Neuberg DIAGNOSTICS | CENTER FOR GENOMIC MEDICINE

BIOINFORMATICS DATA ANALYSIS

Project Deliverables

List of deliverables can be changed,
please reach out to project managers for the custom requirements

Prokaryotic de-novo genome assembly and annotation

Requirements to be provided by client:
1. Organism's scientific name
2. NCBI IDs of the organisms for Phylogenetic and PAN/Core genome analysis

Any suggestions for using specific tools should be mentioned prior to the start of analysis. No further changes will be entertained during the analysis runs.

Note: Read level taxonomy profiling of the processed sample will be done and KRONA plot will be provided. Downstream analysis will be done if at least 60% of the reads support the concerned organism. Otherwise a decision will be taken after discussion with Client.

Deliverables:

- Individual Raw FASTQ FastQC report
- Raw FASTQ MultiQC report
- Individual processed FASTQ FastQC report
- Processed FASTQ MultiQC report
- Assembly completeness and contamination analysis
- De Novo assembly quality assessment report
    - Total number of scaffolds
    - Total length (bp)
    - Max length (bp)
    - N50 (bp)
    - N90 (bp)
    - L50(bp)
    - G+C content (%)

7) Taxonomy analysis
    - 16SRNA sequences
    - Taxonomy assignment through 16s Sequence
    - Taxonomy assignment through Average Nucleotide Identity (ANI)

8) Genome Annotation
    - Genome Annotation
    - Comprehensive protein annotation
    - Number of Hypothetical proteins
    - Number of Proteins with EC number assignments
    - Number of Proteins with GO assignments
    - Number of Proteins with pathway assignments

9) Phylogenetic Analysis

10) Circos plot

Customised Analysis

1. PAN/Core genome analysis
2. Prediction ofIdentification AMR / Virulence factors / Bacterial toxins / Pathogenetic Islands
3. Annotation of AMR genes and representation of genes in circular plot
4. Annotation of secondary metabolites

Eukaryotic de-novo genome assembly and annotation

Requirements to be provided by client:

1. Organism's scientific name

Any suggestions for using specific tools should be mentioned prior to the start of analysis. No further changes will be entertained during the analysis runs.

Note: Read level taxonomy profiling of the processed sample will be done and KRONA plot will be provided. Downstream analysis will be done if at least 60% of the reads support the concerned organism. Otherwise a decision will be taken after discussion with Client.

Deliverables:

- Individual Raw FASTQ FastQC report
- Raw FASTQ MultiQC report
- Individual processed FASTQ FastQC report
- Processed FASTQ MultiQC report
- Assembly completeness analysis
- De Novo assembly quality assessment report
    - Total number of scaffolds
    - Total length (bp)
    - Max length (bp)
    - N50 (bp)
    - N90 (bp)
    - L50(bp)
    - G+C content (%)

- Genome Annotation

Reference-based RNASeq Expression Analysis

Mandatory Requirements to be provided by client:

a. RefSeq / GenBank Accession Number of the Reference Genome

b. In silico rRNA depletion will be carried out in case of no depletion or enrichment step was completed before sequencing. The reference annotation file will be edited to remove all rRNA (as well as tRNA) sequence annotations, to prevent these features from being counted.

c. Sample Grouping information

d. Each condition must have at least one replicate

e. Any suggestions for using specific tools should be mentioned prior to the start of analysis.

f. No further changes will be entertained during the analysis runs.

Example of sample group information file

Sample	Condition
SampleA	Control
SampleB	Control
SampleC	Treated
SampleD	Treated

NOTE for Sample Name:

1. Sample must not start with a number
2. No special characters like "-", "_", space are allowed
3. Sample Name must be unique
4. There must be exactly two groups for condition column (single group or more than 2 group expression analysis is not possible

Note: In case of conditions without replicates, edgeR with exactTest function dispersion guess at 0.01 will be performed.

Deliverables:

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Mapping statistics of individual samples with genome
6. Raw readcount (featureCounts) of individual samples
7. Fragments per million mapped fragments (FPM) in DESeq2
8. &#39;regularized log&#39; transformation (rlog) transformed file in DESeq2
9. Principal component analysis (PCA) plot generated in DEseq2 showing variation within and between groups
10. Sample to Sample distance cluster plot based on their euclidean distance using the regularized log transformed count data
11. Complete set of upregulated and down regulated gene counts will be provided in .xls format.
12. Annotated DESeq2 result file with Chromosome Number, Gene name, Gene Description, start and end position, count values of individual samples, unadjusted p-value, adjusted p-value, fold change, log2foldchange

NOTE: Annotated DESeq2 result will be filtered based on the following 6 cutoffs

1. FDR &lt;0.05 AND LOG2FC 2
2. FDR &lt;0.05 AND LOG2FC 1.5
3. FDR &lt;0.05 AND LOG2FC 1
4. Uncorrected p value &lt;0.05 AND LOG2FC 2
5. Uncorrected p value &lt;0.05 AND LOG2FC 1.5
6. Uncorrected p value &lt;0.05 AND LOG2FC 1

Clients needs to suggest one cut-off value, based on which downstream analysis will be performed

1. Heatmap of 50 most variable genes / transcripts
1. MA plot (based on suggested statistical test and LogFC)
1. Volcano plot (based on suggested statistical test and LogFC)
1. For organisms (if NCBI GeneID is available at NCBI gene2go list), GESA analysis will be performed against GO, KEGG and Reactome database

Exclusively protein coding mRNA and lncRNA based expression analysis (applicable for mouse and human samples)

Mandatory Requirements

1. Sample Grouping information

Example of sample group information file

Sample	Condition
SampleA	Control
SampleB	Control
SampleC	Treated
SampleD	Treated

1. Each condition must have at least one replicate
3) Any suggestions for using specific tools should be mentioned prior to the start of analysis.
4) No further changes will be entertained during the analysis runs.

Note: In case of conditions without replicates, edgeR with exactTest function dispersion guess at 0.01 will be performed.

Method:
Protein-coding transcript sequences and lncRNA transcript sequences will be obtained from GENCODE and quantification will be done using KALLISTO .

Deliverables:

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Summary of rRNA filtration
6. TPM counts of individual samples
7. 'regularized log'; transformation (rlog) transformed file in DESeq2.
8. Annotated DESeq2 result with Chromosome Number, Gene name, Gene Description, start and end position, unadjusted p-value, adjusted p-value, fold change, log2foldchange will befiltered based on the following 6 cutoffs and shared as an interim report. Client needs to suggest **one** cut-off value, based on which downstream analysis will be performed

    1. FDR &lt;0.05 AND LOG2FC 2
    2. FDR &lt;0.05 AND LOG2FC 1.5
    3. FDR &lt;0.05 AND LOG2FC 1
    4. Uncorrected p value <0.05 AND LOG2FC 2
    5. Uncorrected p value <0.05 AND LOG2FC 1.5
    6. Uncorrected p value <0.05 AND LOG2FC 1

1. Principal component analysis (PCA) plot generated in DEseq2 showing variation within and between groups
1. Sample to Sample distance cluster plot based on their euclidean distance using the regularized log transformed count data.
1. Heatmap of 50 most variable genes / transcripts
1. MA plot (based on suggested statistical test and LogFC)
1. Volcano plot (based on suggested statistical test and LogFC)
1. GESA analysis will be performed against GO, KEGG and Reactome database

De novo RNAseq Analysis of non-model organisms

Mandatory Requirements to be provided by client:

1. Organism's scientific name
2. NCBI Taxonomy ID of the Order / Family / Genera based on which annotation need to done against NCBI's NRDB
3. Sample Grouping information
4. Each condition must have at least one replicate
5. Any suggestions for using specific tools should be mentioned prior to the start of analysis.
6. No further changes will be entertained during the analysis runs.

Method:

1. De Novo transcriptome assembly will be done using Trinity
2. Clustering of transcripts will be done by evidential gene pipeline
3. Annotation will be done using BLASTX against a. NCBI's NRDB (Client's specified taxonomy ID), b. UniProt and c. InterproScan

Deliverables:

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Summary of rRNA filtration
6. Corset counts of individual samples (or any alternate tool)
7. “regularised log” transformation (rlog) transformed file in DESeq2
8. Principal component analysis (PCA) plot generated in DEseq2 showing variation within and between groups
9. Sample to Sample distance cluster plot based on their euclidean distance using the regularised log transformed count data.
10. Annotated DESeq2 results with Gene name, Gene Description, subject name, GO terms, unadjusted p-value, adjusted p-value, fold change, log2foldchange will be filtered based on the following 6 cutoffs and shared as an interim report. Client needs to suggest **one** cut-off value, based on which downstream analysis will be performed

    1. FDR &lt;0.05 AND LOG2FC 2
    2. FDR &lt;0.05 AND LOG2FC 1.5
    3. FDR &lt;0.05 AND LOG2FC 1
    4. Uncorrected p value <0.05 AND LOG2FC 2
    5. Uncorrected p value <0.05 AND LOG2FC 1.5
    6. Uncorrected p value <0.05 AND LOG2FC 1

1. Heatmap of 50 most variable genes / transcripts
1. MA plot (based on suggested statistical test and LogFC)
1. Volcano plot (based on suggested statistical test and LogFC)

Identification and Annotation of PHAGE From WGS data

Method:

1. The processed WGS data will be subjected to genome detective online server for detection and classification of viral reads.
2. Samples with detected PHAGE, will be subjected to Assembly using unicycler
3. Unicycler assembly will be subjected to PHASTER for annotation of PHAGE

Deliverables:

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Genome Detective report for individual samples
6. Unicycler assembled FASTA
7. QUAST analysis summary report
8. PHASTER analysis results

METAGENOME BASED SERVICES & DELIVERABLES

16S Analysis

Requirements

1. Sample Grouping information
2. Any suggestions for using specific tools should be mentioned prior to the start of analysis.
3. No further changes will be entertained during the analysis runs.

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Taxonomic bar plots across samples at Phylum, Class, Order, Family, Genus and Species level
6. OTU heatmaps across samples at Phylum, Class, Order, Family, Genus and Species level
7. Rarefaction curve
8. Alpha and Beta Diversity
9. Krona plots (samplewise and groupwise)

Custom Analysis

1. Picrust Functional Analysis
2. Linear Discriminant Analysis (not applicable for single sample comparison)
3. OTU Network Analysis
4. Group wise comparison at Phylum, Class, Order, Family, Genus and Species level (statistical testing)

NOTE : In case of statistical testing, it will be considered as secondary analysis.

Mandatory requirement for Statistical testing:

1. Each condition must have at least one replicate.
2. In case of conditions without replicates
   a- Two Sample Fisher’s exact test will be performed for functional analysis
   b- metstat will be performed for differential abundance testing

18S/ITS2 Sequence Analysis

Requirements

1. Sample Grouping information
2. Each condition must have at least one replicate
3. Any suggestions for using specific tools should be mentioned prior to the start of analysis.
4. No further changes will be entertained during the analysis runs.

Note: In case of conditions without replicates, Two Sample Fisher&#39;s exact test will be performed.

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Taxonomic bar plots across samples at Phylum, Class, Order, Family, Genus and Species level
6. OTU heatmaps across samples at Phylum, Class, Order, Family, Genus and Species level
7. Rarefaction curve
8. Alpha and Beta Diversity
9. Krona plots (samplewise and groupwise)

Custom Analysis

1. Picrust Functional Analysis
2. Linear Discriminant Analysis
3. OTU Network Analysis
4. Group wise comparison at Phylum, Class, Order, Family, Genus and Species level

Prokaryotic de-novo genome assembly and annotation (Applicable for pure bacterial cultures)

Requirements to be provided by client:

1. Organism Name
2. NCBI IDs of the organisms for Phylogenetic and PAN/Core genome analysis
3. Any suggestions for using specific tools should be mentioned prior to the start of analysis.
4. No further changes will be entertained during the analysis runs.

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Assembly completeness and contamination analysis
6. De Novo assembly quality assessment report
   - Total number of scaffolds
   - Total length (bp)
   - Max length (bp)
   - N50 (bp)
   - N90 (bp)
   - L50(bp)
   - G+C content (%)

7) Taxonomy analysis
   - 16SRNA sequences
   - Taxonomy assignment through 16s Sequence
   - Taxonomy assignment through Average Nucleotide Identity (ANI)

8) Genome Annotation
   - Genome Annotation using PROKKA
   - Comprehensive protein annotation
   - Number of Hypothetical proteins
   - Number of Proteins with EC number assignments
   - Number of Proteins with GO assignments
   - Number of Proteins with pathway assignments

9) Phylogenetic Analysis
10) Circos plot

Note: In case of conditions without replicates, Two Sample Fisher&#39;s exact test will be performed.

Custom Analysis

1. PAN/Core genome analysis
1. Prediction of AMR / Virulence factors / Bacterial toxins / Pathogenetic Islands
1. Annotation of AMR genes and representation of genes in circular plot
1. Identification and annotation of secondary metabolites

Whole Metagenome (Shotgun) Analysis

Requirements

1. Sample Grouping information
2. Each condition must have at least one replicate

a. Note: In case of conditions without replicates, Two Sample Fisher's exact test will be performed.

2. Any suggestions for using specific tools should be mentioned prior to the start of analysis.
3. No further changes will be entertained during the analysis runs.

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Metagenome assembly
6. Binning of assembled metagenome
7. Quality assessment of the genome bins
8. Screening and selection of bins
9. Annotation of the selected bins
10. Taxonomy identification
11. Functional enrichment
12. Taxonomic bar plots, heatmaps

Custom Analysis

1. Antimicrobial resistance screening (Based on client requirement)
2. Virulence screening
3. Toxicity assessment
4. Secondary metabolite screening
5. Differential enrichment analysis

miRNA analysis

Requirements

a. Samples with replicates

b. Control

c. Reference organism Genbank/Refseq/ENSEMBL ID

d. Sample grouping information

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. miRNA quality assessment statistics
6. Identification of novel and known miRNA (R1 reads will be used)
7. Absolute miRNA counts (R1 reads will be used)
8. Differential expression analysis (R1 reads will be used). For samples without replicates, edgeR with exactTest will be performed. Alternatively for samples with replicates, DESeq2 will be performed.
9. Heatmap of top 20 up and down-regulated gene counts (R1 reads will be used)
10. MA plot (R1 reads will be used)
11. Volcano plot (R1 reads will be used)
12. miRNA target enrichment (R1 reads will be used). Target enrichment will be done forhuman, mouse, rat, Danio rario, Drosophila melanogaster, C elegans.

Note: Standard deliverables applicable for species listed at mirbase (https://mirbase.org/browse/)

ChIP-Seq analysis

Requirements

a. Samples with replicates
b. Control
c. Antibody information
d. Reference organism Genbank/Refseq/ENSEMBL ID
e. Sample grouping information

Example for sample grouping information

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Alignment summary
6. Chip-Seq QC Metrics
7. Peak Calls (MACS2 --broad parameter)
8. Peak quantification
9. Differential binding analysis
10. MA Plot
11. Volcano Plot
12. PCA plot
13. Sample to sample distance

Bisulfite Sequencing

Requirements

1. Reference genome ENSEMBL/RefSeq ID
2. Illumina paired-end reads

Standard Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Alignment:
    1. Bismark output.
        a. Aligned reads in BAM format.
        b. Log file giving summary statistics about alignment.
    2. BWA-meth output
        a. Aligned reads in a sorted BAM file.
        b. Summary file giving metrics about the aligned BAM file.

1. Methylation Extraction:
    - Coverage text file summarising cytosine methylation values.
    - Methylation statuses in bedGraph format, with 0-based genomic start and 1-based end coordinates.
    - QC data showing methylation bias across read lengths.

1. Biskmark Report: Bismark generates a HTML reports describing results for each sample, as well as a summary report for the whole run

Bulk ATAC sequencing

Requirements

1. ATAC-seq paired-end FastQ reads
2. Sample Grouping Information
3. Reference Organism: Human/Mouse

Deliverables

1. Individual Raw FASTQ FastQC report
2. Raw FASTQ MultiQC report
3. Individual processed FASTQ FastQC report
4. Processed FASTQ MultiQC report
5. Alignment
6. Peak Calls
7. Peak Annotation relative to gene features
8. Consensus peakset
9. Functional Analysis

MeDIP sequencing

1. Raw and processed data QC Reports
2. Alignment against reference genome
3. Whole genome region distribution
4. Peak distribution statistics
5. Peak associated gene screening
6. Functional annotation
7. Differential expression level analysis of peak associated genes

dd-RAD sequencing (GBS)

Requirements

1. Reference genome scientific name for reference based variant calling
2. sample to population mapping file (for population diversity analysis)

Deliverables

1. FASTQC and multiQC reports of raw data
2. FASTQC and multiQC reports of processed data
3. Filtered VCF files
4. Population genetics summary statistic
5. Principal Component Analysis

NOTE: Deliverables 4 and 5 will be applicable if sample to population mapping file isprovided.

dd-RAD sequencing (GBS) with population diversity and GWAS

Requirements

1. Reference genome scientific name for reference based variant calling
2. Sample to population mapping file (for population diversity analysis)

Deliverables

1. FASTQC and multiQC reports of raw data
2. FASTQC and multiQC reports of processed data
3. Filtered VCF files
4. Population genetics summary statistic
5. Principal Component Analysis
6. Phenotype Distribution Plot
7. SNP Density Plot
8. PCA Plot
9. Manhattan plot in Circular fashion
10. Manhattan plot in Rectangular fashion for single trait
11. Manhattan plot in Rectangular fashion for multiple traits
12. Q-Q plot for single trait
13. Q-Q plot for multiple traits
14. FarmCP based SNP statistics (p-values and effect)
15. General linear model (GLM) SNP statistics (p-values and effect)

NOTE:

1. Deliverables 4 and 5 will be applicable if sample to population mapping file is provided.

2. Deliverables 14 and 15 applicable for phenotype file with multiple trait

GWAS analysis

Requirements

1. phenotype file
2. variant call file in VCF / PLINK binary / HAPMAP format
3. Kinship matrix (optional)

Deliverables

1. Phenotype Distribution Plot
2. SNP Density Plot
3. PCA Plot
4. Manhattan plot in Circular fashion
5. Manhattan plot in Rectangular fashion for single trait
6. Manhattan plot in Rectangular fashion for multiple traits
7. Q-Q plot for single trait
8. Q-Q plot for multiple traits
9. FarmCP based SNP statistics (p-values and effect)
10. General linear model (GLM) SNP statistics (p-values and effect)

Note: Deliverables 6 and 8 applicable for phenotype files with multiple traits

WES sequencing

1. QC analysis of raw read data
2. Mapping of reads to reference genome using BWA
3. Aligenement Statistics
4. SNP and InDel detection based on the mapping results (based on GATK pipeline)
5. Annotation of all variants
6. Filtration of variants based on client’s proposed parameters

Requirements: BED file

Cut&Run / Cut&Tag

Requirements

Control Dataset (FASTQ)

Deliverables

1. Individual raw and processed data QC reports
2. Target genome alignment summary
3. Spike-in genome alignment summary
4. bedGraph files
5. bigWig coverage files
6. Peak Calls (SEACR)
7. Heatmap peak analysis
8. Genome browser session (IGV)
9. peak-based QC reports

Neuberg DIAGNOSTICS | CENTER FOR GENOMIC MEDICINE

NEUBERG CENTER FOR GENOMIC MEDICINE (NCGM)

(A Unit Of Neuberg Supratech Reference Laboratories Private Limited)

GTPL House Lane, Near East Ebony, Sindhu Bhavan Road, Bodakdev,
Ahmedabad -380059 | Phone : 079-61618111, 6357244307
Email : contact@ncgmglobal.com | Website : www.ncgmglobal.com