Introduction to Large-Scale Biomedical Data Analysis

A grand overview of the course

- Introductory course created for the BIG (Bioinformatics, Imaging and Genetics) concentration.
- Purpose of the course: introduce modern high-dimensional biomedical **data** from:
  - Bioinformatics and computational biology.
  - Biomedical imaging.
  - Statistical genetic.

Contents of the course

- Focus on:
  - Scientific background: questions and motivations.
  - Technologies.
  - Data and their characteristics.
  - Brief overview of some statistical methods, opportunities and challenges for statisticians.

  There will be a lot of materials and new terminologies!
- **Not** covered in this course: detailed statistical theories and methods for data analyses.
- This is a knowledge-centric class, big picture and concepts are more important.

Format of the course

- Co-taught by multiple instructors.
- Students are evaluated by 3 reading assignments: you need to write reading reports!

Lecture 1:
Introduction to next generation sequencing

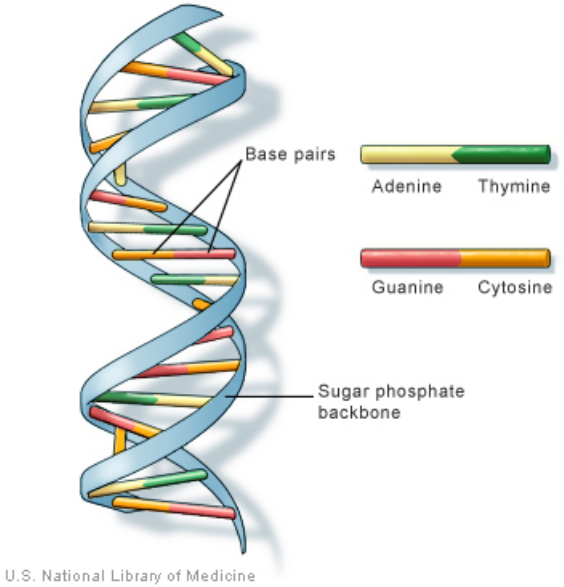
Outline

- Biological Backgrounds: DNA and DNA sequencing.
- Next generation sequencing (NGS) technologies.
- Data from NGS.
- Typical NGS data analysis workflow.
- An application of NGS: RNA-seq.

Background: DNA and sequencing

DNA (DeoxyriboNucleic Acid)

- A molecule contains the genetic instruction of all known living organisms and some viruses.
- Resides in the cell nucleus, where DNA is organized into long structures called **chromosomes**.
- Most DNA molecule consists of two long polymers (**strands**), where two strands entwine in the shape of a double helix.
- Each strand is a chain of simple units (**bases**) called **nucleotides**: A, C, G, T.
- The bases from two strands are complementary by **base pairing**: **A-T, C-G.**



DNA sequence

- The order of occurrence of the bases in a DNA molecule is called the **sequence** of the DNA. The DNA sequence is usually store in a big text file:

```
ACAGGTTTGCTGGTGACCAGTTCTTCATGAGGGACCATCTATCACAACAG
AGAAAGCACTTGGATCCACCAGGGGCTGCCAGGGGAAGCAGCATGGGAGC
CTGAACCATGAAGCAGGAAGCACCTGTCTGTAGGGGGAAGTGATGGAAGG
ACATGGGCACAGAAGGGTGTAGGTTTTGTGTCTGGAGGACACTGGGAGTG
GCTCCTGGCATTGAAACAGGTGTGTAGAAGGATGTGGTGGGACCTACAGA
CAGACTGGAATCTAAGGGACACTTGAATCCCAGTGTGACCATGGTCTTTA
AGGACAGGTTGGggccaggcacagtggctcatgcctgtaatcccagcact
```

- Some interesting facts:
  - Total length of the human DNA is **3 billion bases**.
  - Difference in DNA sequence between two individuals is less than 1%.
  - Human and chimpanzee have 96% of the sequences identical. Human and mouse: 70%.

Genome size (total length of DNA)

Organism	Genome size (bp)	# genes
E. coli	4.6M	4,300
S. cerevisiae (yeast)	12.5M	5,800
C. elegans (worm)	100M	20,000
A. thaliana (plant)	115M	28,000
D. melanogaster (fly)	123M	13,000
M. musculus (mouse)	3G	23,800
H. sapiens (human)	3.3G	25,000

DNA sequencing

- Technologies to determine the nucleotide bases of a DNA molecule.
- Motivation: decipher the genetic codes hidden in DNA sequences for different biological processes.
- **Genome projects**: determine DNA sequences for different species, e.g., human genome project.
- **Genomic research** (in a nutshell): study the functions of DNA sequences and related components.

Sequencing technologies

- Traditional technology: **Sanger sequencing**.
  - Slow (low throughput) and expensive: it took Human Genome Project (HGP) 13 years and $3 billion to sequence the entire human genome.
  - Relatively accurate.
- New technology: different types of **high-throughput sequencing**.

Next generation sequencing (NGS)

- Aka: high-throughput sequencing, second-generation sequencing.
- Able to sequence large amount of short sequence segments in a short period:
  - High-throughput: hundreds of millions sequences in a run.
  - Cheap: sequence entire human genome costs a few thousand dollars now.
  - Higher error rate compared to Sanger sequencing.

NGS technologies

- Complicated and involves a lot of biochemical reactions.
  - Sequencing by synthesis.
  - Sequencing by ligation.
  - Pyrosequencing.
- In a nutshell:
  - Cut the long DNA into smaller segments (several hundreds to several thousand bases).
  - Sequence each segment: start from one end and sequence along the chain, base by base.
  - The process stops after a while because the noise level is too high.
  - Results from sequencing are many sequence pieces. The lengths vary, usually a few thousands from Sanger, and several hundreds from NGS.
  - The sequence pieces are called “reads” for NGS data.

Single-end vs. paired-end sequencing

- Sequence one end or both ends of the DNA segments.
- Paired-end sequencing: reads are “paired”, separated by certain length (the length of the DNA segments).
- Paired-end data can be used as single-end, but contain extra information for reads pairing.
- Useful in some cases, for example, detecting structural variations in the genome.

Available NGS platforms

- Major player:
  - Illumina: HiSeq, Genome Analyzer (GA)
  - LifeTech: SOLiD, IonTorrent
  - Roche 454
- Others:
  - Complete Genomics
  - Pacific Bioscience
  - Helicose

Introduction to NGS Applications

Applications of NGS

- NGS has a wide range of applications.
  - DNA-seq: sequence genomic DNA.
  - RNA-seq: sequence RNA products.
  - ChIP-seq: detect protein-DNA interaction sites.
  - Bisulfite sequencing (BS-seq): measure DNA methylation strengths.
  - A lot of others.

Technology	Brief description
ChIP-seq	Locate protein-DNA interaction or histone modification sites.
CLIP-seq	Map protein-RNA binding sites
RNA-seq	Quantify expression
SAGE-seq	Quantify expression
RIP-seq	capture TF-bound transcripts
GRO-seq	evaluate promoter-proximal pausing
BS-seq	Profile DNA methylation patterns
MeDIP-seq	Profile DNA methylation patterns
TAB-seq	Profile DNA hydroxyl-methylation patterns
MIRA-seq	Profile DNA methylation patterns
ChiRP-seq	Map lncRNA occupancy
DNase-seq	Identify regulatory regions
FAIRE-seq	Identify regulatory regions
FRT-seq	Quantify expression
Repli-seq	Assess DNA replication timing
MNase-seq	Identify nucleosome position
Hi-C	Infer 3D genome organization
ChIA-PET	Detect long distance chromosome interactions
4C-seq	Detect long distance chromosome interaction
Sono-seq	Map open-chromatin sites
NET-seq	determine *in vivo* position of all active RNAP complexes.
NA-seq	Map Nuclease-Accessible Sites

DNA-seq

- Sequence the untreated genomic DNA.
  - Obtain DNA from cells, cut into small pieces then sequence the segments.
- Goals:
  - **Genome re-sequencing**: compare to the **reference genome** and look for genetic variants:
    - Single nucleotide polymorphisms (SNPs)
    - Insertions/deletions (indels),
    - Copy number variations (CNVs)
    - Other structural variations (gene fusion, etc.).
  - ***De novo assembly*** of a new unknown genome.

Variations of DNA-seq

- Targeted sequencing, e.g., exome sequencing.
  - Sequence the genomic DNA at targeted genomic regions.
  - Cheaper than whole genome DNA-seq, so that money can be spent to get bigger sample size (more individuals).
  - The targeted genomic regions need to be "captured" first using technologies like microarrays.
- Metagenomic sequencing.
  - Sequence the DNA of a mixture of species, mostly microbes, in order to understand the microbial environments.
  - The goal is to determine number of species, their genome and proportions in the population.
  - De novo assembly is required. But the number and proportions of species are unknown, so it poses challenge to assembly.

ChIP-seq

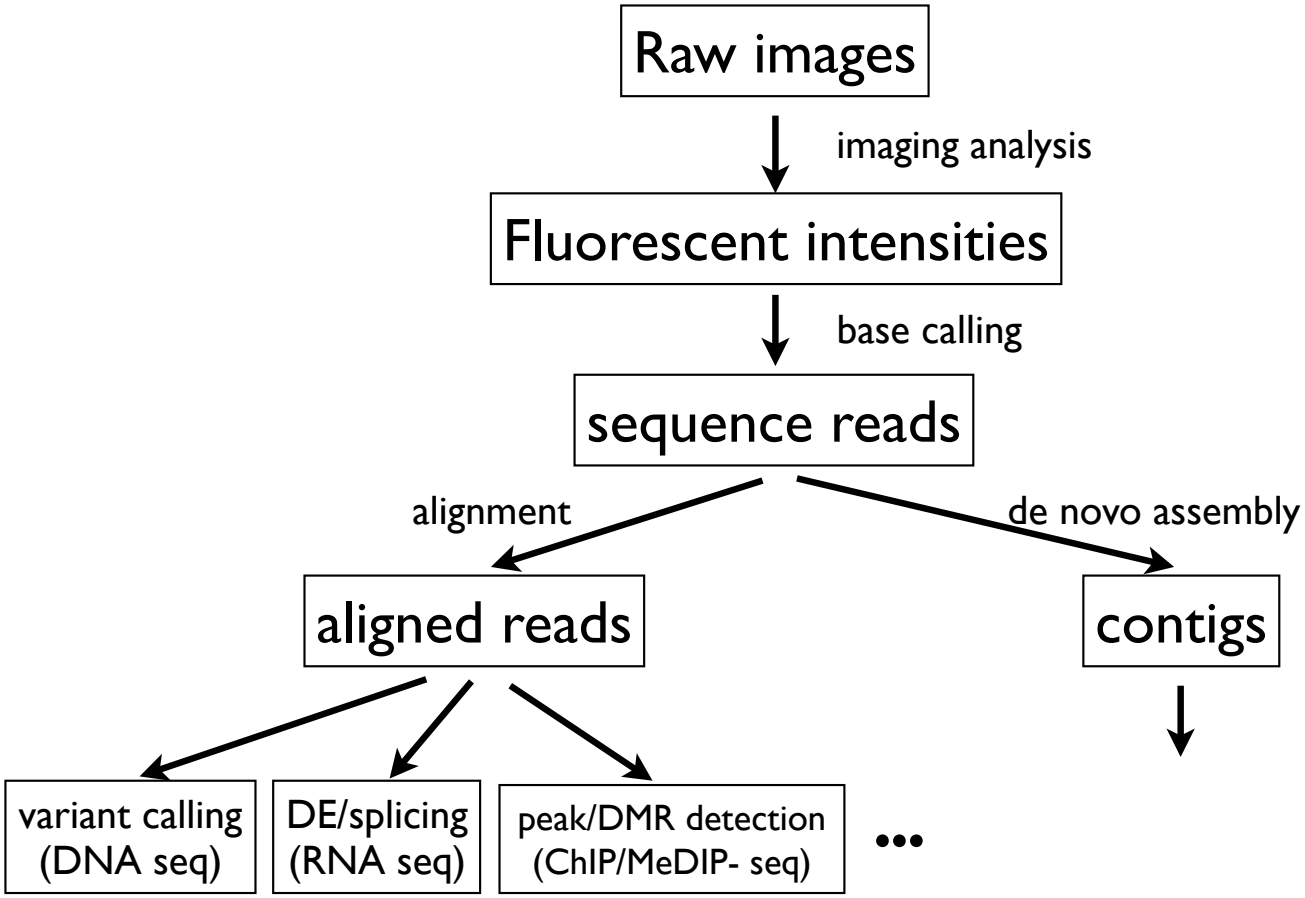
- Chromatin-Immunoprecipitation (ChIP) followed by sequencing (seq): sequencing version of ChIP-chip.
- A type of "captured" sequencing. ChIP step is to capture genomic regions of interest.
- Used to detect locations of certain "events" on the genome: genome-wide location analysis (GWLA). Example of events:
  - Transcription factor binding.
  - DNA methylations and histone modifications.

RNA-seq

- Sequence the “transcriptome”: the set of RNA molecules.
- Goals:
  - Catalog RNA products.
  - Determine transcriptional structures: alternative splicing, gene fusion, etc.
  - Quantify gene expression: the sequencing version of gene expression microarray.

NGS data and analyses workflow

Workflow of NGS data analysis

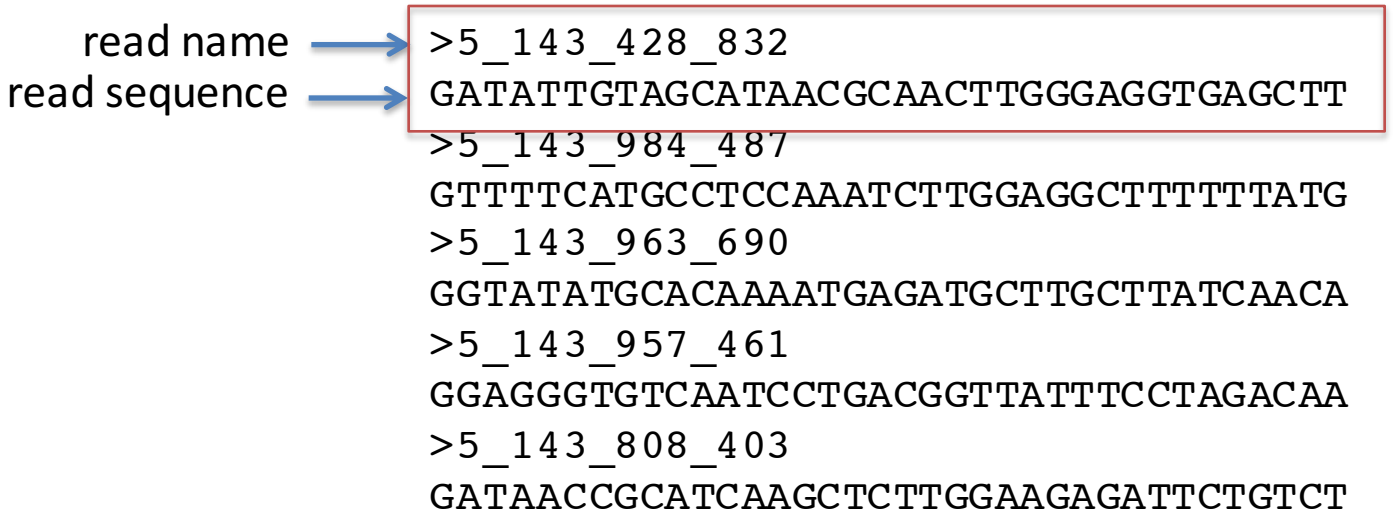


Base calling

- The “raw” data from sequencer are many images.
- To obtain the sequence reads, need a “base calling” step:
  - First extract data (fluorescent intensities) from images
  - Next convert intensities (continuous) to a base calls (categorical).
  - Both steps involve many statistical methods to extract signals from noisy data.
- Sequencing data analyses start from the sequence reads (results of base calling) for most people.

Raw sequence reads data from NGS after base calling

- Large text file (millions of lines) with simple format.
  - Most frequently used: fasta/fa format for storing the sequences, or fastq format storing both the sequence and corresponding quality scores.
- fa format:



fastq format

```
@HWI-EAS165:1:1:50:908:1
CTGCGGTCTCTAAAGTGCCATCTCATTGTGCTTTGTATCAGTCAGTGCTGGA
+
BCCBCB8ABBBBBBBB:BC=8@BBA:@BB@BBBCBB<9BBAC;A<C?BAAB<#
@HWI-EAS165:1:1:50:0:1
NCAACCCCCACAGTAATATGTAAAACAAAAACTAAAACCAGGAGCTGAAGGG
+
#BABABBBBBB@08<@?A@7:A@CCBCCCCBBBCCBB=?BBBB@7@B=A>:2
@HWI-EAS165:1:1:50:708:1
GGTCAGCATGTCTTCTGTTAAGTGCTTGCACAAGCTAGCCTCTGCCTATGGG
+
BB@A;B>@A@@=BB=BB?A>@@>B?ABBA=A?@@>@@A:=?>?A@=B8@@AB
@HWI-EAS165:1:1:50:1494:1
CTGGTGTCACACAAGCAGGTCTCCTGTGTTGACTTCACCAGACACTGTCATT
+
BCBB@AB@1ABBBBBBAAB?BBBBAB<A?AA>BB@?1ABBA@BBBA@;B>>:
```

read name

read sequence

separator

quality scores

Sequence alignment and assembly

- Sequence a known genome --- Alignment
  - Use the known genome (called “reference genome”) as a blue print.
  - Determine where each read is located in the reference genome.
- Sequence a whole new genome --- Assembly
  - New genome: a species with unknown genome, or the genome is believed to be very different from reference (e.g., cancer).
  - Basically the short reads are “stitched” together to form long sequences called “contigs”.
  - Overlaps among sequence reads are required, so it needs a lot of reads (deep coverage).
  - More computationally intensive.

Alignment

- For most people, alignment is the first step for sequencing data analysis, since the machine output raw sequence read file (fastq).
- Need: sequence reads file and a reference genome.
- It is basically a string search problem: where is the short (50-letter) string located within the reference string of 3 billion letters.
- Brute-force searching is okay for a single read, but computationally infeasible to alignment millions of reads.
- Clever algorithms are needed to preprocess the reference genome (indexing), which is beyond the scope of this class.

Popular alignment software

- Bowtie: fast, but less accurate.
- BWA (Burrows-Wheeler Aligner): same algorithm as bowtie, but allow gaps in alignments.
  - about 5-10 times slower than bowtie, but provide better results especially for paired end data.
- Maq (Mapping and Assembly with Qualities): with SNP calling capabilities.
- ELAND: Illumina's commercial software.
- A lot of others. See http://en.wikipedia.org/wiki/List_of_sequence_alignment_software for more details.
- Our suggestion: use bowtie for single-end, and bwa for paired-end.

Once you have the reads aligned

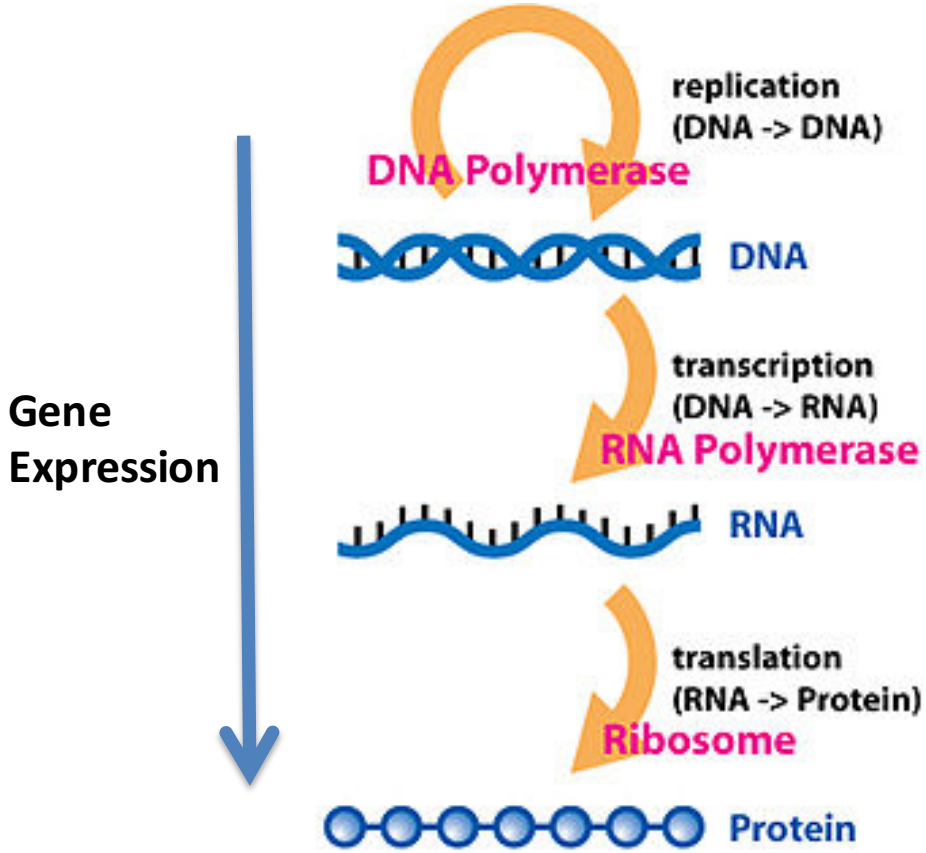
- Downstream analyses depend on purpose.
- Often one wants to manipulating and visualizing the alignment results. There are several useful tools:
  - file manipulating (format conversion, counting, etc.): samtools/Rsamtools, BEDTools, bamtools, IGV tools.
  - Visualizing: samtools (text version), IGV (Java GUI).

RNA-seq data Analysis

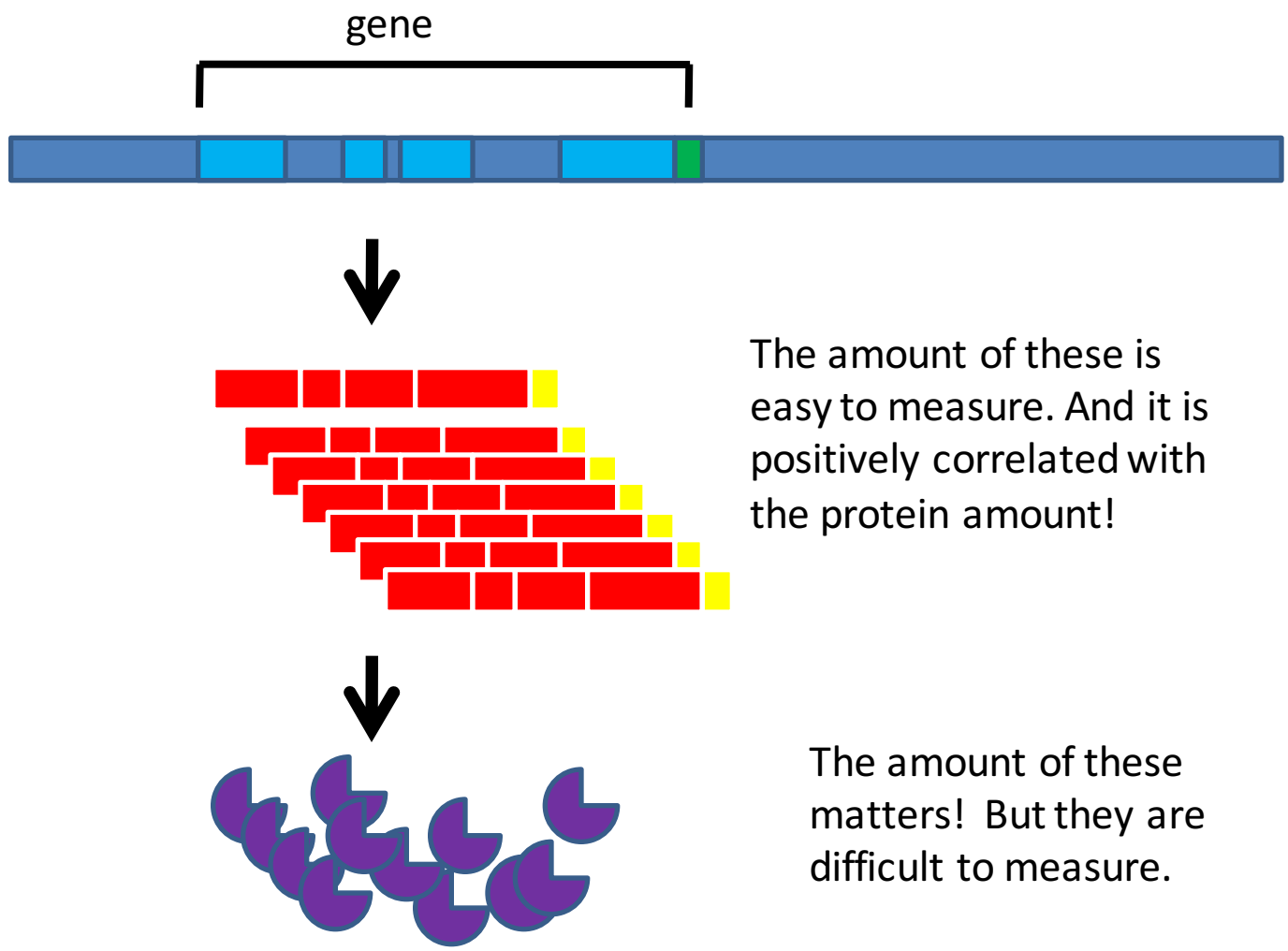
A little biological background

Central dogma of molecular biology:

"The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that information cannot be transferred back from protein to either protein or nucleic acid."



Gene expression levels are measured through their mRNA abundance.



High-throughput technologies to measure mRNA abundance

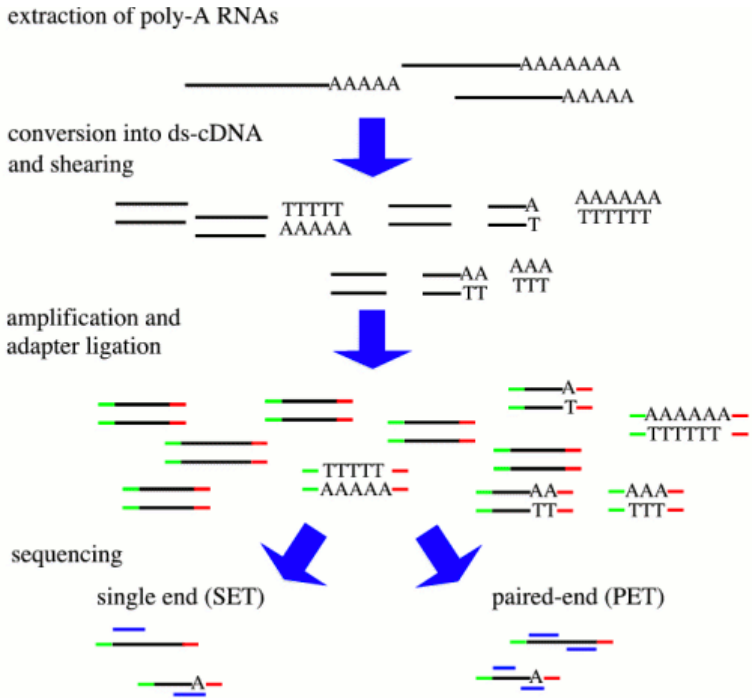
- Traditionally: gene expression microarray.
- Now: RNA-seq.
  - Data have greater dynamic range and higher signal-to-noise ratios.
  - Contain more information: structural change of genes.
- Results show good correlation between microarray and RNA-seq data.

RNA-seq

- Sequence the set of RNA molecules.
- Scientific goals:
  - Catalogue RNA products.
  - Determine transcriptional structures: alternative splicing, gene fusion, etc.
  - Quantify gene expression: the sequencing version of gene expression microarray.

Experimental procedures

- Procedures:
  - RNA -> cDNA.
  - Shearing and amplification of cDNA.
  - Sequence the cDNA segments.
- Compare to expression microarray:
  - Clearer signal without hybridization.
  - Better dynamic ranges for expressions.
  - Provide richer information.



RNA-seq data analyses

- **Gene expression analysis**: compare expressions of genes under different biological conditions.
- Structural analysis of transcriptome: discover and compare the structural variations of mRNA:
  - Alternative splicing.
  - Gene fusion.
- Small RNA analyses.

RNA-seq gene expression analysis

- Goal: compare gene expressions between different samples.
- Steps:
  1. Summarization: get a number for each gene to represent its expression level.
  2. Normalization: remove technical artifacts so that data from different samples are comparable.
  3. **Differential expression detection**: gene by gene statistical test, usually based on negative binomial model.
- Popular software:
  - R/Bioconductor packages: DESeq, edgeR, DSS.
  - Stand alone: cufflink and cuffdiff.

Differential expression

- Biological motivation is the same as in gene expression microarray: find differentially expressed (DE) genes.
- Microarray methods are not directly applicable: continuous vs. count data, but ideas can be borrowed.
- Usually needs multiple replicates per sample, so that means and variances can be evaluated.
- Test is carried gene by gene.
- Two major tasks of developing a test:
  - Within group variance estimation.
  - Test procedure and inference.

RNA-seq data for DE analysis

- A matrix of integers, rows are genes, columns are samples:

	Cancer	Cancer	Cancer	Cancer	Normal	Normal	Normal	Normal
Gene 1	20	25	16	7	4	1	21	11
Gene 2	493	628	133	450	28	97	462	225
Gene 3	128	146	121	25	30	26	106	80
Gene 4	1	0	0	0	1	0	0	0
Gene 5	58	105	27	40	19	28	88	54
Gene 6	0	0	1	0	0	0	0	0
...								

Data model for replicated samples

- For a sample with M replicates, the counts for gene *i* replicate *j,* denoted by Y_{ij}, is often modeled by following hierarchical model:

$$Y_{ij} \mid \lambda_i \sim Poisson(\lambda_i), \lambda_i \sim Gamma(\alpha, \beta)$$

- Marginally, the Gamma-Poisson compound distribution is **Negative binomial**. So the counts for a gene from multiple replicates is modeled as Negative binomial:

$$Y_{ij} \sim NB(\alpha, \beta)$$

A little more about the NB distribution

- NB is over-dispersed Poisson:
  - Poisson: $var = \mu$
  - NB: $var = \mu + \mu^2\phi$
- Dispersion parameter ϕ approximates the squared coefficient of variation: $\phi = \frac{var-\mu}{\mu^2} \approx \frac{var}{\mu^2}$
- Dispersion ϕ represents the biological variance.
- NB distribution can be parameterized by mean and dispersion.

Simple ideas for RNA-seq DE analysis

- Transform data into continuous scale (e.g., by taking log), then use microarray methods:
  - Troublesome for genes with low counts.
- For each gene, perform two group Poisson or NB test for equal means. But:
  - Number of replicates are usually small, asymptotic theories don't apply so the results are not reliable.
  - Like in microarray, information from all genes can be combined to improve inferences (e.g., variance shrinkage).

DEseq (Anders *et al.* 2010, *GB*)

- Counts are assumed to following NB, parameterized by mean and variance: $K_{ij} \sim \mathrm{NB}(\mu_{ij}, \sigma_{ij}^2)$,
- The variance is the sum of shot noise and raw variance:

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}} .$$

- The raw variance is a smooth function of the mean, so it is assumed that genes with same means will have the same variances.
- Hypothesis testing using exact test:

$$p_i = \frac{\sum_{\substack{a+b=k_{iS} \\ p(a,b) \le p(k_{iA}, k_{iB})}} p(a,b)}{\sum_{a+b=k_{iS}} p(a,b)}.$$

Bioconductor package `DEseq`

- Inputs are:
  - integer matrix for gene counts, rows for genes and columns for samples.
  - experimental design: samples for the columns.

```
library(DESeq)
conds=c(0,0,0,1,1,1)
cds=newCountDataSet(data, conds )
cds=estimateSizeFactors( cds )
cds=estimateVarianceFunctions( cds )
fit=nbinomTest( cds, 0, 1)
pval.DEseq=fit.DEseq$pval
```

edgeR

- From a series of papers by Robinson et al.(the same group developed limma): 2007 *Bioinformatics*, 2008 *Biostatistics*, 2010 *Bioinformatics*.
- Empirical Bayes ideas to "shrink" gene-specific estimations and get better estimates for variances.
- The parameter to shrink is over-dispersion (Φ) in NB, which controls the within group variances.
- There is no conjugate prior so a shrinkage is not straightforward.
- Used a conditional weighted likelihood approach to establish an approximate EB estimator for Φ.

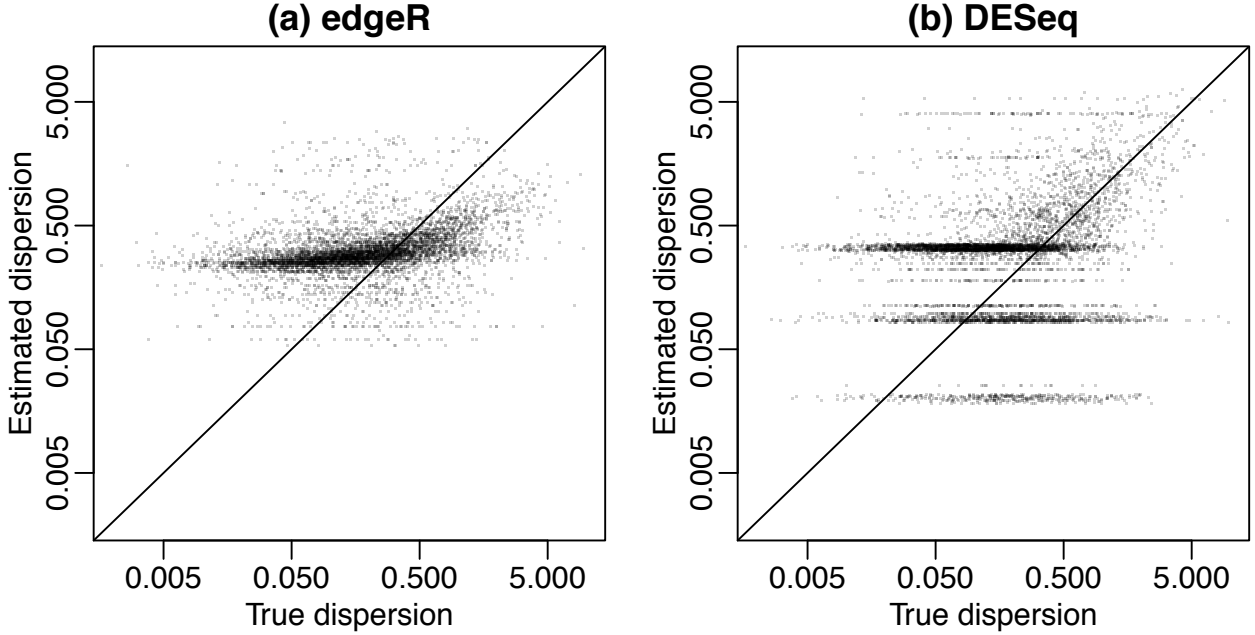
Bioconductor package `edgeR`

- Inputs are the same as DEseq: an integer matrix for counts and column labels for design.

```
library(edgeR)
d = DGEList(counts=data, group=c(0,0,0,1,1,1),
          lib.size=colSums(data))
d = calcNormFactors(d)
d = estimateCommonDisp(d)
d = estimateTagwiseDisp(d, trend=TRUE)
fit.edgeR = exactTest(d)
pval.edgeR = fit.edgeR$table$p.value
```

DSS (Wu *et al.* 2012, *Biostatistics*)

- Found that the shrinkage from DESeq and edgeR are too strong.



A hierarchical model for the counts

$$Y_{gi}|\theta_{gi} \sim \text{Poisson}(\theta_{gi} s_i)$$

$$\theta_{gi}|\phi_g \sim \text{Gamma}(\mu_{g,k(i)}, \phi_g)$$

$$\phi_g \sim \text{log-normal}(m_0, \tau^2)$$

- Y_{gi}: observed counts for gene *g*, sample *i*
- Θ_{gi}: unobserved true expression for gene *g*, sample *i*
- Φ_g: dispersion (related biological variance) for gene *g*.
- s_i: library size for sample *i*.

The posterior

- Negative binomial is parameterized by mean and dispersion, then the posterior for dispersion is:

$$\begin{aligned}\log[p(\phi_g|Y_{gi},\nu_{gi},i=1,\ldots,n)] \propto & \sum_i \psi(\phi_g^{-1}+Y_{gi}) - n\psi(\phi_g^{-1}) - \phi_g^{-1}\sum_i \log(1+\nu_{gi}\phi_g) \\ & + \sum_i Y_{gi}[\log(\nu_{gi}\phi_g) - \log(1+\nu_{gi}\phi_g)] \\ & - \frac{[\log(\phi_g)-m_0]^2}{2\tau^2} - \log(\phi_g) - \log(\tau), \qquad (4.1)\end{aligned}$$

- Here, $\nu_{gi} = \mu_{g,k(i)} s_i$ is the expected value for Y_{gi}.
- It's a penalized likelihood to penalize (1) dispersions far away from prior mean; and (2) large dispersions.

Testing and inference in two-group comparison

- Wald test: $t_g = \frac{\hat{\mu}_{g,1} - \hat{\mu}_{g,2}}{\sqrt{\hat{\sigma}^2_{g,1} + \hat{\sigma}^2_{g,2}}}$
  - With dispersions, variances can be computed according to NB distribution: $var = \mu + \mu^2\phi$
  - The variance for $\hat{\mu}_{g,1}$ is: $\hat{\sigma}^2_{g,1} \equiv \frac{1}{n_1^2}\left[\hat{\mu}_{g,1}\left(\sum_{j:k(j)=1}\frac{1}{s_j}\right) + n_1\hat{\mu}^2_{g,1}\tilde{\phi}_g\right]$
- Inferences: use normal P-values and local FDR.

DSS Bioconductor package

- Inputs are the same as DEseq and edgeR: an integer matrix for counts and column labels for design.

```
conds=c(0,0,0,1,1,1)
seqData=newSeqCountSet(X, conds)
seqData=estNormFactors(seqData)
seqData=estDispersion(seqData)
result=waldTest(seqData, 0, 1)
```

Summary for RNA-seq DE test

- Based on my experiences and simulation results:
  - All methods provide very similar results when the variation of dispersions is small.
  - When variation of dispersions is large, DSS outperforms.
- Testing procedure for multiple-factor design is still not well developed. Mostly based on glm.
- Some rooms for statistical researches.

Review of this class

- I have covered some aspects of next generation sequencing :
  - Biological backgrounds.
  - Sequencing technologies.
  - Brief introduction to NGS applications: DNA-seq, RNA-seq and ChIP-seq.
  - RNA-seq analyses: motivation, statistical methods for DE analysis (very briefly), useful software.