# The common procedures of bioinformatic analysis

## Contents

Part I Projects

# 1 Meta-analysis

## 1.1 Introduction to Meta-analysis

Meta-analysis refers to methods focused on contrasting and combining results from different studies, in the hope of identifying patterns among study results, sources of disagreement among those results, or other interesting relationships that may come to light in the context of multiple studies. Typically, meta-analysis is used:

1) To increase statistical power for primary end points and for subgroups;
2) To improve estimates of the size of the effect;
3) Resolve uncertainty when reports disagree;
4) Deduce new conclusion according to subgroup analysis;
5) Find new hypothesis.

Meta-analysis is now a basic statistic method for systemic review of research articles in evidence based medicine.

### 1.1.1 Advantages

Advantages of meta-analysis (e.g. over classical literature reviews, simple overall means of effect sizes etc.) include:

- Shows if the results are more varied than what is expected from the sample diversity
- Derivation and statistical testing of overall factors / effect size parameters in related studies
- Generalization to the population of studies
- Ability to control for between-study variation
- Including moderators to explain variation
- Higher statistical power to detect an effect than in 'n=1 sized study sample'
- Deal with information overload: the high number of articles published each year
- It combines several studies and will therefore be less influenced by local findings than single studies will be.
- Makes it possible to show if a publication bias exists.
- Using information from published studies sufficiently, most work will be done with statistical analysis, resulting in shorter project period.

### 1.1.2 Research pipeline

- Study design

- Search of literature
- Selection of studies
- Decide which dependent variables or summary measures are allowed
- Assessment and description of each study
- Statistical analysis
  - Heterogeneity analysis
  - Effect size analysis
  - Combined results illustration
  - Sensitivity analysis
  - Publication bias analysis
- Results interpretation and meta-analysis quality evaluation

## 1.2 Meta-analysis of association studies

### 1.2.1 Significance

Single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide — A, T, C or G — in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual. SNPs are one of the most common types of genetic variation. The increasing ease and plummeting costs of genotyping and sequencing make it easy to conduct association studies which aim to find disease related polymorphisms. Nevertheless, statistical power of single independent study is limited to its sample size and other factors. Inevitably, inconsistence among different studies exists. Meta-analysis of several SNP based association studies would reach higher statistical power, providing better insights into the mechanism for the pathogenesis of complex disease. Consequently, meta-analysis will help identify disease related locus precisely, constituting a great advance toward new molecular diagnostic tests for the development of potential therapies.

### 1.2.2 Research contents

After searching and selecting SNP based association studies, conduct meta-analysis and provide weighted average results. Specific aims are as follows:

- Evaluate the reliability of previous studies;
- Seek new disease related loci and mechanisms.

### 1.2.3 Examples

Dunlop *et al.*[1] (2012) performed a meta-analysis of five genome-wide association studies to identify common variants influencing colorectal cancer (CRC) risk comprising 8,682 cases and 9,649 controls. Replication analysis was performed in case-control sets totaling 21,096 cases and 19,555 controls. We identified three new CRC risk loci at 6p21 (rs1321311, near CDKN1A; P = 1.14 × 10(-10)), 11q13.4 (rs3824999, intronic to POLD3; P = 3.65 × 10(-10)) and Xp22.2 (rs5934683, near SHROOM2; P = 7.30 × 10(-10)) This brings the number of independent loci associated with CRC risk to 20 and provides further insight into the genetic architecture of inherited susceptibility to CRC. Their work was published on Nature Genetics (IF=36.377).

## 1.3 Meta-analysis of gene expression studies

### 1.3.1 Significance

In the field of molecular biology, gene expression profiling is the measurement of the activity (the expression) of thousands of genes at once, to create a global picture of cellular function. These profiles can, for example, distinguish between cells that are actively dividing, or show how the cells react to a particular treatment. As the costs of gene expression microarray and RNA sequencing plummeting, gene expression studies which aim to find differentially expressed genes in different tissue (e.g. tumor and normal) become popular. Nevertheless, statistical power of single independent study is limited to its sample size and other factors. Inevitably, inconsistence among different studies exists. Meta-analysis of gene expression studies would reach higher statistical power, providing better insights into the mechanism for the pathogenesis of complex disease. Consequently, meta-analysis will help identify disease related genes precisely, constituting a great advance toward new molecular diagnostic tests for the development of potential therapies.

### 1.3.2 Research contents

After searching and selecting gene expression studies, conduct meta-analysis and provide weighted average results. Specific aims are as follows:

- Evaluate the reliability of previous studies;
- Seek new disease related genes and mechanisms.

### 1.3.3 Examples

Kavak *et al.*[2] applied a novel meta-analysis approach to multiple sets of merged serial analysis of gene expression and microarray cancer data in order to analyze transcriptome alterations in human cancer. They  identified 81 co-regulated regions on the human genome (RIDGEs) . These findings engender a deeper understanding of cancer biology by demonstrating the existence of a pool of under-studied multi-cancer genes and by highlighting the cancer-specificity of some TA-RIDGEs. Their work was published on Nucleic Acids Research (IF=7.48).

# 2 Integration analysis

## 2.1 Introduction to integration analysis

Integration analysis refers to reanalyzing combined data of published studies, in the hope of overcoming previous bias result from small sample size of independent study or analytical problems due to limited computing resource. Different from meta-analysis, integration analysis involves raw data rather than statistical results of previous studies. Specific aims are as follows:

1) Increase statistical power with enlarged sample size;
2) Design new analysis pipeline to seek new findings;
3) Evaluate the reliability of previous studies.

### 2.1.1 Advantages

Advantages of integration analysis (e.g. over classical literature reviews, single independent study etc.) include:

- Higher statistical power based on combined data from previous studies with enlarged sample size;
- Evaluate the inconsistence among published studies with better analysis pipeline and combined data set;
- Enlarge the possibility of unravel new molecular mechanism underlying complex disease;
- Using data from published studies sufficiently; □
- Most work will be done with statistical analysis, resulting in shorter project period.

### 2.1.2 Research pipeline

- Study design
- Search of literature
- Selection of studies
- Download raw data from databases
- Statistical analysis
    - ◆ Identify disease related loci
    - ◆ Identify new molecular mechanism
- Results interpretation and integration analysis quality evaluation

## 2.2 Integration analysis of association studies

### 2.2.1 Significance

Statistical power of single independent study is limited to its sample size and other factors. Inevitably, inconsistence among different studies exists. Integration analysis of several SNP based association studies would reach higher statistical power based on combined raw data of previous studies, providing better insights into the mechanism for the pathogenesis of complex disease. Consequently, integration analysis will help identify disease related locus precisely, constituting a great advance toward new molecular diagnostic tests for the development of potential therapies.

### 2.2.2 Research contents

After downloading raw data from previous SNP based association studies, conduct integration analysis and provide new results. Specific aims are as follows:

- Evaluate the reliability of previous studies;
- Seek new disease related loci and mechanisms.

### 2.2.3 Examples

Tuna *et al.*[3] (2010) retrieved large genomic data sets from the Gene Expression Omnibus (GEO) database to perform genome-wide analysis of aUPD in breast tumor samples and cell lines using approaches that can reliably detect aUPD, providing valuable insights into breast tumorigenesis. aUPD was identified in 52.29% of the tumor samples. The most frequent aUPD regions were located at chromosomes 2q, 3p, 5q, 9p, 9q, 10q, 11q, 13q, 14q and 17q. Their work was published on PLoS one (IF=4.092).

## 2.3 Integration analysis of gene expression studies

### 2.3.1 Significance

Statistical power of single independent study is limited to its sample size and other factors. Inevitably, inconsistence among different studies exists. Integration analysis of several gene expression studies would reach higher statistical power based on combined raw data of previous studies, providing better insights into the mechanism for the pathogenesis of complex disease. Consequently, integration analysis will help identify disease related genes precisely, constituting a great advance toward new molecular diagnostic tests for the development of potential therapies.

### 2.3.1 Research contents

After downloading raw data from previous gene expression studies, conduct integration analysis and provide new results. Specific aims are as follows:

- Evaluate the reliability of previous studies;
- Seek new disease related loci and mechanisms.

### 2.3.2 Examples

With a novel gene signature derived from mouse models, He *et al.*[4] (2009) conducted integration analysis on breast cancer and lung cancer data sets downloading from the GEO database and determined six-gene model in predicting survival in breast cancer patients. In addition, the model was able to stratify poor from good prognosis for lung cancer patients in majority of the datasets analyzed. Their work was published on Clinical Cancer Research (IF=7.742).

## 3 Omics analysis

### 3.1 Significance

In molecular biology, the terminology omics refers to a totality of some sort, including genomics, proteinomics, metabolomics, transcriptomics, lipidomics, immunomics, glycomics and RNomics. As the development of scientific research, studies on single field or single level (e.g. genomics) can’t solve all medical problems. Omics analysis aims to combine the objects of different fields and decode the interactions among genes, molecular and proteins, thus providing new insights into the exploration of pathogenesis for all kinds of diseases.

### 3.2 Research contents

Analyze data sets at different levels (DNA, RNA and protein) for a specific disease combined with corresponding clinical data, which including family history, symptoms and prognosis, for following aims:

- Identify disease related genes;
- Investigate the relationships of genes, RNA and proteins, explore mechanisms under organ metabolism and etiopathology.

### 3.3 Advantages

Compared with studies on single level, advantages of omics analysis are as follows:

- Evaluate previous studies from the view of different biological levels;
- Explore new mechanism for disease prevention and treatments with data sets from different levels;
- Propose new prospectives for further studies;
- Use published data sets and clinical data sufficiently; □
- Most work will be done with statistical analysis, resulting in shorter project period.

### 3.4 Research pipeline

- Study design
- Search of literature
- Selection of studies
- Download data sets and clinical data from databases
- Statistical analysis
    - ◆ Inference of possible disease related loci using different models;
    - ◆ Validate the effects of inferred sites with data sets from different biological levels;
    - ◆ Using clinical data, explore underlying mechanism for possible

prognosis.

- Results interpretation and omics analysis quality evaluation

## 3.5 Examples

Nibbe *et al*.[5] (2010) demonstrated that integration of mRNA expression data and proteomic profiling data sources with a "proteomics-first" approach can enhance the discovery of candidate sub-networks in cancer that are well-suited for mechanistic validation in disease. With the two data set from GEO (GSE10950 and GSE8671) and proteomic data, they identify new network for regulating human colorectal cancer. Their work was published on PLoS Computational Biology (IF=5.215).

Part II Bioinformatics analysis

# 1 SNP genotyping array

## 1.1 Overview

As a variation at a single site in DNA, SNP is the is the most frequent type of variation in the genome. For example, there are around 50 million SNPs that have been identified in the human genome. Genotypes of different samples can be determined through SNP array. The basic principles of SNP array are the same as the DNA microarray. These are the convergence of DNA hybridization, fluorescence microscopy, and solid surface DNA capture.

## 1.2 Workflow

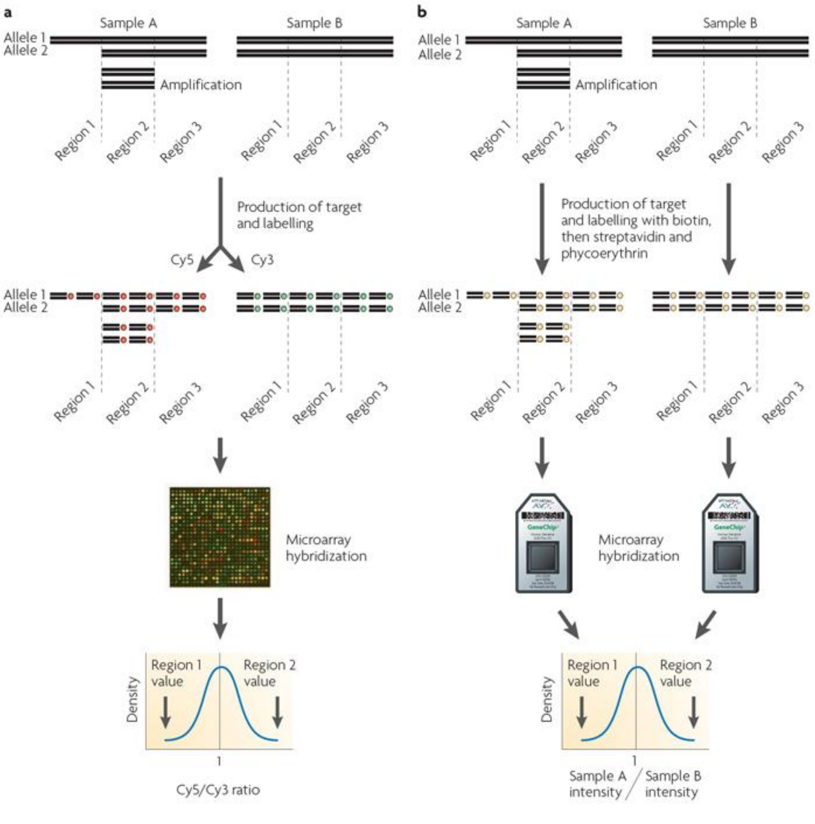
a
Sample A
Sample B
Allele 1
Allele 2
Amplification
Region 1
Region 2
Region 3
Production of target and labelling
Cy5
Cy3
Microarray hybridization
Region 1 value
Region 2 value
Density
1
Cy5/Cy3 ratio
b
Production of target and labelling with biotin, then streptavidin and phycoerythrin
Sample A intensity / Sample B intensity

Figure 1- 1 Workflow for SNP genotyping

## 1.3 Bioinformatics analysis

### 1.3.1 Standard analysis

- Study design according to diseases and available samples;

- Quality control (QC) for samples: according to call rate, genotyping concordance rate and contamination rate, remove duplicated samples and samples failed to be genotyped; check the gender for each sample;
- QC for sites: Preprocessing SNP sites according to call rate, Minor Allele Frequency (MAF) and Hardy-Wenberg Equilibrium (HWE) .

### 1.3.2 Advanced analysis

- Association analysis: Pipeline is illustrated in figure 1-2.
  a) Stratification analysis: Use various statistical methods to check population structure; adjust for population stratification to avoid false positive and negative results in association analysis.
  b) Association analysis (single locus, multiple loci, haplotype): Association analysis with allelic, genotypic, dominant and recessive models for qualitative or quantitative traits. Example is illustrated in figure 1-3.
  c) Evaluation for significant results: Evaluate significant sites according to genotyping quality, MAF and HWE.
  d) Annotation for significant loci；
  e) Fine mapping analysis: Select tagSNP for candidate regions and construct haplotype (illustrated in Figure1-3);
- Linkage analysis: linkage analysis can be done for pedigree samples to identify risk sites for Mendel disease. (Figure 1-4).

Figure 1- 2 Pipeline for association analysis

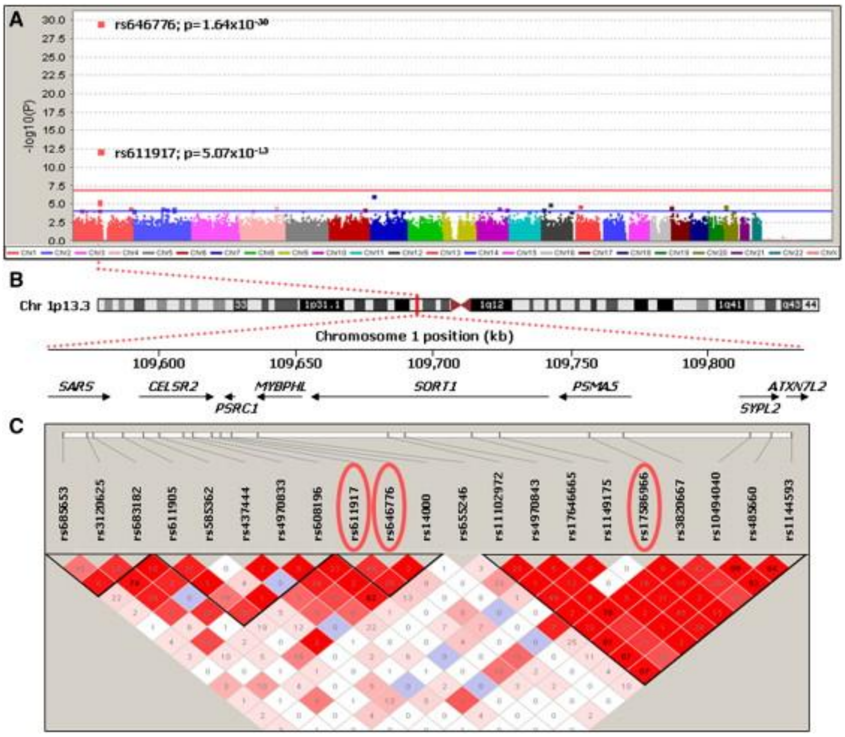

Figure 1- 3 Paradigm for haplotype and association results[6]

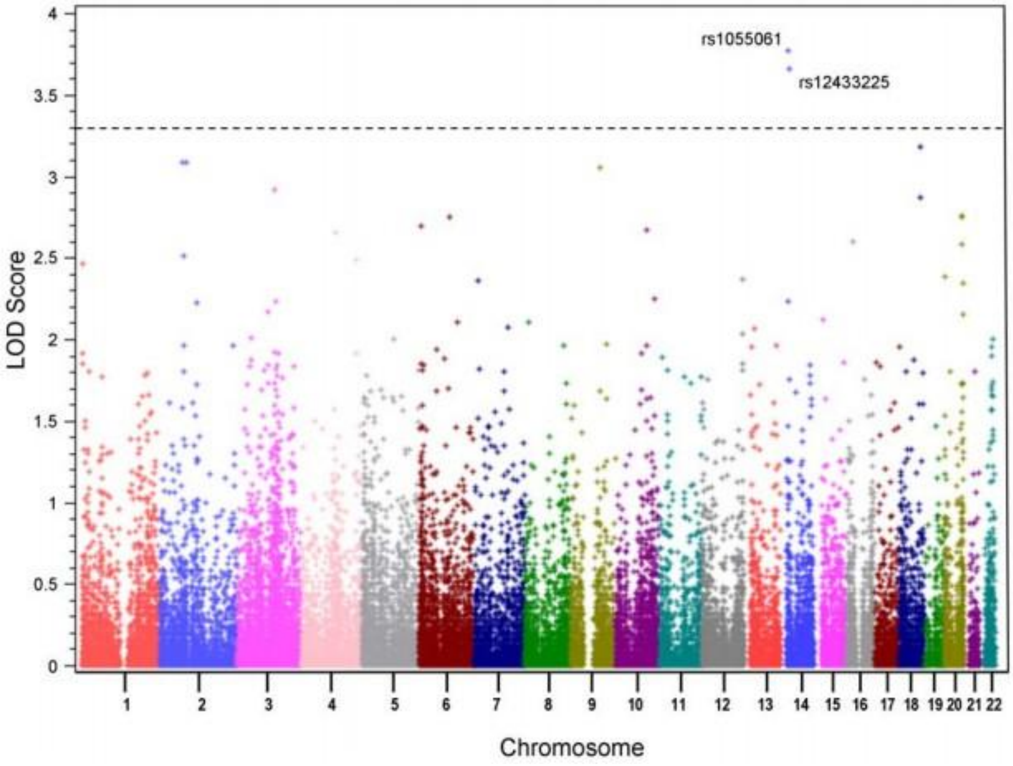

Figure 1-4 Paradigm for linkage analysis results[7]

# 2 Gene expression array

## 2.1 Overview

In the field of molecular biology, gene expression profiling is the measurement of the activity (the expression) of thousands of genes at once, to create a global picture of cellular function. Expression array measures the relative activity of previously identified target genes with cDNA samples.

Popular expression arrays include: NimbleGen Expression Arrays, Affymetrix Expression Arrays, Agilent Expression Arrays, SuperArray Expression Assays and Panomics Gene Expression.

## 2.2 Workflow

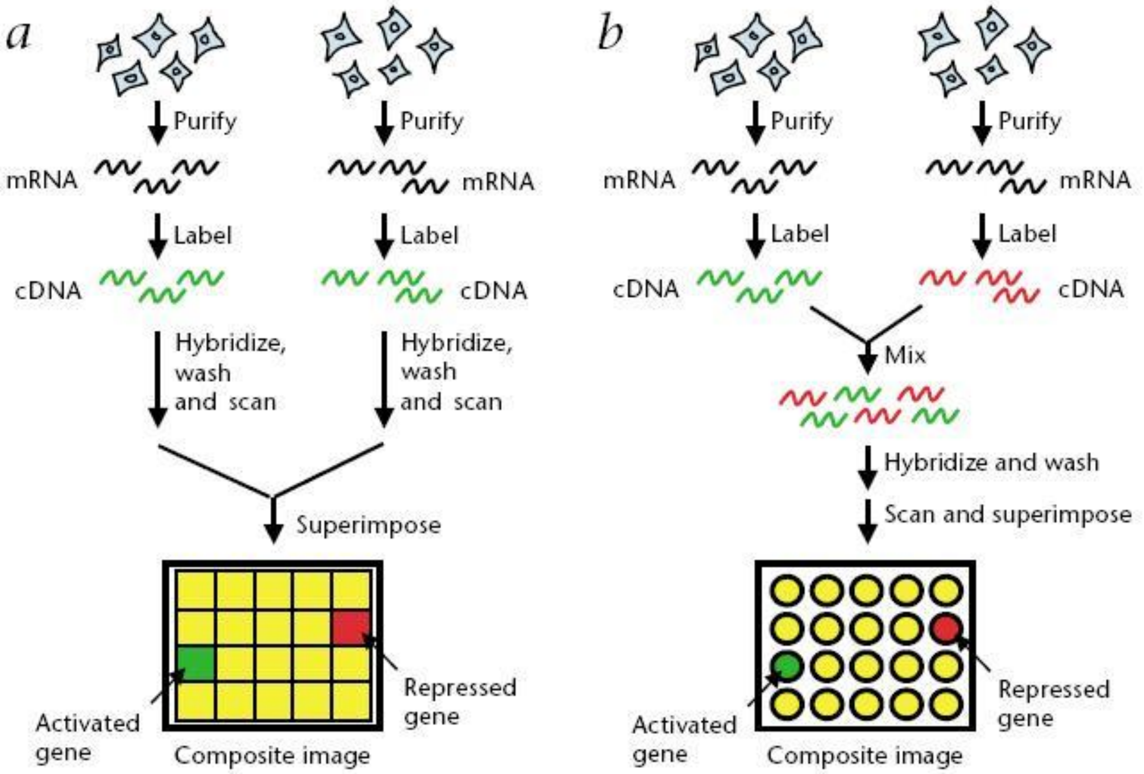

Figure 2- 1 Workflow for gene expression array[8]

## 2.3 Bioinformatics analysis

### 2.3.1 Standard analysis

- Data QC
- Raw data normalization

### 2.3.2 Advanced analysis

- Gene expression differentiation analysis
- Cluster analysis
- Discriminatory analysis
- Gene Ontology (GO, Figure 2-2)
- Pathway analysis: construct interaction network for interested genes
- Enrichment analysis
    - GO enrichment analysis
    - Pathway or protein domain enrichment analysis
- Protein-Protein interaction network analysis

.

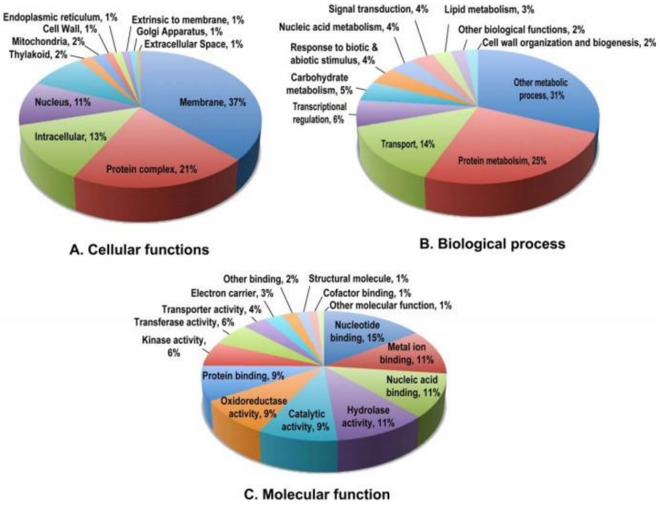

Figure 2- 2 Gene Ontology results[9]

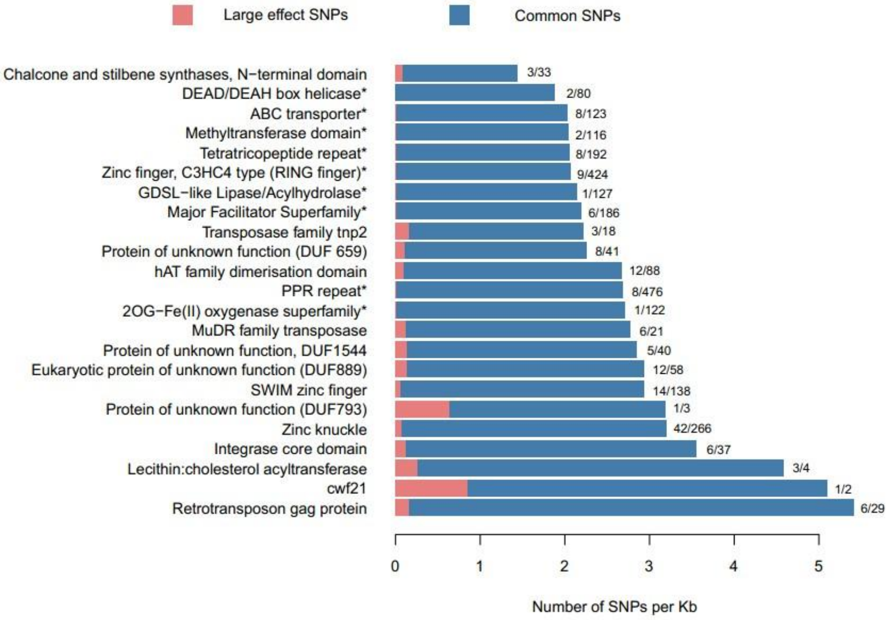

Figure 2- 3 Enrichment analysis results[10]

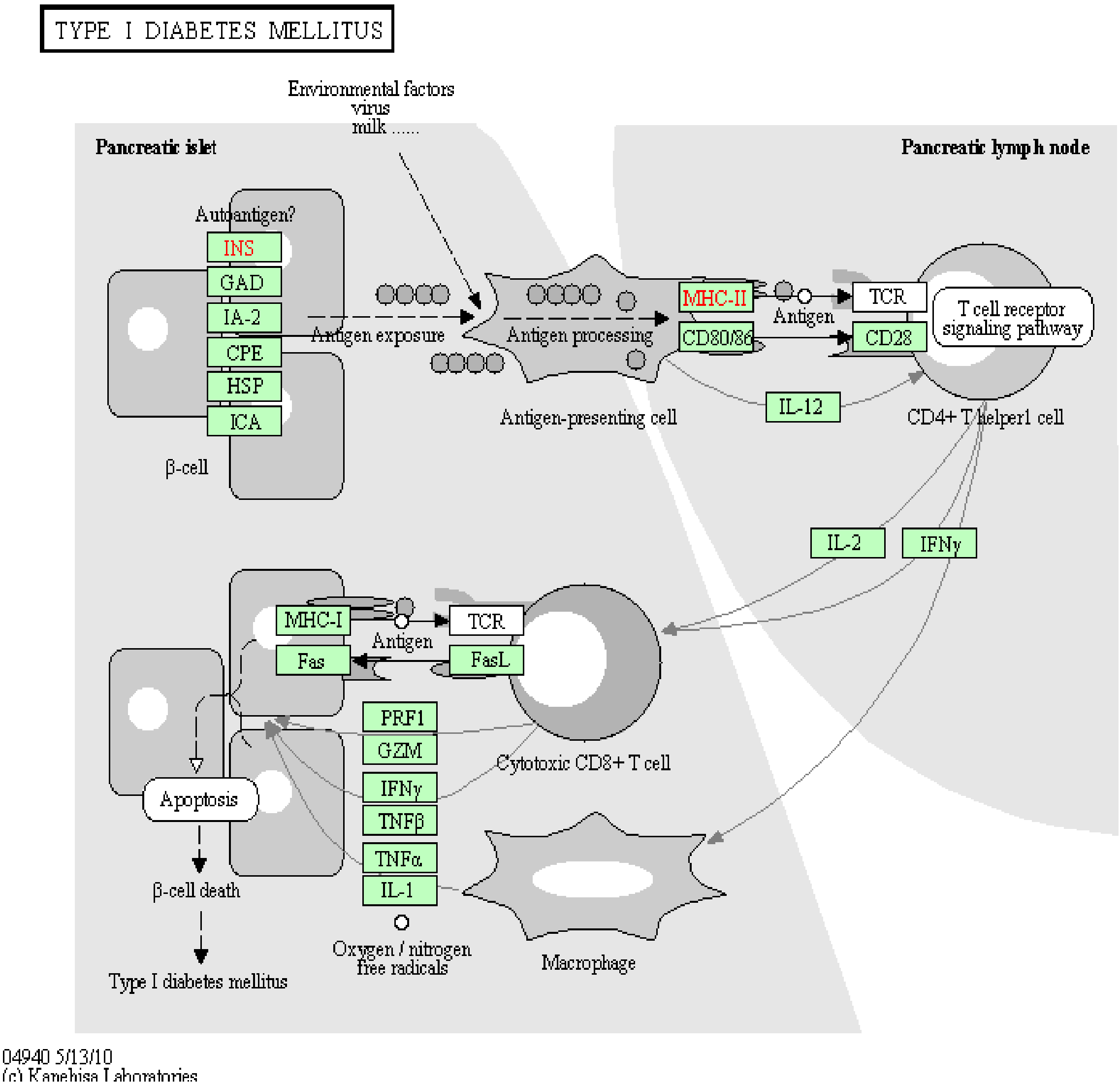

Figure 2- 4 Pathway for I type diabetes [11]

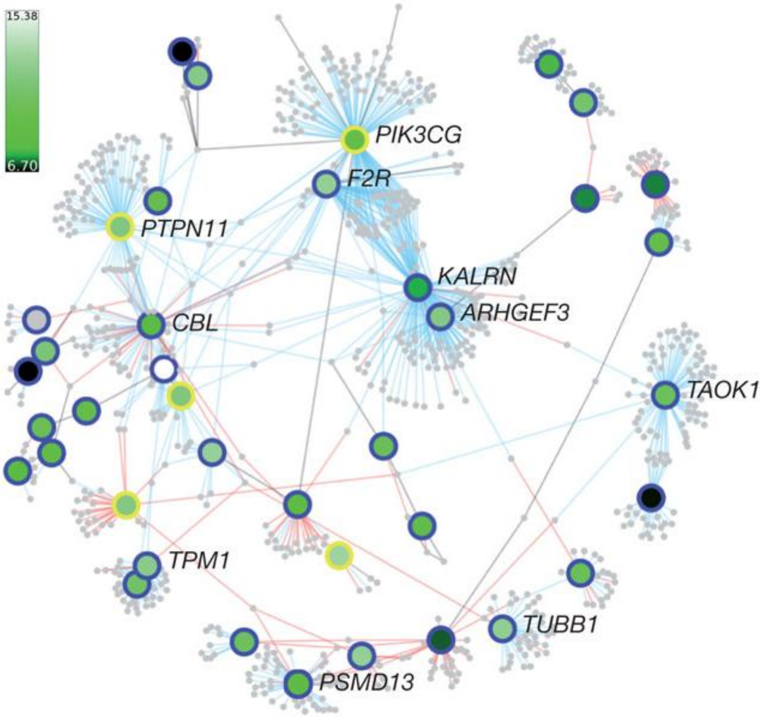

Figure 2-5 Paradigm for protein-protein interaction network [12], different colors correspond to different gene expression levels

# 3 DNA methylation array

## 3.1 Overview

DNA methylation plays a significant role in the epigenetic regulation of chromatin structure, which in the last decade has been recognized to be important in the regulation of gene expression, development and genetic imprinting in vertebrates. Changes in the methylation pattern and level have been shown to contribute to cancer and various developmental diseases.

Popular arrays for DNA methylation include: Roche-NimbleGen DNA Methylation Assay, Illumina Methylation Assay, Agilent DNA Methylation and CpG assay.

## 3.2 Workflow

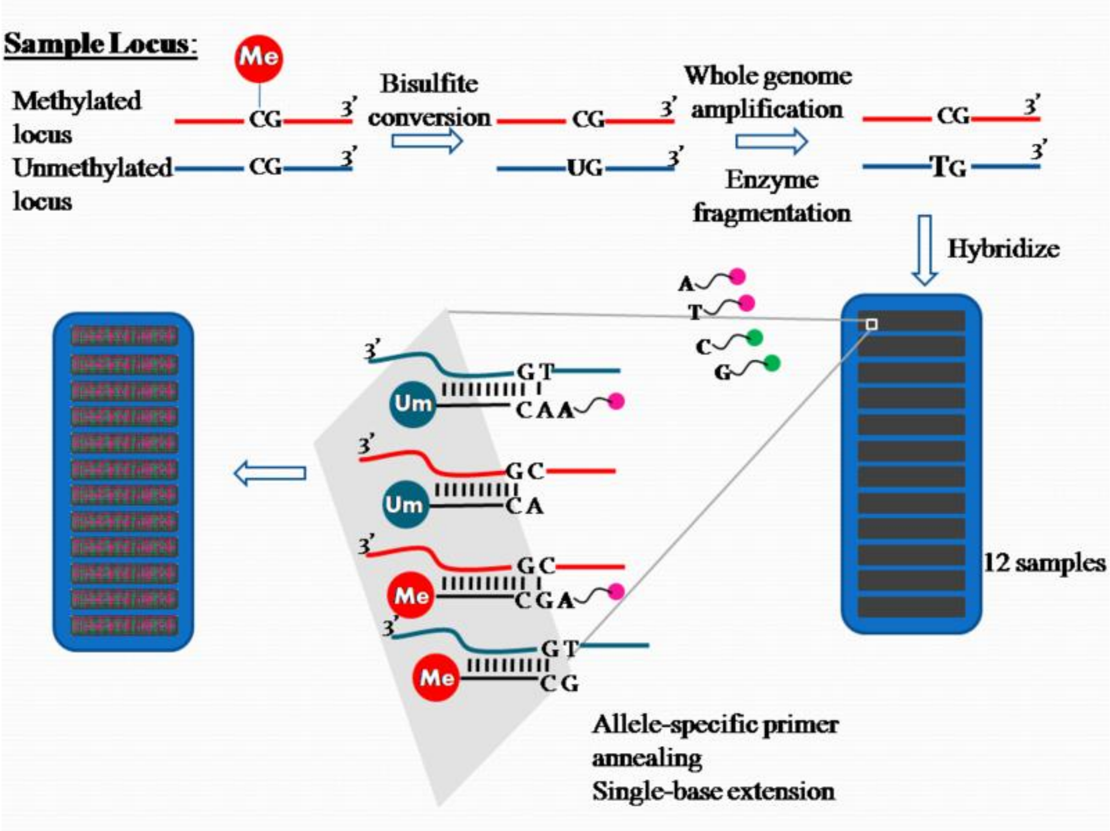

Figure 3-1 Workflow for Illumina Methylation Assay [13]

## 3.3 Bioinformatics analysis

### 3.3.1 Standard analysis

- Data QC
- Raw data normalization

### 3.3.2 Advanced analysis

- Relevancy and recurrence rate analysis
- Relativity between case and control
- Differentiation analysis
- Cluster analysis
- Building model for disease prediction based on methylation pattern
- Pathway analysis
- Gene Ontology
- Association analysis between methylation pattern and regulation of gene expression
- Motif identification for regions adjacent differential methylated sites
- Methylation pattern around TTS region
- Methylation pattern across the whole genome

# 4 array CGH

## 4.1 Overview

Comparative genomic hybridization (CGH) is a molecular-cytogenetic method for the analysis of copy number changes (gains/losses) in the DNA content of a given subject's DNA and often in tumor cells. CGH will detect only unbalanced chromosomal changes. Structural chromosome aberrations such as balanced reciprocal translocations or inversions cannot be detected, as they do not change the copy number. In array CGH, arrays of genomic BAC,P1,cosmid or cDNA clones are used for hybridization instead of metaphase chromosomes in conventional CGH. Fluorescence ratios at arrayed DNA elements provide a locus-by-locus measure of DNA copy-number variation, represents a means of achieving increased mapping resolution technique.

## 4.2 Workflow

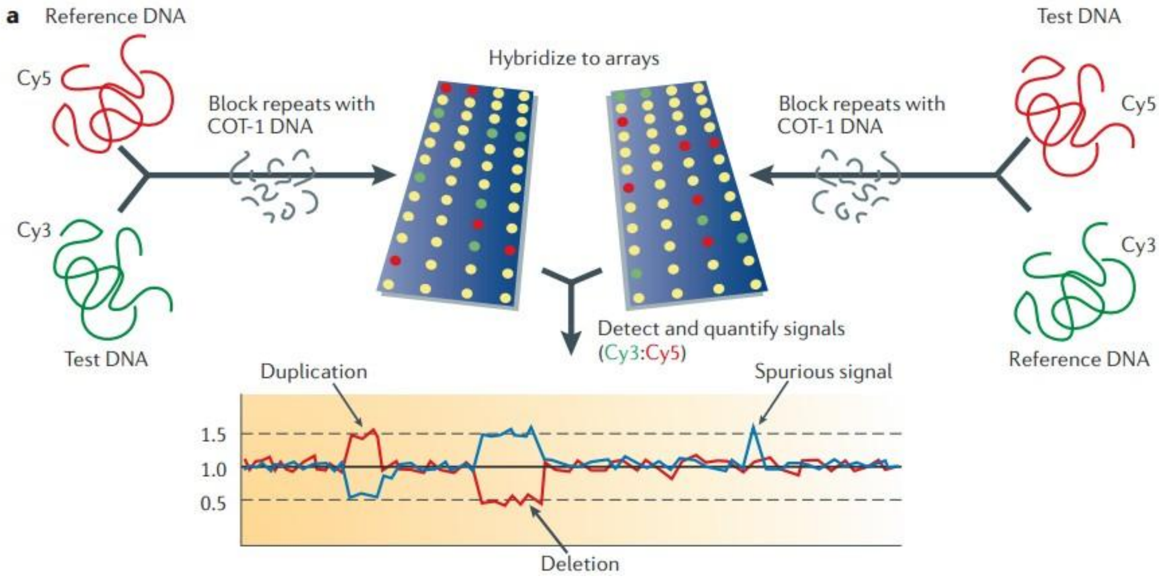

Figure 4- 1 Workflow for array CGH [14]

## 4.3 Bioinformatics analysis

### 4.3.1 Standard analysis

- Data extraction and preprocessing
- Determination of copy number variation (CNV)
- Location of CNV
- Annotation

### 4.3.2 Advanced analysis

- Gene Ontology
- Pathway analysis
- Analysis of shared CNVs for different samples

# 5 ChIP-chip

## 5.1 Overview

Chromatin Immunoprecipitation-chip, (ChIP-chip) is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray technology ("chip"). Like regular ChIP, ChIP-chip is used to investigate interactions between proteins and DNA in vivo. Specifically, it allows the identification of the cistrome, sum of binding sites, for DNA-binding proteins on a genome-wide basis. Whole-genome analysis can be performed to determine the locations of binding sites for almost any protein of interest. The goal of ChIP-chip is to localize protein binding sites that may help identify functional elements in the genome.

## 5.2 Workflow

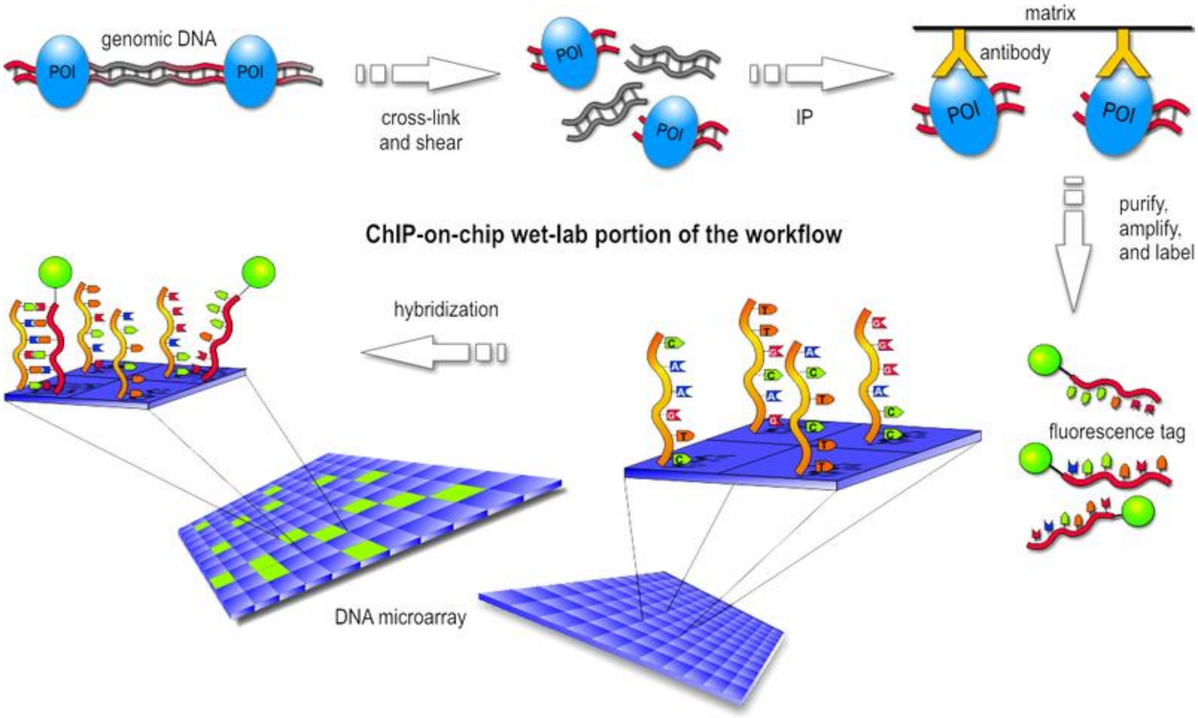

Figure 5- 1 Workflow for ChIP-chip

## 5.3 Bioinformatics analysis

### 5.3.1 Standard analysis

- Data QC
- Raw data normalization
- Peak calling
- Peak annotation

### 5.3.2 Advanced analysis

- Gene Ontology
- Pathway analysis
- Peak differentiation between samples

# 6 miRNA array

## 6.1 Overview

A microRNA (abbreviated miRNA) is a short ribonucleic acid (RNA) molecule found in eukaryotic cells. A microRNA molecule has very few nucleotides (an average of 22) compared with other RNAs. miRNAs are post-transcriptional regulators that bind to complementary sequences on target messenger RNA transcripts (mRNAs), usually resulting in translational repression or target degradation and gene silencing. The human genome may encode over 1000 miRNAs, which may target about 60% of mammalian genes and are abundant in many human cell types. Aberrant expression of miRNAs has been implicated in numerous disease states, and miRNA-based therapies are under investigation.

## 6.2 Workflow

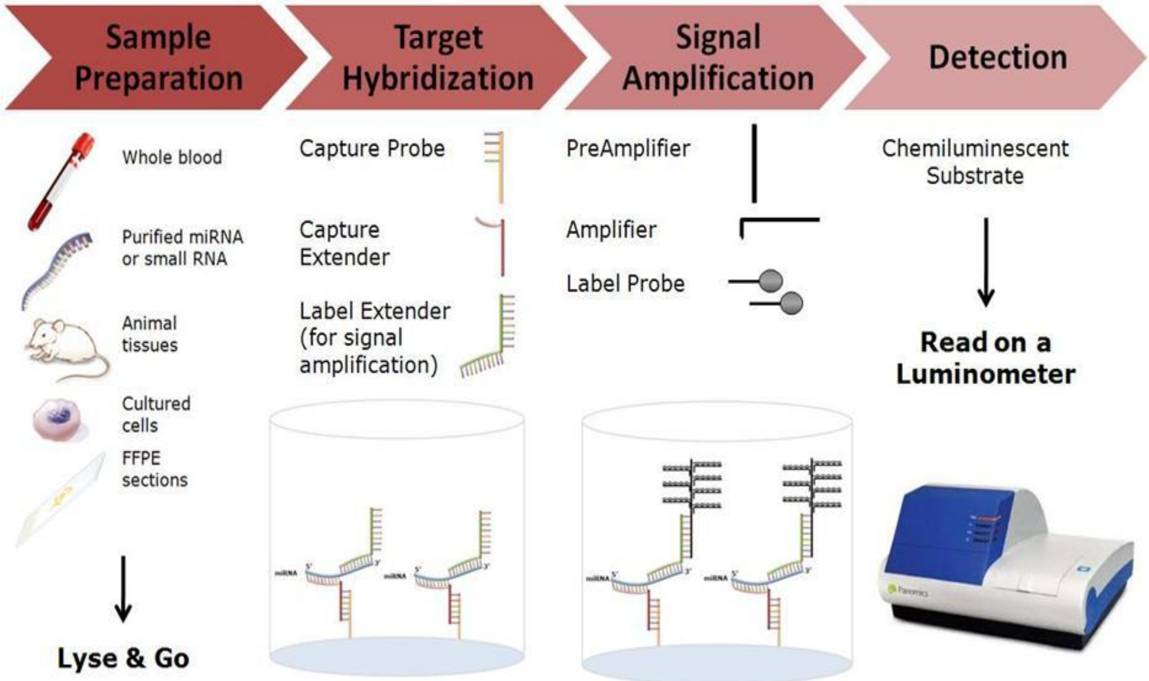

Figure 6- 1 Workflow for QuantiGene 2.0 miRNA array [15]

## 6.3 Bioinformatics analysis

### 6.3.1 Standard analysis

- Data QC
- Raw data normalization

### 6.3.2 Advanced analysis

- Differentiation analysis
- Cluster analysis
- Search for genes regulated by miRNA
- Gene Ontology
- Pathway analysis

# 7 LncRNA array

## 7.1 Overview

Long non-coding RNA (lncRNA) are in general considered (somewhat arbitrarily) as non-protein coding transcripts longer than 200 nucleotides. This limit is due to practical considerations including the separation of RNAs in common experimental protocols. In addition, this limit distinguishes lncRNAs from small regulatory RNAs such as microRNAs, etc. LncRNAs are known to play important roles in gene transcription regulation, post-transcriptional regulation and epigenetic regulation. LncRNA array analysis would provide information for detection of lncRNAs associated with diseases.

## 7.2 Bioinformatics analysis

### 7.2.1 Standard analysis

- Data QC
- Raw data normalization

### 7.2.2 Advanced analysis

- Differentiation analysis
- Cluster analysis
- Search for genes regulated by lncRNA
- Gene Ontology
- Pathway analysis

# 8 Whole genome resequencing

## 8.1 Overview

Whole genome resequencing refers to sequencing different individuals in organism with reference genome sequence. Based on the resequencing data, researchers could detect various variants including SNP, Copy Number Variation (CNV), Insertion/Deletion(Indels) and Structure Variation (SV), providing new opportunities to identify disease related variants.

## 8.2 Workflow

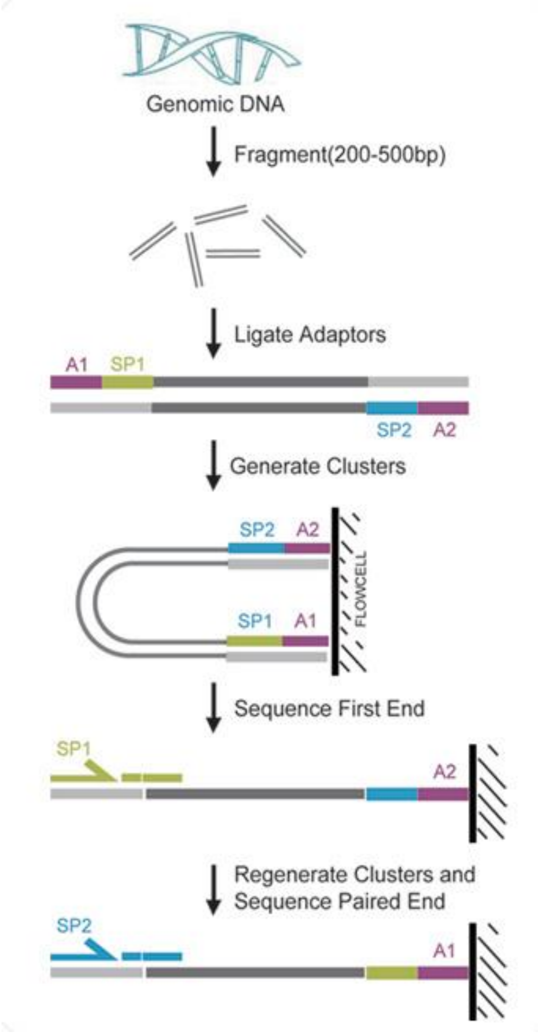

Figure 8- 1 Workflow for Illumina (pair-end) sequencing

## 8.3 Bioinformatics analysis

### 8.3.1 Standard analysis

- Statistics for sequencing data

  After sequencing, total bases, GC-content, distribution for insert size and sequencing quality will be summarized.

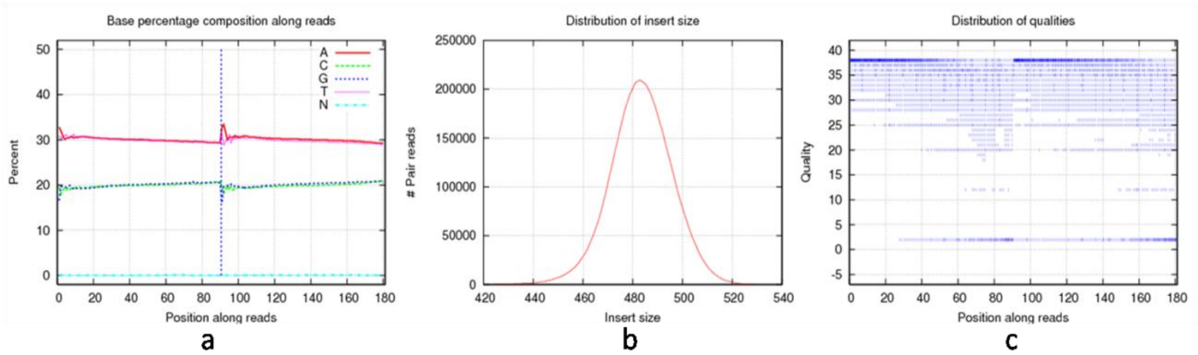

Figure 8- 2 a: distribution for different bases; b: insert size distribution; c: sequencing quality distribution.

- Data filtering: Reads with low quality or high ratio of Ns (more than 50%) will be remove before alignment.
- Alignment

  Mapping reads on to human genome reference (hg18 or hg19) using BWA[16] (Burrows-Wheeler Aligner). PCR duplication will be removed by Samtools[17]. Statistical results for alignment include follows:

  a) Clean reads number;
  b) Unmapped reads number;
  c) Mismatch bases number;
  d) Clean bases number;
  e) Match rate;
  f) Mapped bases number
  g) Mismatch rate;
  h) Coverage;
  i) Depth.
- SNP detection, annotation and summary

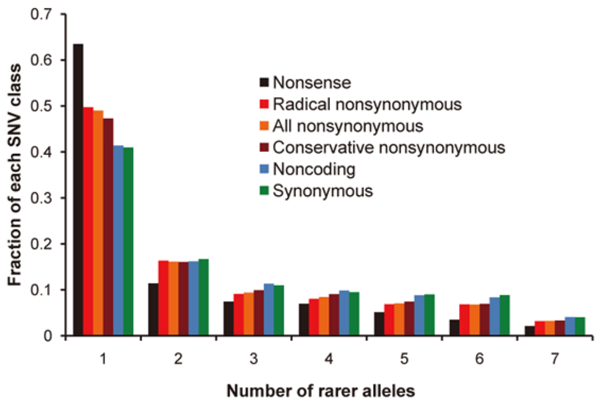

Figure 8- 3 SNP distribution [18]

## 8.3.2 Advanced analysis

### 8.3.2.1 Advanced analysis for individual resequencing

- Indels detection, annotation and summary

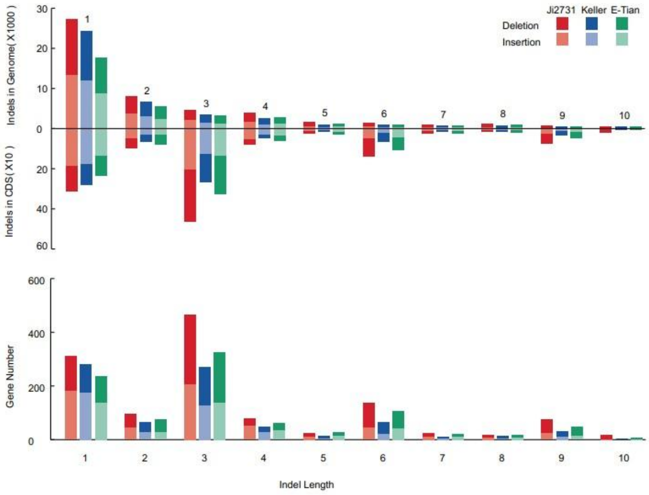

Figure 8- 4 Indels distribution across the whole genome[10]

- SV detection, annotation and summary
- CNV detection, annotation and summary
- Evaluate relationship between variants and diseases or interested phenotype according to available database
- Ancestry analysis

### 8.3.2.2 Advanced analysis for population resequencing

- Indels detection, annotation and summary
- CNV detection, annotation and summary (over 15X per sample)
- Population SNP detection
- Unbiased frequency spectrum estimation (based on population SNPs)

- Population Indels detection
- Haplotype analysis
- Linkage disequilibrium analysis
- Haplotype block prediction
- Demographic analysis (depend on data)
  - Prediction of divergence time
  - Prediction of migration
- Population structure and phylogenetics
  - Population structure
  - Phylogenetic tree
  - Principal components analysis
- Selection analysis (depend on data)
  - Tajima'D
  - Fst

#### 8.3.2.3 Advanced analysis for resequencing of individuals suffered from cancer

- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- Somatic SNP/Indels/SV detection, annotation and summary for normal-tumor pairs
- SNV detection, annotation and summary for normal-tumor pairs
- CNV detection, annotation and summary for normal-tumor pairs
- Prediction of amino acid replacement
- Pathway analysis
- Gene Ontology
- Compare with available database (cosmic, dbSNP etc)
- Viral sequences detection (depend on data)
- Rearrangement detection and annotation (depend on data)
- Selection analysis to detect driver gene (depend on data)
- Mutation target network (depend on data)

#### 8.3.2.4 Advanced analysis for resequencing of individuals suffered from complex disease

- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- NGS-GWAS
  - Association analysis based on low-depth sequencing of large sample size
  - Association analysis based on high-depth sequencing of large or feasible small sample size
- De novo mutation detection based on pedigree samples.

#### 8.3.2.5 Advanced analysis for resequencing of individuals suffered from monogenic disease

- Gender determination
- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- Screening variants in non-coding regions (exon, splicing sites and UTR regions will be remained)
- Compared with available databases (dbSNP, HapMap, The 1000 Genomes project and YH)

# 9 Exome/target region sequencing

## 9.1 Overview

Exome sequencing (also known as targeted exome capture) is an efficient strategy to selectively sequence the coding regions of the genome as a cheaper but still effective alternative to whole genome sequencing. It is estimated that the protein coding regions of the human genome constitute about 85% of the disease-causing mutations. Both exome and target region sequencing use target-enrichment methods, which allow one to selectively capture genomic regions of interest from a DNA sample prior to sequencing.

## 9.2 Workflow

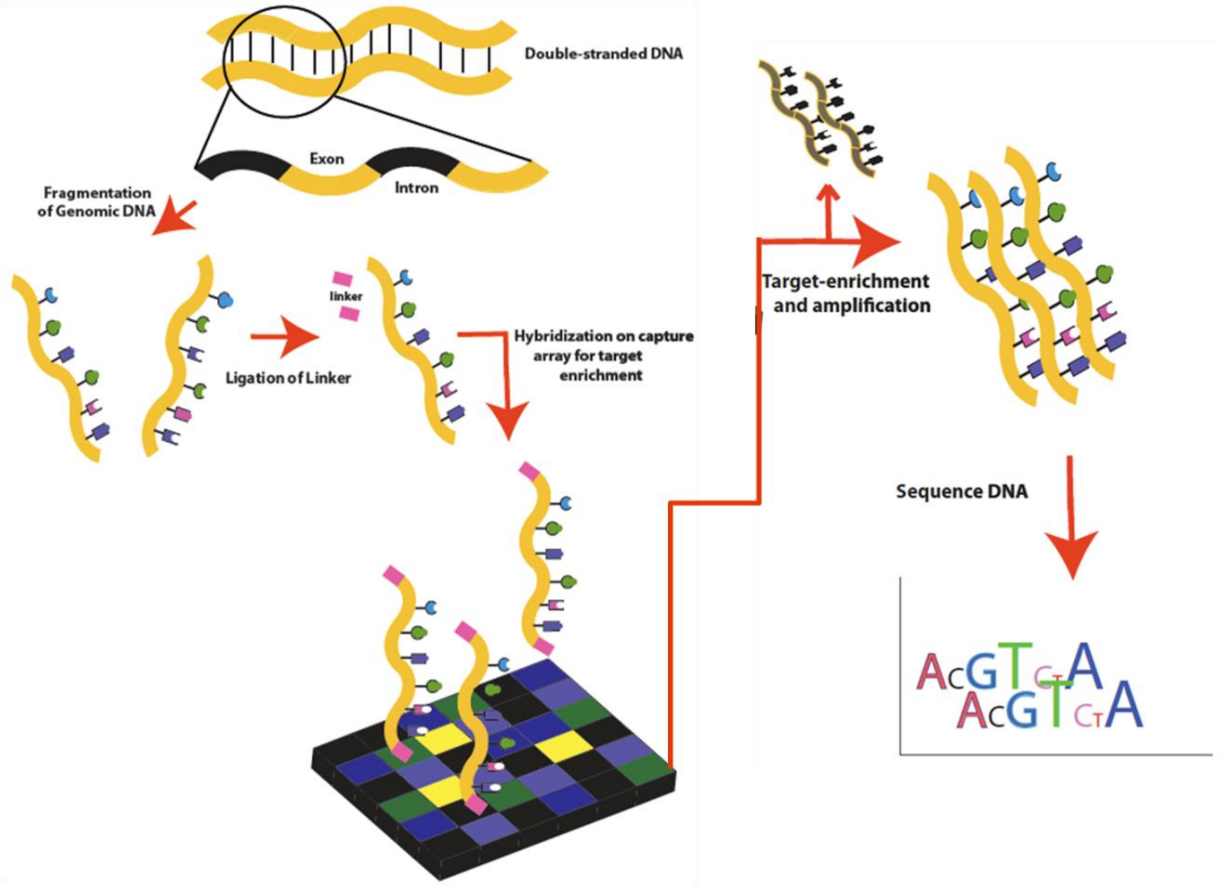
Double-stranded DNA
Exon
Intron
Fragmentation of Genomic DNA
linker
Ligation of Linker
Hybridization on capture array for target enrichment
Target-enrichment and amplification
Sequence DNA
ACGTCTA
ACGTCTA

Figure 9- 1 Workflow for exome sequencing [19]

## 9.3 Bioinformatics analysis

Similar to whole genome sequencing, we provide following analysis procedures for Exome/target region sequencing.

### 9.3.1 Standard analysis

- Data summary
- Data filtering
- Alignment
- SNP detection, annotation and summary

### 9.3.2 Advanced analysis

#### 9.3.2.1 Advanced analysis for individual resequencing

- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- Evaluate relationship between variants and diseases or interested phenotype according to available database
- Ancestry analysis

#### 9.3.2.2 Advanced analysis for population resequencing

- Indels detection, annotation and summary
- CNV detection, annotation and summary (over 15X per sample)
- Population SNP detection
- Unbiased frequency spectrum estimation (based on population SNPs)
- Population Indels detection
- Haplotype analysis
- Linkage disequilibrium analysis
- Haplotype block prediction
- Demographic analysis (depend on data)
- Population structure and phylogenetics
- Selection analysis (depend on data)

#### 9.3.2.3 Advanced analysis for individuals suffered from cancer

- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- Somatic SNP/Indels/SV detection, annotation and summary for normal-tumor pairs
- SNV detection, annotation and summary for normal-tumor pairs
- CNV detection, annotation and summary for normal-tumor pairs
- Prediction of amino acid replacement
- Pathway analysis

- Gene Ontology
- Compare with available database (cosmic, dbSNP etc)
- Viral sequences detection (depend on data)
- Rearrangement detection and annotation (depend on data)
- Selection analysis to detect driver gene (depend on data)
- Mutation target network (depend on data)

#### 9.3.2.4 Advanced analysis for individuals suffered from complex disease

- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- NGS-GWAS
- De novo mutation detection based on pedigree samples.

#### 9.3.2.5 Advanced analysis for individuals suffered from monogenic disease

- Gender determination
- Indels detection, annotation and summary
- SV detection, annotation and summary
- CNV detection, annotation and summary
- Screening variants in non-coding regions (exon, splicing sites and UTR regions will be remained)
- Compared with available databases (dbSNP, HapMap, The 1000 Genomes project and YH)

# 10 ChIP-Seq

## 10.1 Overview

ChIP-sequencing, also known as ChIP-seq, is used to analyze protein interactions with DNA. ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to map global binding sites precisely for any protein of interest.

## 10.2 Workflow

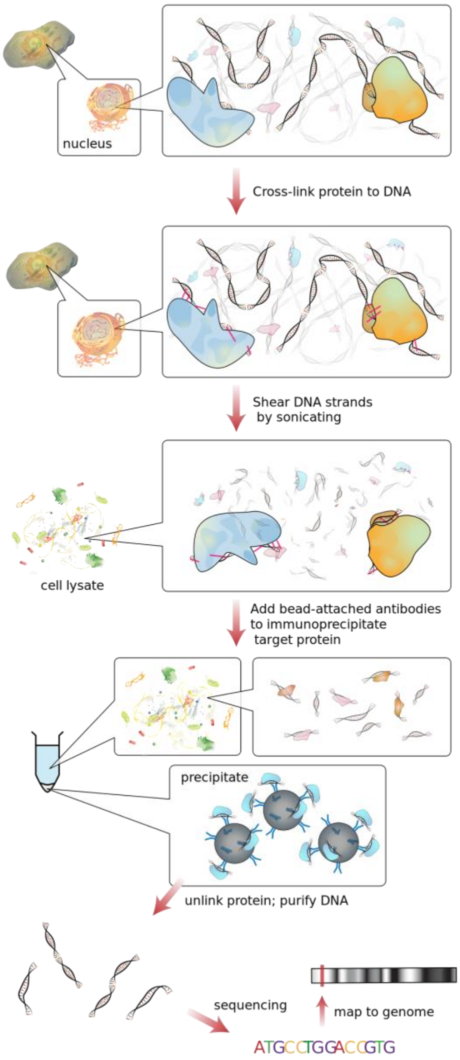
nucleus
Cross-link protein to DNA
Shear DNA strands by sonicating
cell lysate
Add bead-attached antibodies to immunoprecipitate target protein
precipitate
unlink protein; purify DNA
sequencing
map to genome
ATGCCTGGACCGTG

Figure 10- 1 Workflow for ChIP-Seq[20]

## 10.3 Bioinformatics analysis

### 10.3.1 Standard analysis

- Remove adaptor contamination and low quality reads, data summary

- Alignment
- Distribution of reads across the whole genome
  - Distribution of unique mapped reads in repeats region
  - Distribution of unique mapped reads in genes
  - Depth of unique mapped reads across the whole genome
- Peak analysis
  - Peak identification
  - Distribution of Peak length
  - Peak coverage across the whole genome
  - Pattern of Peak in genes
- Manual of UCSC Genome Browser

### 10.3.2 Advanced analysis

- Peak related genes and GO analysis
  - Peak related genes
  - Gene ontology
- Differential analysis among samples
  - Differential analysis based on peak
  - Differential analysis based on peak related genes
- Reads enrichment of genes around TSS region
- Motif analysis
  - Identify new Motif
  - Analysis for known motif (motif information needed）

# 11 MBD-Seq/ MeDIP-Seq

## 11.1 Overview

DNA methylation is an epigenetic modification involved in both normal developmental processes and disease states through the modulation of gene expression and the maintenance of genomic organization. Methylated DNA immunoprecipitation (MeDIP) is used to isolate methylated DNA fragments for input into DNA detection methods, such as DNA sequencing (MeDIP-Seq) . Similar to MeDIP-seq, Methylated DNA Binding Domain-Sequencing (MBD-Seq) combines precipitation of methylated DNA by recombinant methyl-CpG binding domain of MBD2 protein and sequencing of the isolated DNA by a massively parallel sequencer.

## 11.2 Workflow

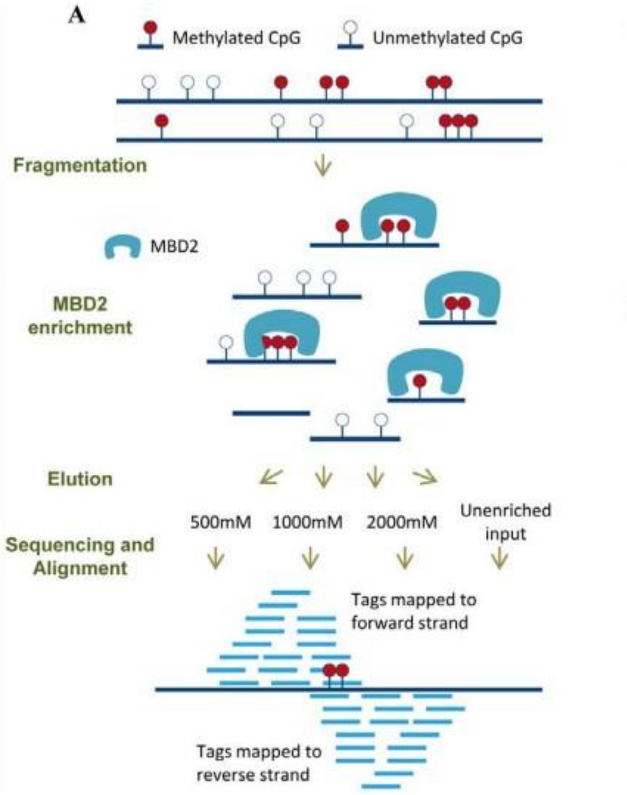

Figure 11- 1 Workflow for MBD-Seq [21]

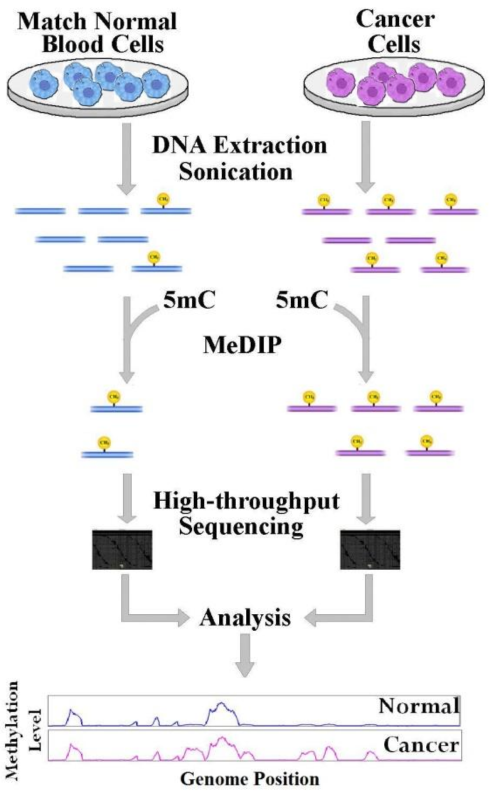

Figure 11- 2 Workflow for MeDIP-seq

## 11.3 Bioinformatics analysis

Bioinformatics analysis for MBD-seq/MeDIP-seq is as follows:

### 11.3.1 Standard analysis

- Remove adaptor contamination and low quality reads, data summary
- Alignment
- Distribution of reads across the whole genome
    - ◆ Distribution of reads on every chromosome
    - ◆ Depth across the whole genome
    - ◆ Depth of CG, CHG and CHH sites
    - ◆ Distribution of reads in regions with different CpG density
    - ◆ Distribution of reads on genes
    - ◆ Mean depth distribution on genes and flanking regions (~2000bp)
- Enrichment analysis
    - ◆ Peak identification
    - ◆ Distribution of Peak length
    - ◆ Peak distribution in regions with different CpG density
    - ◆ Peak related genes
    - ◆ Pattern of Peak in genes

### 11.3.2 Advanced analysis

- Differential analysis across the genome based on peak
  - Peak related genes
  - Gene Ontology
  - Pathway analysis

# 12 RNA-seq

## 12.1 Overview

RNA-seq, also called "Whole Transcriptome Shotgun Sequencing" ("WTSS"), refers to the use of high-throughput sequencing technologies to sequence cDNA in order to get information about a sample's RNA content. The technique has been rapidly adopted in studies of diseases like cancer. RNA-seq can be done with a variety of platforms to test many ideas and hypotheses. For example, using the Illumina Genome Analyzer platform, recent applications include sequencing mammalian transcriptomes.

## 12.2 Workflow

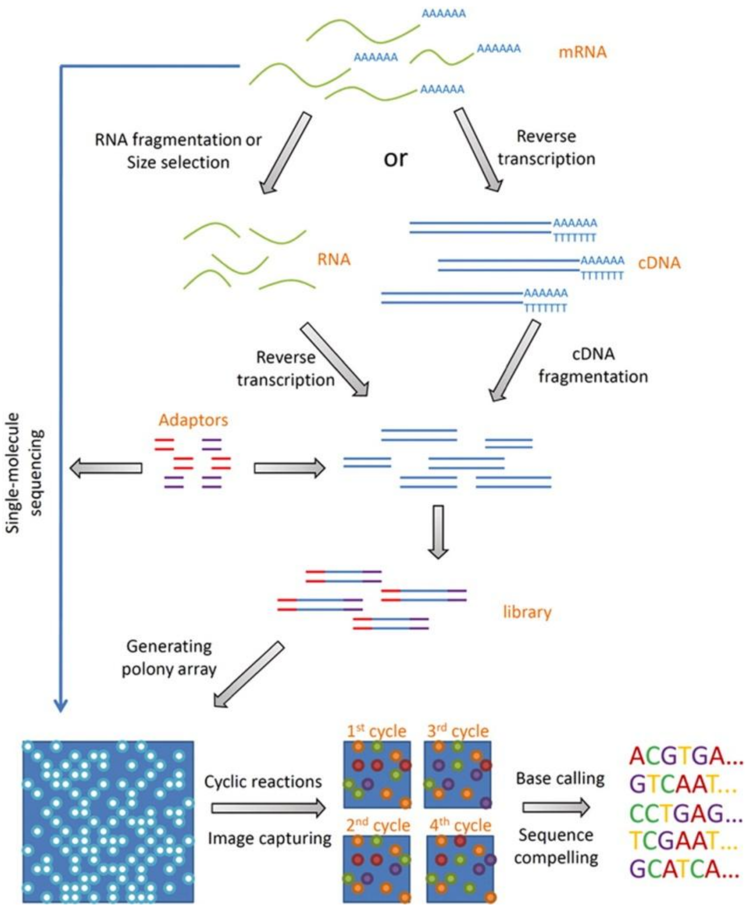

Figure 12- 1 Workflow for RNA-seq[22]

## 12.3 Bioinformatics analysis

### 12.3.1 Standard analysis (with reference genome sequence)

- Remove adaptor contamination and low quality reads, data summary

- Evaluation for RNA-seq ( summary for alignment, randomness assessment, reads distribution across the whole genome)
- Annotation(coverage, depth)
- Differential analysis
- Optimization of gene structure
- Alternative splicing sites identification
- Prediction of new transcript
- SNP analysis

### 12.3.2 Standard analysis (without reference genome sequence)

- Remove adaptor contamination and low quality reads, data summary
- Data summary and evaluation for RNA-seq
- Assembly (Distribution of contig and unigene)
- Unigene annotation
- Unigene GO analysis
- Unigene COG analysis
- Unigene pathway analysis
- CDS prediction
- Differential expression analysis for unigene
- Gene Ontology and pathway analysis for differential expressed unigenes among samples

# 13 Small RNA sequencing

## 13.1 Overview

Small RNA (sRNA) , which includes microRNA (miRNA), short interfering RNA (siRNA), piwi-interacting RNA (piRNA) and siRNA RNA (rasiRNA), are typically only ~18–40 nucleotides in length, however their effect on cellular processes is profound. They have been shown to play critical roles in developmental timing, cell fate, tumor progression, neurogenesis. Input material for sRNA sequencing is often enriched for sRNAs.

## 13.2 Workflow

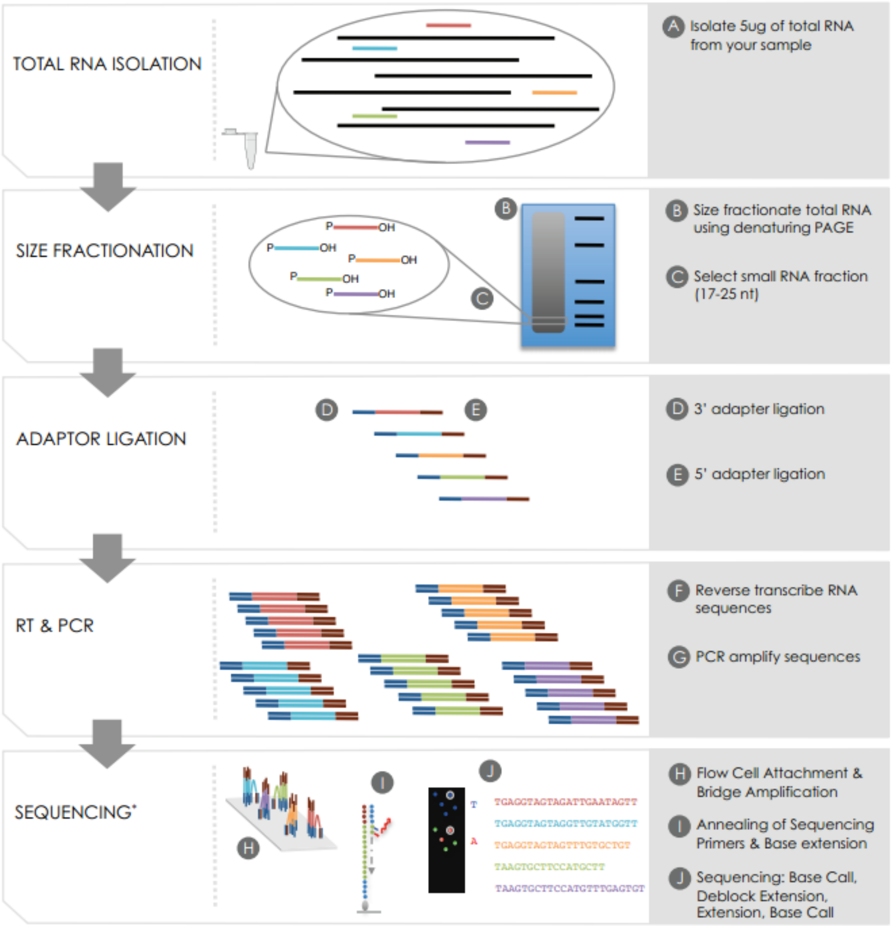
TOTAL RNA ISOLATION
A Isolate 5ug of total RNA from your sample
SIZE FRACTIONATION
P OH
B
C
B Size fractionate total RNA using denaturing PAGE
C Select small RNA fraction (17-25 nt)
ADAPTOR LIGATION
D
E
D 3' adapter ligation
E 5' adapter ligation
RT & PCR
F Reverse transcribe RNA sequences
G PCR amplify sequences
SEQUENCING*
H
I
J
T
A
TGAGGTAGTAGATTGAATAGTT
TGAGGTAGTAGGTTGTATGGTT
TGAGGTAGTAGTTTGTGCTGT
TAAGTGCTTCCATGCTT
TAAGTGCTTCCATGTTTGAGTGT
H Flow Cell Attachment & Bridge Amplification
I Annealing of Sequencing Primers & Base extension
J Sequencing: Base Call, Deblock Extension, Extension, Base Call

Figure 13- 1 Workflow for small RNA sequencing

## 13.3 Bioinformatics analysis

### 13.3.1 Standard analysis

- Remove adaptor contamination and low quality reads, data summary
- Length distribution for 18~30 nt small RNA

- Common and specific sequences for samples
- Distribution of small RNA on preselected reference genome
- Comparison of small RNA with rRNA, tRNA, snRNA and snoRNA
- Comparison of small RNA and repetitive sequence (need preselected reference genome and corresponding annotation information)
- Comparison of small RNA and exon/intron (need preselected reference genome and corresponding annotation information)
- Comparison of small RNA and known miRNA in miRBase (need preselected species, all plants, all animals, or all species can also be an option)
- Annotation of small RNA according to priority
- Predict new miRNA using Mireap and provide their secondary structure plot
- Expression statistics for known miRNAs
- Family analysis for known miRNAs (need the Latin name of reference genome)

### 13.3.2 Advanced analysis

- Differential analysis for miRNAs (more than 2 samples) and cluster analysis (more than 3 samples)
- Target gene prediction for miRNA (need gene coding sequences)
- Gene Ontology and KEGG pathway analysis for miRNA target genes
- Base edit analysis for miRNA (known miRNA)

# Part III References

1. Dunlop, M.G., et al., *Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk.* Nat Genet, 2012.

2. Kavak, E., et al., *Meta-analysis of cancer gene expression signatures reveals new cancer genes, SAGE tags and tumor associated regions of co-regulation.* Nucleic Acids Res, 2010. **38**(20): p. 7008-21.

3. Tuna, M., et al., *Association between acquired uniparental disomy and homozygous mutations and HER2/ER/PR status in breast cancer.* PLoS One, 2010. **5**(11): p. e15094.

4. He, M., et al., *Expression signature developed from a complex series of mouse models accurately predicts human breast cancer survival.* Clin Cancer Res, 2010. **16**(1): p. 249-59.

5. Nibbe, R.K., M. Koyuturk, and M.R. Chance, *An integrative -omics approach to identify functional sub-networks in human colorectal cancer.* PLoS Comput Biol, 2010. **6**(1): p. e1000639.

6. Carrasquillo, M.M., et al., *Genome-wide screen identifies rs646776 near sortilin as a regulator of progranulin levels in human plasma.* Am J Hum Genet, 2010. **87**(6): p. 890-7.

7. McGregor, T.L., et al., *Consanguinity mapping of congenital heart disease in a South Indian population.* PLoS One, 2010. **5**(4): p. e10286.

8. Stears, R.L., T. Martinsky, and M. Schena, *Trends in microarray analysis.* Nat Med, 2003. **9**(1): p. 140-5.

9. Sharma, M.K., et al., *A genome-wide survey of switchgrass genome structure and organization.* PLoS One, 2012. **7**(4): p. e33892.

10. Zheng, L.Y., et al., *Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor).* Genome Biol, 2011. **12**(11): p. R114.

11. http://www.genome.jp/kegg/pathway/hsa/hsa04940.html.

12. Gieger, C., et al., *New gene functions in megakaryopoiesis and platelet formation.* Nature, 2011. **480**(7376): p. 201-8.

13. http://en.wikipedia.org/wiki/Illumina_Methylation_Assay.

14. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.

15. http://www.panomics.com/index.php?id=product_75.

16. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

17. http://samtools.sourceforge.net/.

18. Fujimoto, A., et al., *Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing.* Nat Genet, 2010. **42**(11): p. 931-6.

19. http://en.wikipedia.org/wiki/Exome_sequencing.

20. http://en.wikipedia.org/wiki/ChIP-sequencing.

21. Lan, X., et al., *High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology.* PLoS One, 2011. **6**(7): p. e22226.

22. Wang, L., P. Li, and T.P. Brutnell, *Exploring plant transcriptomes using ultra high-throughput sequencing.* Brief Funct Genomics, 2010. **9**(2): p. 118-28.