Review Article

# Overview of a typical next generation sequencing (NGS) project

**Nancy Bhura, Atul Kumar Upadhyay***
School of Bioengineering and Biosciences, Division of bioinformatics, Lovely Professional University, Phagwara, India.
***Corresponding author's E-mail:** atul4606iabt@gmail.com

**Received:** 30-07-2017; **Revised:** 15-09-2017; **Accepted:** 06-10-2017.

**ABSTRACT**

With the advent of Next Generation Sequencing (NGS) there is huge amount of genome-transcriptome sequence data available in public domain from all most all the kingdom of life such as microbe, plant and animal. To make any meaning from this data, we need to analyse these sequences for the presence of the genes and other regulatory elements in these sequences. All most all the NGS project produces enormous amount of raw data and there is big gap between amount of data generated and computational experts to make meaning from these data. Considering the above problem this study is performed to make an intensive survey of tools and techniques used for the analysis of genomes. Eukaryotic genomes are more complex as compared to prokaryotic so require different set of tools. Most of the available tools are specific to specific genome types, such as small genomes require different set of tools to assemble it as compared to bigger genomes. Gene prediction tool use different features such as start site, stop site, intron-exon signatures etc. to predict genes on DNA sequence. Gene prediction tools such as MAKER, FgeneSH etc. also depends on the genome types and related available resources. There are still many open questions to be invertigated in the field of NGS data analysis.

**Keywords:** Genome, Annotaions, Assembly, Next Generation Sequencing, DNA, Protein.

## INTRODUCTION

High throughput genome and transcriptome sequencing of organisms produce lot of sequence data. To make meaning from this data one needs to analyze this data in detail and it needs human expertise to gandle such data. The analysis of genome data requires several steps (Figure 1).

Major steps of NGS data analysis can be broadly classify into following four groups:

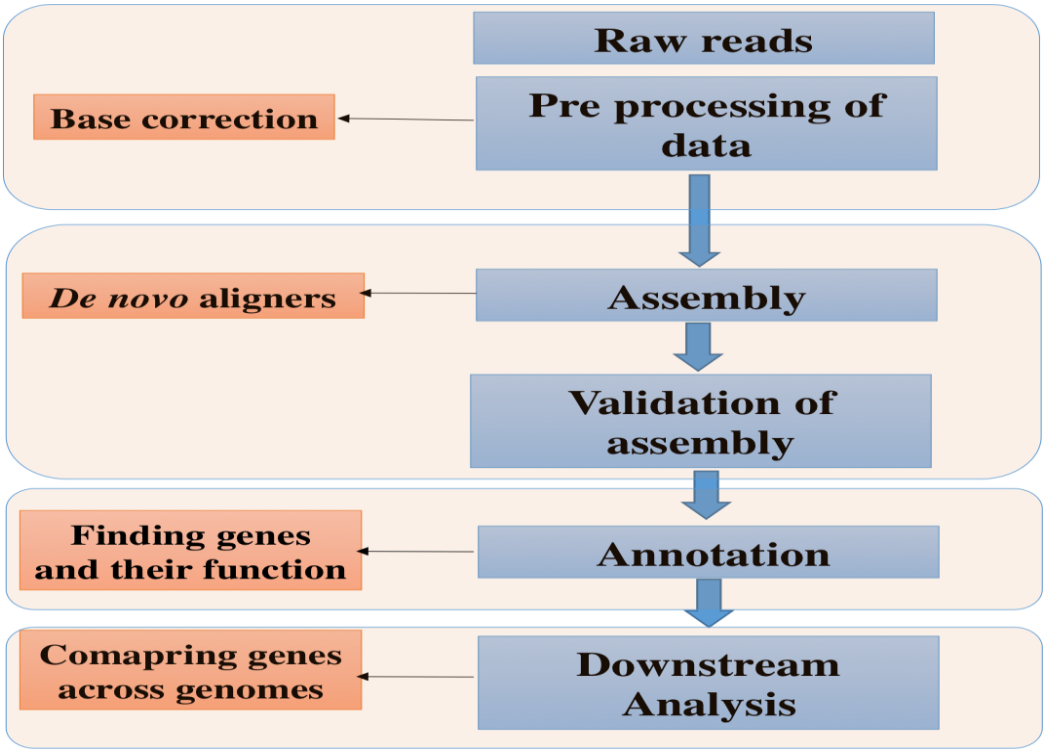


**Figure 1:** Cartoon representation of major steps of a typical NGS data analysis. Starting from pre-processing of raw reads before assembly followed by genome annotation and downstream analysis.

**Pre-Assembly Steps**

This step is also known as pre-processing of raw data for its quality and authenticity. Firstly, raw reads need to be processed by filtering low quality reads based on Phred score. Minimum average phred score of 20 is essential for a read to qualify for further analysis. The filtering of low qaulity reads depends on quality of the sequencing data (phred score), overall GC content and abundance of repeats or duplicated reads. There are many open source tools and softwares available for the pre-processing of data (Table 1).

**Table 1:** List of commonly used softwares for the pre-processing of NGS raw data.

| S. No. | Tool/Software | Year | Link |
|---|---|---|---|
| 1 | NGS short(2.1) | 2014 | http://research.bioinformatics.udel.edu/genomics/ngsShoRT/ |
| 2 | NGS QC toolkit (2.3.2) | 2012 | http://www.nipgr.res.in/ngsqctoolkit.html |
| 3 | Fastx toolkit(0.0.13.2) | 2010 | http://hannonlab.cshl.edu/fastx_toolkit/ |
| 4 | Seqtrim | | http://www.shigatox.net/stec/cgi-bin/seqtrim |
| 5 | Cutadapt | 2017 | https://pypi.python.org/pypi/cutadapt |
| 6 | Btrim | 2011 | https://omictools.com/btrim-tool |
| 7 | Solexaqa (v.2.2) | 2011 | http://solexaqa.sourceforge.net/ |
| 8 | Sickle | 2011 | https://omictools.com/sickle-tool |
| 9 | Scythe | 2011 | https://omictools.com/scythe-tool |
| 10 | Trimmomatic (v.0.32) | 2014 | http://www.usadellab.org/cms/?page=trimmomatic |

**Genome Assembly (DE NOVO OR Reference Based)**

Atere filtering raw reads the good quality reads are assembled in draft or complete genome. Draft is incomplete genome where as complete genomes are very less number. The performance of tools used for genome assembly differes a lot on the basis of speed, scalability and quality[1-3]. There are thousands of tools and softwares freely available for genome assembly but it is knowledge and experience of user to select best suited software. Here, we have provided a short list of available softwares that are commonly used for genome assembly of different types of genomes such as small, medium or large genomes (Table 2).

**Table 2:** List of freely available and commonly used genome assemly softwares along with their URL and brief description.

| S. No. | Tool/Software | Link | Description |
|---|---|---|---|
| 1 | Ray | https://sourceforge.net/projects/denovoassembler/ | De novo genome assembler |
| 2 | SSAKE | http://www.bcgsc.ca/platform/bioinfo/software/ssake | De novo genome assembly With short reads |
| 3 | ABySS | http://www.bcgsc.ca/platform/bioinfo/software/abyss | De novo genome assembly With short PE-reads |
| 4 | DISCOVER de novo | https://software.broadinstitute.org/software/discovar/blog/ | A large and small genome assembler |
| 5 | SOAPDENOVO | https://github.com/aquaskyline/SOAPdenovo2 | Short read de novo assembler |
| 6 | HGAP | https://github.com/ben-lerch/HGAP-3.0 | High quality microbial genome assembler |
| 7 | VirusTap | https://gph.niid.go.jp/cgi-bin/virustap/index.cgi | Virus genome-targeted assembly pipeline |
| 8 | ALLPATHS-LG | ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/ | Assembles whle genome shortgun sequence |
| 9 | Velvet | http://www.molecularevolution.org/software/genomics/velvet | De novo genome assemble for small genome |
| 10 | MaSuRCA | http://www.genome.umd.edu/masurca.html | Allows assembly of variable length reads |

### Annotation of Assembled Genomes

After assembly and selection of best assembly from the filtered reads, one needs to find genes present in the genome. The process of finding genes and its functions in the genome is known as genome annotations. Genome annotation can be broadly grouped in two categories viz., structural annotation and functional annotation. Under structural annotation we look for ORF and its location, gene structure, coding region and its regulatory motifs. In second category i.e., function annotation assigning of biochemical functions, biological functions, regulation of interactions and expression profiles are taken care. For genome annotation also there are hundreds of softwares and pipelines. A brief list of softwares and pipeline is provided in tabular form (Table 3).

**Table 3:** List of genome annotation tools and pipelines, their link and description.

| S. No. | Tools/Pipeline | Link | Description |
|---|---|---|---|
| 1 | Genome sequence annotation server | https://www.gensas.org/ | A pipeline for structural and functional annotation |
| 2 | GATU | https://virology.uvic.ca/virology-ca-tools/gatu/ | Genome annotation transfer utility |
| 3 | Annotation of microbial genome sequence | https://omictools.com/annotation-of-microbial-genome-sequences-tool | Microbial genome annotation |
| 4 | PANTHER | http://www.pantherdb.org/ | Protein ANalysis Through Evolutionary Relationship |
| 5 | Genomation | https://github.com/BIMSBbioinfo/genomation | A R package for analysis and visualization of genome data set |
| 6 | MAKER | http://www.yandell-lab.org/software/maker.html | A genome annotation pipeline |
| 7 | AUGUSTUS | http://augustus.gobics.de/ | Online server for gene annotation |
| 8 | OncoCis | http://149.171.80.192/OncoCis/ | An annotaion tool for cis-regulatory element in cancer |
| 9 | LookSeq | http://www.sanger.ac.uk/science/tools/lookseq | Annotation tool for structural and sequential variation |
| 10 | PlantReactome | http://plantreactome.gramene.org/ | Provides pathway information to support genome annotations |

## DOWNSTREAM ANALYSIS

Based on the question being investigated in the particular genome project downstream analysis is performed such as expression of genes related to secondary metabolites[4]. Comparison of homologous genes across the genomes to get an understanding of evolution of genes can be performed at genome level. Genome wide identification of some peculiar phenomenon in genes and proteins such as prediction of 3D-domain swapping in human genome[5].

### Detail Description of Above Mentioned Steps

Important technologies which in use nowadays are Illumina (www.illumina.com), 454 Life Sciences (http://www.454.com/), Helicos Biosciences (http://seqll.com/) etc. Most of the present technologies have limitation of short read lengths, inability to generate paired-end sequences. Other major issue with plants is polyploidy of genomes, and sometimes very huge genomes. Majority of the genome projects are unique in them as the data structures, genome size, base-composition, repeat content and polymorphism level.

### Pre-Processing and Assembly of Genomic Data

The raw data coming from high throughput sequencer machines needs to be processed and checked for quality before used in assembly and other analysis. One of the most commonly used programe tools for quality checks and processing of reads is Fastx-Tools[6] and FastQC[7]. Fast QC provides quality score for each residue of each read. On the basis of pred score the read positions with low quality could be trimmed and or filtered. Next, the filtered and processed reads are taken for genome or transcriptome assembly.

There are hundreds of genome assembly and annotation software available such as SOAPDENOVO, Velvet etc[8, 9]. Draft genome assembly is an iterative process with several rounds of assembly, evaluation and parameter tweaking. Initial step of assembly is contig building, and after that to combine the contigs in scaffolds long-insert (mate-pair) libraries are used. Genome annotations are influenced by the available gene, ESTs data of that or related organism.

Correctness and accuracy of the genome assemblies are determined by several parameters such as genome coverage i.e., percentage of genome assemble covered in

the assembly is one of the basic parameter. Genome size estimation can be done by two methods viz., C-value and K-mer frequency. Contig N50, scaffold N50, percentage GC content, size of the scaffolds and gap percentages are other parameter to evaluate genome assembly.

N50 is a parameter which tells about the average length of the scaffolds. It is defined as length of contig or scaffold which covers 50% of the assembled genome by adding together all the contigs/scaffold larger than or equal to it, when they are arranged in ascending or descending order. N50 should not be used to validate the assembly it indicate merely contiguity and contains no information on assembly accuracy. Mapping of the paired-end or mate-pair data could be used to detect errors in the assembly. REAPR, SAMTools, amosvalidate pipeline are few of the tools which could be used for mapping of the reads to reference genome[10, 11, 12]. Lower coverage regions or mis-orientaion of reads suggests mis-assemblies on the other hand aberrant insert sizes indicate small insertions or deletions. By considering all the above parameters best assembly could be selected.

**Genome Annotation**

Genome annotation is next, which technically involves finding the genes in whole genome sequences and assigning function with biologically relevant information that could be from Gene Ontology (GO) terms, or KEGG pathways[13, 14]. Genome annotations also include prediction of other functional elements such as regulatory regions.

In case of non model organisms or *de novo* genomes, annotation is often confined to protein-coding sequences (CDS) or transcripts, which could be achieved by several tools such as MAKER-pipeline, AGUSTUS etc[15, 16]. A near complete genome annotation constitutes a considerable effort and requires expertise in bio informatics. Masking of repetitive sequences including low-complexity regions and transposable elements is very essential step before genome annotation. Repeat masking can be achieved by RepeatMasker[17].

Accuracy of gene prediction is key to functional annotation of genes. A typical gene of eukaryoyes and prokaryotes have peculiar features which differentiate each other such as introns, exons etc (Figure 2). Most of the gene prediction methods searches typical gene features such as start and stop sites, exone-intron boundaries to predcit gene on a DNA sequence. The field of gene prediction is an active and important research field in bioinformatics. With the pouring of enormous genomic data in public domain, there is an urgent need for computational tools and approaches that could help in identification of accurate gene location, structure and function. Small genomes, mostly of prokaryotic organism like bacteria, ranges from 0.5 to 10 Mbp (Million base pairs), have high gene density with less than 10% non coding sequences. Prokaryotic genes are without introns with several unique patterns. Intron less genes are unique feature of small genomes. A regulatory region such as ribosomal binding site helps in locating genes.

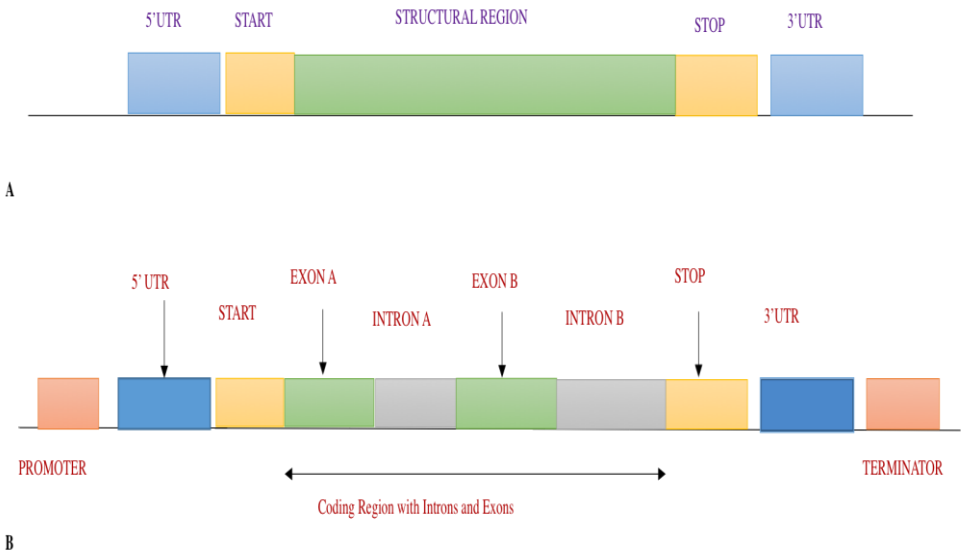


**Figure 2:** A typical A) Prokaryotic and B) Eukaryotic gene with its prominent features used in genome annotation.

It is more complicated to predict gene in eukaryotes, because of complex genomes, than in prokaryotes. Eukaryotic genomes are bigger in size ranging from 10 Mbp to 670 Gbp, with low gene density. There is high percentage of repeat sequences in eukaryotes and the gene is composed of intron and exons. Exons are coding region whereas introns are non-coding region. The primary objective in gene prediction is the identification of exons, introns and functional sites. There are several features such as nucleotide composition, codon bias, and ploy A site signal, used for gene prediction to separate coding and non-coding regions.

**Downstream Analysis**

Once complete proteome of a genome is annotated many downstream analysis could be performed such as finding pathways of important metabolism, comparative analysis across genomes. Knowledge of metabolic pathways,

genes involved and their arrangement in the genome have great implication in genetic engineering of these pathways for the improvement of quality and quantity of the end products. Comparative analysis of these genes across genomes will shed light on the conservation of pathways and genes in particular organism as compared to others.

DNA polymorphism is other aspect wich could be studied in detail with the knowledge of genome and proteome. Polymorphism in plant genomes such as SNP, SSRs could be identified from the genomic data and their effect could be studied in great detail for the improvement of quality and quantity of the product of plant/crop.

## CONCLUSION

Analysis of genomic data requires some exceptional compuational biology skills and expreinece. There are thousands of tools and web-servers freely abvailable to public domain, but it needs a through knowledge for the better selection of the set of tools for a prticular genome project. In general a genome project generates huge amount of data, which needs to be examined carefully to extract some meaning full result and conclusions.

Genomic or NGS data analysis is basically categorised into four major classes viz., pre-processing of raw data, assembling of the good quality data in form a bigger sequences (Scaffolds), Annotation of assembled genomes and down stream analysis. During first step i.e., pre-processing step, raw reads has to be filtered on the basis of phred scores. The low quality reads are discarded and reads with phred score more than 20 are used for further steps. Next step of NGS data analysis in most of the cases is assembly of these good quality reads. Finding genes on the assembled sequences is known as genome annotation and is the third major step of analysis. Accuracy of genome annotation depends on the availability of good gene models and ESTS related to the organism. It is still a daunting task to predict exact exon boundires, which is a open problem in the field of genome annotation. Identifying and locating short exons is another challenge in task in genome annotation as discriminating characteristic features of such exons are not easy in small sequences. Problem magnifies to other level in cases where the coding exon is a multiple of three as missing of such exon would not have any effect on the genome or gene assembly. Alternative splicing of genes also pose severe problem to genome annotation as it may give rise to genes with more than one possible exon assembly. Although some programs like GENESCAN has dealt the issue by identifying sub-optimal exons, the problem remains to be researched. It is need of the hour to devise a comprehensive criterion for assessing the quality of gene prediction programmes.

Furthermore, downstream analysis is based on the question posed in the project and related problems in the concerned field. In this review, we have discussed major steps and issues of NGS data analysis briefly and challenges of gene prediction in the genomic data in deatil.

## REFERENCES


1. Miller JR, Koren S, Sutton G. Assembly algorithm for Next-Ganeration Sequencing data. Genomics. 95(6), 2010, 315–27.
2. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Sébastien B, Jarrod AC, Guillaume C, Rayan C, Hamidreza C, Wen-Chi C, Jacques C, Cristian DF, Roderick DT, Richard D, Dent E, Scott E, Pavel F, Nuno AF, Ganeshkumar G, Richard AG, Sante G, Élénie G, Steve G, Matthias H, Giles H, David H, Joseph BH, Isaac YH, Jason H, Martin H, Shaun DJ, David BJ, Erich DJ, Huaiyang J, Sergey K, Paul JK, Jacob OK, James RK, Sergey K, Tak-Wah L, Dominique L, François L, Yingrui L, Zhenyu L, Binghang L, Yue L, Ruibang L, Iain M, Matthew DM, Nicolas M, Sergey M, Delphine N, Zemin N, Thomas DO, Benedict P, Octávio SP, Adam MP, Francisco PM, Michael P, Dariusz P, Xiang Q, Carson Q, Filipe JR, Stephen R, Daniel SR, Graham RJ, Simone S, Michael C, David CS, Alexey S, Ted S, Timothy IS, Jay S, Yujian S, Jared TS, Henry S, Fedor T, Francesco V, Riccardo V, Bruno MV, Jun W, Kim CW, Shuangye Y, Siu-Ming Y, Jianying Y, Guojie Z, Hao Z, Shiguo Z, Ian FK. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience. 2(1), 2013, 10.
3. Narzisi G, Mishra B. Comparing De Novo genome assembly: The long and short of it. PLoS One. 6(4), 2011.
4. Upadhyay AK, Chacko AR, Gandhimathi A, Ghosh P, Harini K, Joseph AP, Adwait GJ, Snehal DK, Swati K, Nagesh K, Chandana SL, J. Mahita, Ramya M, Sony M, Manoharan M, Oommen KM, Eshita M, Mahantesha N, Sathyanarayanan N, Shaik NP, Upadhyayula SR, Anantharamanan R, Shilpa S, Prashant NS, Heikham RS, Anshul S, Margaret SS, Manojkumar S, Ramaswamy S, Malali G, Ramanathan S. Genome sequencing of herb Tulsi (Ocimum tenuiflorum) unravels key genes behind its strong medicinal properties. BMC Plant Biol. 15(1), 2015, 212.
5. Upadhyay AK, Sowdhamini R. Genome-wide prediction and analysis of 3D-domain swapped proteins in the human genome from sequence information. PLoS One. 11(7), 2016, 1–20.
6. Gordon A, Hannon GJ, Gordon. FASTX-Toolkit. [Online] http://hannonlab. cshl. edu/fastx_ toolkit http://hannonlab. cshl. edu/fastx_ toolkit. 2014.
7. Andrews S. FastQC: A quality control tool for high throughput sequence data. Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. 2010.
8. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1(1), 2012 Jan, 18.
9. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18(5), 2008, 821–9.
10. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. BioMed Central Ltd; 14(5), 2013, R47.

11. Heng Li1, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan NH, Gabor Marth, Goncalo Abecasis RD. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16), 2009, 2078–9.

12. Schatz MC, Delcher a L, Roberts M, Marcçais G, Pop M, Yorke J a. GAGE: A critical evaluation of genome assemblies and assembly algorithms Steven L. Salzberg, Adam M. Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J. Treangen. Genome Res. 22, 2012, 557–67.

13. Harris M a, Clark J, Ireland a, Lomax J, Ashburner M, Foulger R. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32(Database issue), D 2004 Jan 1, 258-61.

14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1), 2000 Jan, 27–30.

15. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Carson Holt, Alejandro Sánchez Alvarado, Mark Yandell. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1), 2008 Jan; 188–96.

16. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005 Jul 1, 33(Web Server issue):W465-7.

17. Smit, AFA, Hubley, R & Green PR. RepeatMasker Open-3.0.;1996-2010 http://www.repeatmasker.org [Internet]. 2010 [cited 2014 Aug 5]. Available from: http://www.repeatmasker.org/faq.html#faq3

**Source of Support:** Nil, **Conflict of Interest:** None.