughput analysis of NGS data using the globus genomics system", Computational and Structural Biotechnology Journal 13, 2015, 64–74.
[7] Genome Analysis Toolkit (GATK). [Online] Available: hnp://www.broadinstitute.org/gatk/guide/best-practices, 2016.
[8] wAnnovar for functional Annotation [Online] Available: http://wannovar.usc.edu/.
[9] National Central for Biotechnology Information (NCBI). [Online] Available: http://www.ncbi.nlm.nih.gov/, 2016.
[10] National Central for Biotechnology Information. SRA Handbook, 2014 [Online] Available: http://www.ncbi.nlm.nih.gov/sra/.
[11] UCSC Genome Browser (UCSC). [Online] Available: https://genome.ucsc.edu/, 2016
[12] Angela Mc Ardle, Brian Flatley, Stephen R. Pennington and Oliver FitzGerald. "Early biomarkers of joint damage in rheumatoid and psoriatic arthritis" Arthritis Research & Therapy, 2015 17:141
[13] Graham B Wiley, Jennifer A Kelly and Patrick M Gaffney. "Use of next-generation DNA sequencing to analyse genetic variants in rheumatic disease" Arthritis Research & Therapy, 2014,6:490.