WISSEN

TECHNOLOGY PAPER

Usage of Spark to increase the efficiency of GENOMIC DATA COMPARISON for Anthropological Genetics data analysis

Authors

Anika Ranjan, Balaji Sankar Konakalla

www.wissen.com

Contents

www.wissen.com

Introduction

The study of genomic data has always intrigued scientists and geneticists across the world. By studying the genetics data of a species across various population groups, one can determine the degree of relatedness and genetic mutations among the different groups of species. It helps us in anthropological genetics to understand human evolution. Thanks to the various research efforts worldwide, we have detailed sequencing information for various population groups widely and freely available for data analysis. By analyzing this genomic data, one can identify common genetic patterns within various populations, which can lead to breakthrough revelations and also help solve many diseases. The size of the genomic data, however, causes an impediment in this analysis. Even with the use of modern programming languages, and readily available bioinformatics tools like BLAST (Basic Local Alignment Search Tool), the analysis of such data often becomes difficult and very slow due to the sheer size of the genomic data and limitations of computational resources on one machine. In order to solve this challenge and make the analysis of genomic data more efficient and faster, we can run the tools in a cluster environment such as Spark.

In this whitepaper, we propose the use of Spark to implement highly scalable and efficient computational systems that can be used in anthropological genetics analysis. We conducted a proof of concept by executing BLAST and other tools on human genomic data from the 1000 genomes project using Spark. Horizontal scaling in execution times was observed on increasing the number of executors thereby demonstrating that Spark could serve as an effective computational platform for genomic data analysis.

Genomic Data Analysis

Efficient and Faster Analysis

Computational Platforms

Background

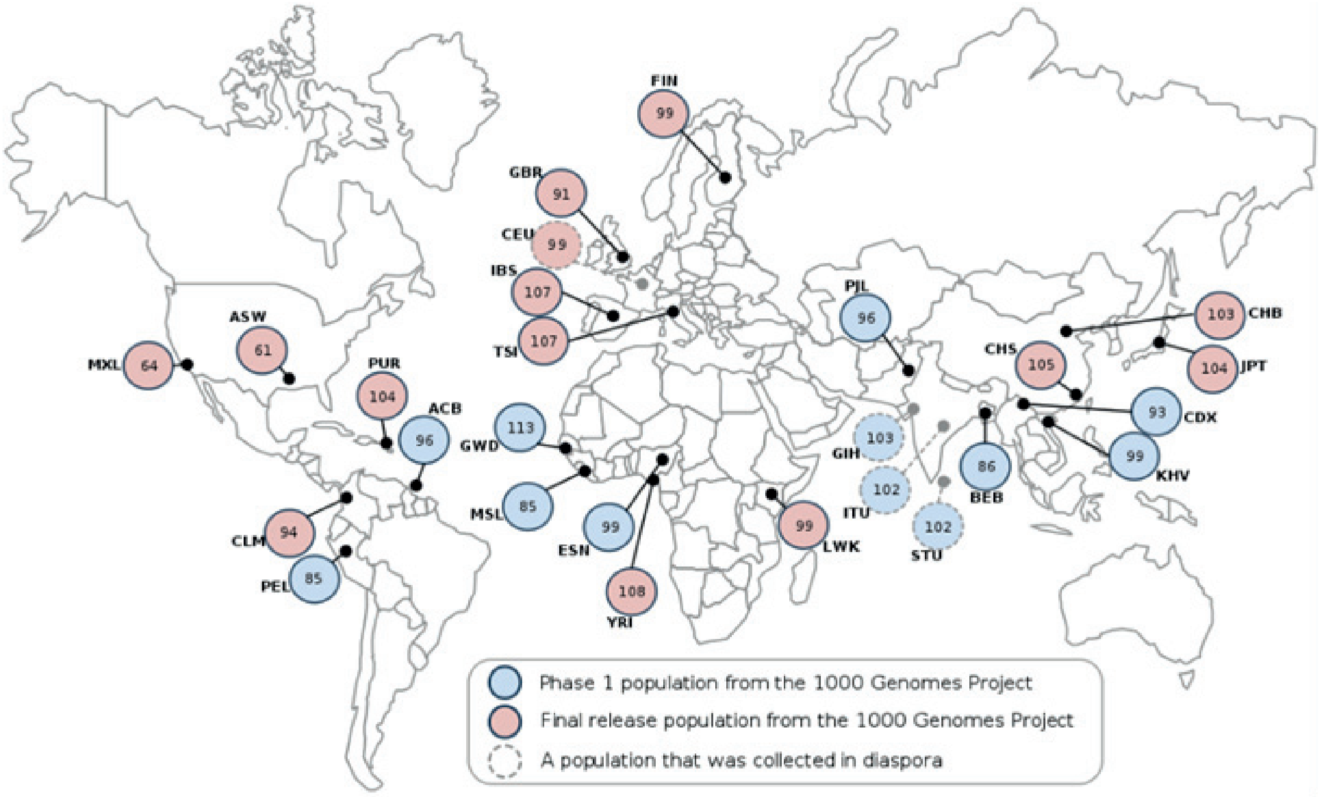
Anthropological Genetics

Anthropological genetics is a synthetic area of study where the evolutionary theory is examined while applying genetic methodologies and is of interest to anthropologists (Crawford and Beaty, 2013). The patterns of genetic variations and commonalities among human populations are studied in this field which helps anthropological geneticists to determine the degree of relatedness among different population groups and learn about a society's mating structures, fluctuations in size, population history, and the amount of mixing, or admixture, that occurred between various population groups. A lot of excitement in this area is around genetic research on the origins of Homo sapiens, still, genetic anthropologists are very much interested in looking at the more recent events in the story of the human race, such as the history and patterns of human migrations and identifying correlations from them (Ben-Ari, 1999).

1000 Genomes Project

The 1000 genomes project was a research effort carried out at an international level to establish a very detailed database of human genetic variation. The project utilized resources from multidisciplinary research teams from worldwide institutes. All the teams contributed to an enormous sequence dataset which is freely accessible through public databases to the scientific community for any kind of research undertaking (Wikipedia contributors, 2022).

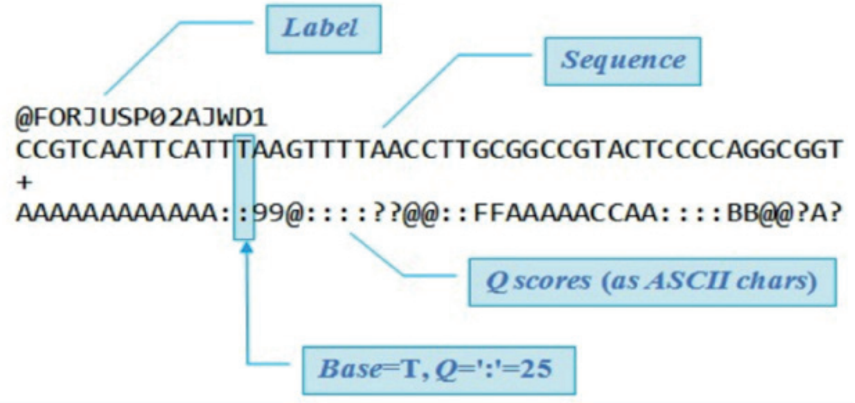
Below is a map that identifies the locations of population samples of the 1000 Genomes Project. The circle in the map represents the number of sequences that were in the final release (Wikipedia contributors, 2022).



FASTQ files

FASTQ files are used for nucleotide storage where separated nucleotide strings are compacted in four sentence format per sequence string.

Example of a FASTQ file (FASTQ Files, n.d.)



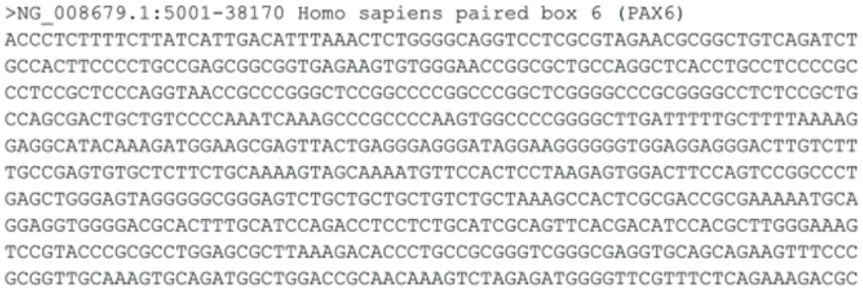
```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

The label is an assigned group of numbers that talk about the location of the nucleotide relative to the genome. The Q scores or PHREAD scores are assigned as a measure of quality. They indicate the accuracy of the sequenced base pair being the exact base pair and also containing the nucleotide sequence. A FASTQ sequence for a genome is split into two sister files: the forward and reverse sequences.

FASTA files

FASTA files only contain the sequence strings of the FASTQ file. While FASTA files are smaller and less condensed, the quality of the base pair index is lost and also unattainable. Because of being a simpler format than FASTQ, FASTA files can be converted from FASTQ files using outside tools. Because of this reason, most sequences sequenced for the public domain will be in a FASTQ format - to benefit any projects that do require PHREAD score information. The first sentence at the beginning of the FASTA file indicates information on the context of the genome: species, chromosomes, etc. (Akalin, 2020).

Example of FASTA format (Akalin, 2020)



```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

BLAST files

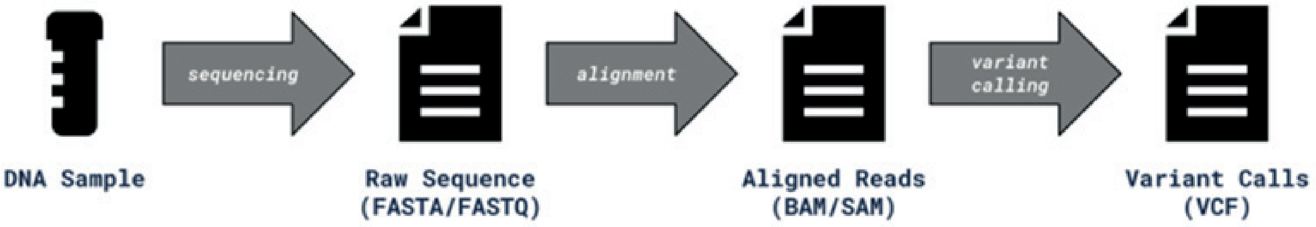
BLAST is a tool designed to compare two genomes in the FASTA file format. The program compares two sequences (nucleotide or protein) and can give output in a variety of formats. The output indicates the results of comparing the sample genomic data with the reference genome, to identify the location of variants.

BLAST can be executed using the following options:

- Run it online on NCBI website by uploading the genome that needs to be compared (Basic Local Alignment Search Tool, n.d.).
- Run it from our local computer by using tools like Biopython calling the BLAST online version, which runs BLAST over the internet connection, using NCBI database (Biopython - Overview of BLAST, n.d.).
- Run standalone BLAST, by downloading the standalone BLAST tool, installing the reference genomic database, and then comparing your sample data with this reference data. This does not involve any data transfer over the internet connection and hence could be faster. Also, this gives us the flexibility to create our own reference database to search against sequences (Biopython - Overview of BLAST, n.d.).

VCF (Variant Call Format) file

VCF (Variant Call Format) file: VCF is the abbreviation of Variant Call Format. This file format is used for various research projects such as the 1000 Genomes project to encode structural genetic variants. A variant call format file is the result of a bioinformatics pipeline. A VCF file is a text file that is used for storing gene sequence variations. The process of creating a VCF file involves multiple steps (Maurer, 2022). Firstly, a DNA sample is sequenced through a next-generation sequencing system (NGS system) that creates a raw sequence file. Next, the raw sequence data is aligned which creates a BAM/SAM file. After this, the variant calling process identifies changes to a particular genome as compared to the reference genome (Maurer, 2022). This results in a variant call format (VCF) output file. VCF has genotype data stored in a tab-delimited file format. Below is a high-level flow of a pipeline that produces a VCF file in bioinformatics (Maurer, 2022):



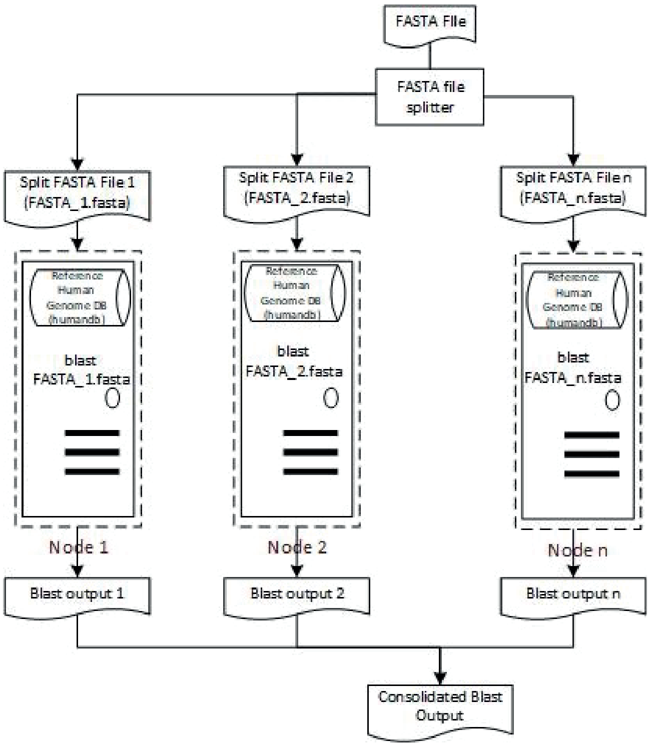
Implementation

Project 1

Running BLAST on a sample human genome against the reference genome using Spark to identify variants

The purpose of this project was to run BLAST locally in an efficient way on a sample human genome against the reference human genome and create an output indicating locations of genetic variations and matches. For this project, BLAST was executed for a variety of sizes of the sample file, using a different number of executors, to identify how the efficiency/ performance of the execution of BLAST changes with the change in the number of executors in the cluster.

Diagram


FASTA File
FASTA file splitter
Split FASTA File 1 (FASTA_1.fasta)
Split FASTA File 2 (FASTA_2.fasta)
Split FASTA File n (FASTA_n.fasta)
Reference Human Genome DB (humandb)
blast FASTA_1.fasta
blast FASTA_2.fasta
blast FASTA_n.fasta
Node 1
Node 2
Node n
Blast output 1
Blast output 2
Blast output n
Consolidated Blast Output

Steps Involved:

1. Install BLAST locally (Index of /Blast/Executables/Blast+/LATEST, n.d.) .
2. Get the reference human FASTA file (GRCh38_latest_genomic.fna) to make a reference local database (Human Genome Resources at NCBI, n.d.).
3. Create a local database using this reference genome by running the following makeblastdb -in GRCh38_latest_genomic.fna -dbtype nucl
4. Download the low coverage WGS (Whole Genomic Sequencing) data FASTQ file for any sample from 1000 genomes project (IGSR: The International Genome Sample Resource, n.d.).
5. Convert FASTQ to FASTA file by running the following command python FASTQ_to_FASTA_converter. py ERR016162_1.fastq ERR016162_1.fasta gz
6. Create multiple sizes of the sample FASTA file (5MB, 10MB and 20MB).
7. Run the BLAST against the local human reference genome data for various sizes of the sample data on a Hadoop cluster with a different number of executors.
8. Track the amount of time taken for Blast to complete for each execution.

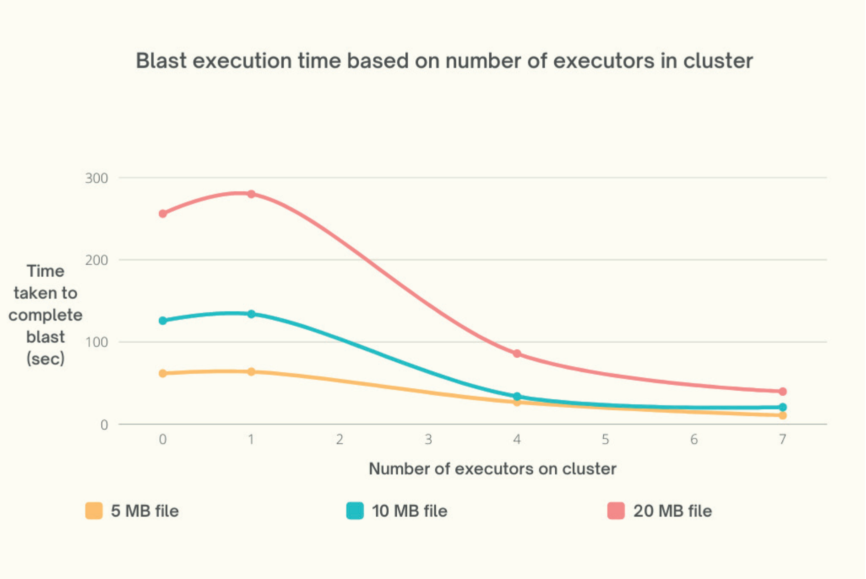
The above diagram depicts the data distribution of the sample FASTA file across n nodes in a cluster and running BLAST on all the nodes against the human reference genome data (which is copied in each of the nodes). The query file is evenly split and distributed across all the nodes. Each node has BLAST and human reference genome data locally installed. BLAST is executed, in parallel, individually at each of the nodes, creating separate BLAST output files. The BLAST output is consolidated in the end to one file.

Key Findings

Following was the time taken in minutes to perform BLAST on query FASTA file of varying sizes:

Query FASTA file size	Standalone BLAST execution time (sec)	Execution time on cluster with 1 Executor (sec)	Execution time on cluster with 4 Executors (sec)	Execution time on cluster with 7 Executors (sec)
5 MB	62	64	27	11
10 MB	126	134	34	21
20 MB	256	280	86	40

Below is a graph of the comparison of BLAST execution time taken for each of the sample files and the change in efficiency based on the number of executors in the cluster.



The above graph indicates how increasing the number of nodes on the cluster on which BLAST is being executed drastically improves the efficiency of BLAST and reduces the execution time.

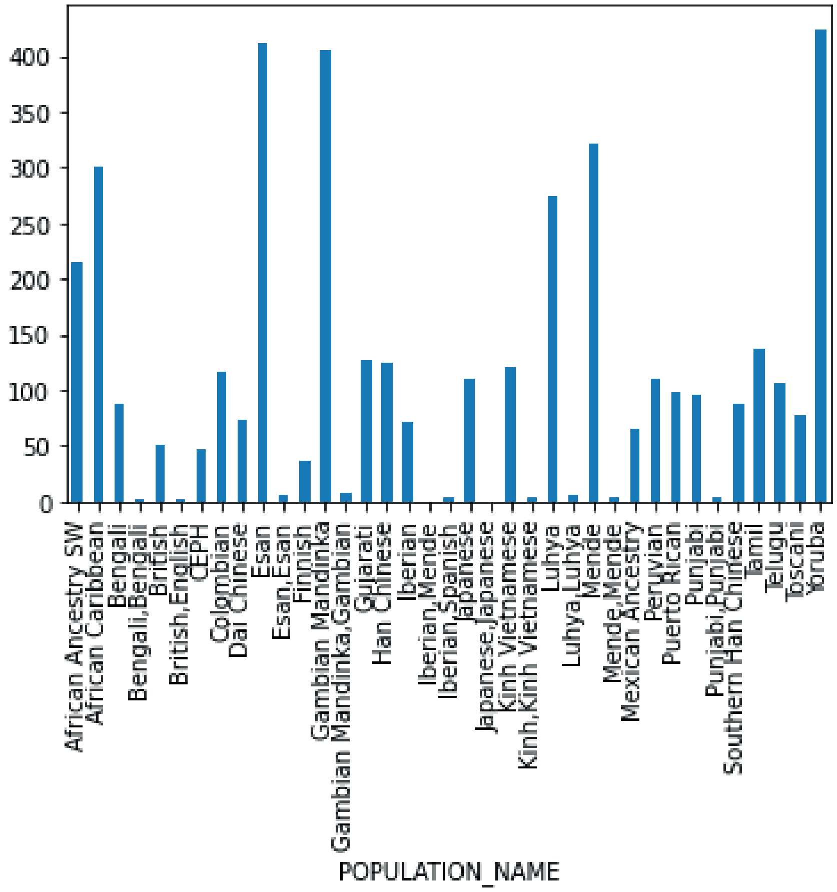
Project 2

Identify mutations on a specific gene across samples of the human population to identify the co-relation between population origins with genetic mutations

The purpose of this project was to create a program to compare the VCF file for various samples of human genomes (2548 samples) from the 1000 genome project targeting a specific location within a chromosome. The VCF file along with the population origin information can be collectively used to identify mutations for a specific gene based on population origins.

Key Findings

Below is a graph of the number of mutations found per the various population origins. Based on this graph, we can identify which population origins had the maximum number of genetic mutations for the gene in consideration.



Steps Involved:

1. Download the VCF file from the 1000 genomes project for a specific chromosome for which we want to analyze the mutations (Index of /Vol1/Ftp/Data_Collections/1000_Genomes_Project/Release/20190312_Biallelic_SNV_and_INDEL, n.d.) .
2. Download the population origin information file (IGSR: The International Genome Sample Resource, n.d.).
3. Create a program in pyspark to read the VCF file for the specific gene and identify the number of mutations within that sequence for various samples.
4. Create a graph on this data set to identify the change in the number of mutations based on population origin.

Conclusion

Based on this experiment, we can conclude that Spark can be effectively utilized to perform large scale genomic data analysis. We observed horizontal scaling with large cluster size, thereby providing a way forward for analysing large genomic data sets. The free availability of genomic sequences along with easy accessibility to computational platforms like Spark will usher a new wave of exciting research in genomics.

Increased Efficiency

Horizontal Scaling

Spark for Genomics

References

- Crawford, M., Beaty, K. (2013, November 18). *DNA fingerprinting in anthropological genetics: past, present, future.* National Library of Medicine. Retrieved October 12, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831593/
- Ben-Ari, E (1999, February 16). *Molecular biographies: Anthropological geneticists are using the genome to decode human history.* BioScience. Retrieved October 11, 2022, from https://academic.oup.com/bioscience/article/49/2/98/240412
- Wikipedia contributors. (2022, August 6). 1000 *Genomes Project.* Wikipedia. Retrieved October 12, 2022, from https://en.wikipedia.org/wiki/1000_Genomes_Project
- *FASTQ files.* (n.d.). Retrieved October 13, 2022, from https://www.drive5.com/usearch/manual/fastq_files.html
- Akalin, A. (2020, September 30). *7.1 FASTA and FASTQ formats | Computational Genomics with R.* Retrieved October 13, 2022, from https://compgenomr.github.io/book/fasta-and-fastq-formats.html
- *Basic Local Alignment Search Tool.* (n.d.). National Library of Medicine. Retrieved October 12, 2022, from https://BLAST.ncbi.nlm.nih.gov/BLAST.cgi
- Biopython - *Overview of BLAST*. (n.d.). Retrieved October 11, 2022, from https://www.tutorialspoint.com/biopython/biopython_overview_of_BLAST.htm
- Maurer, I. (2022, February 16). *What is a Variant Call Format (VCF) file?* Precision Oncology Solutions | GenomOncology. Retrieved October 11, 2022, from https://www.genomoncology.com/blog/what-is-a-variant-call-format-vcf-file
- *Index of /BLAST/executables/BLAST+/LATEST*. (n.d.). Retrieved October 11, 2022, from https://ftp.ncbi.nlm.nih.gov/BLAST/executables/BLAST+/LATEST/
- *Human Genome Resources at NCBI.* (n.d.-c). National Center for Biotechnology Information. Retrieved October 12, 2022, from https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/
- *IGSR: The International Genome Sample Resource.* (n.d.). Retrieved October 11, 2022, from https://www.internationalgenome.org/data-portal/
- *Index of /vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL*. (n.d.). Retrieved October 11, 2022, from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20190312_biallelic_SNV_and_INDEL/

About Wissen

Wissen Group is a leading IT services and solutions company offering services to companies in the domains of Banking & Finance, Telecom, and Healthcare. Established in the year 2000 in the US, Wissen also has offices in India, UK, Australia, Mexico, and Canada, with best-in-class infrastructure and development facilities. Wissen has successfully delivered projects worth $1 billion for more than 25 of the Fortune 500 companies. The Wissen Group includes more than 4000 highly skilled professionals.

Wissen offers an array of services including Application Development, Artificial Intelligence & Machine Learning, Big Data & Analytics, Visualization & Business Intelligence, Robotic Process Automation, Cloud, Mobility, Agile & DevOps, Quality Assurance & Test Automation. Wissen is uniquely positioned to help clients in building Enterprise Systems, implementing Digital Strategy, and gaining a competitive advantage with business transformation.

About Wissen Technology

Great Place To Work. Certified APR 2021–MAR 2022 INDIA

Wissen Technology is a niche global technology consulting and solutions company. It is part of Wissen Group and was established in the year 2015 to deliver critical projects. Our team currently consists of 1200+ highly skilled professionals.

Wissen Technology has been certified as a **Great Place to Work**®. The technology and thought leadership that the company commands in the industry is the direct result of the kind of people Wissen has been able to attract. Wissen is committed to providing them the best possible opportunities and careers, which extends to providing the best possible experience and value to our clients.

✉ wissentechnology.hr@wissen.com

Our service offerings

- Application Development
- Artificial Intelligence and Machine Learning
- Big Data and Analytics
- Visualization and Business Intelligence
- Robotic Process Automation
- Cloud and Mobility
- Agile and DevOps
- Quality Assurance and Test Automation
- Infrastructure Management

WISSEN TECHNOLOGY

www.wissen.com

USA | CANADA | UK | INDIA | AUSTRALIA