# Analysis of High-Throughput Sequencing Data with R and Bioconductor

## Workshop Introduction

Tyler Backman, Rebecca Sun & Thomas Girke

October 29, 2010

# Outline

# History of DNA Sequencing

1977 "DNA Sequencing by Chemical Degradation" is published by Allan Maxam and Walter Gilbert.

1977 "DNA Sequencing by Enzymatic Synthesis" is published by Fred Sanger.

1980 Fred Sanger and Walter Gilbert receive the Nobel Prize in Chemistry.

1982 GenBank starts as a public repository of DNA sequences.

1986 Leroy Hood's laboratory at the California Institute of Technology announces the first semi-automated DNA sequencing machine.

1997 Genome sequence of E. coli is published.

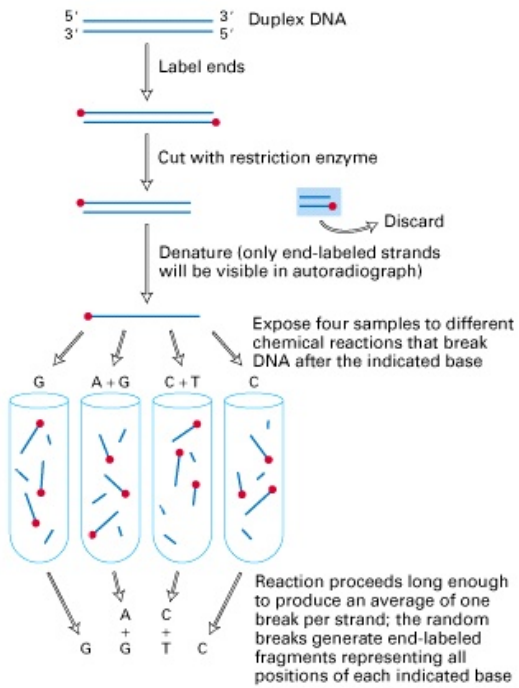
2001 Draft sequence of the Human genome is published.

2004 Next generation sequencing technologies become available to the public.

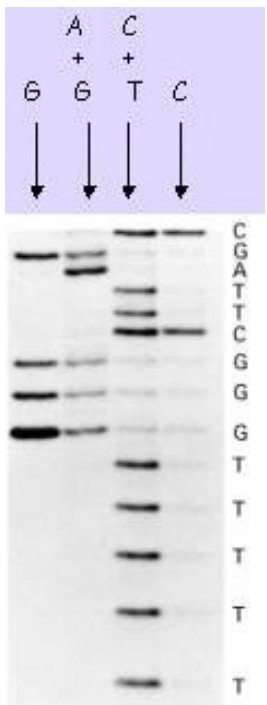
# Outline

# Chemical Sequencing by Maxam & Gilbert

1. Uses radioactive labeled DNA fragments of 500 bp.
2. Four separate chemical treatments generate DNA breaks at the positions: G, A+G, C, C+T.
3. The fragments are size-separated by gel electrophoresis in four separate lanes.
4. Visualization of the fragments by autoradiography on an X-ray film.



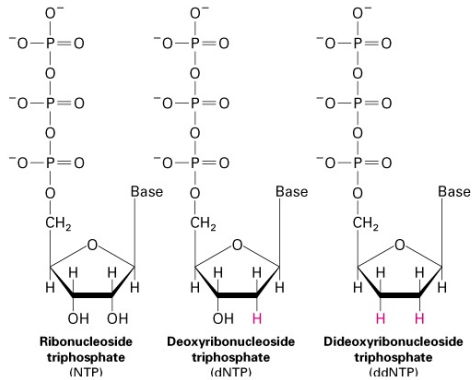

Chemical DNA Degradation

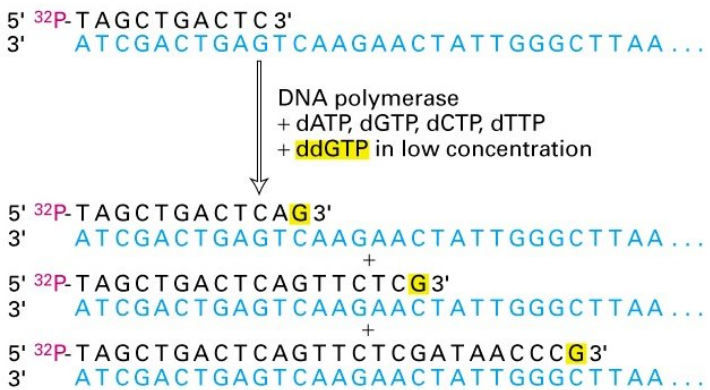



Gel Electrophoresis

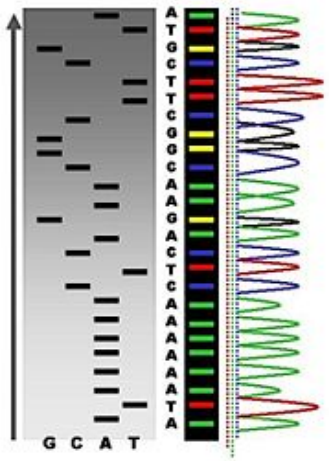
# Outline

# Illustration of Sanger Sequencing



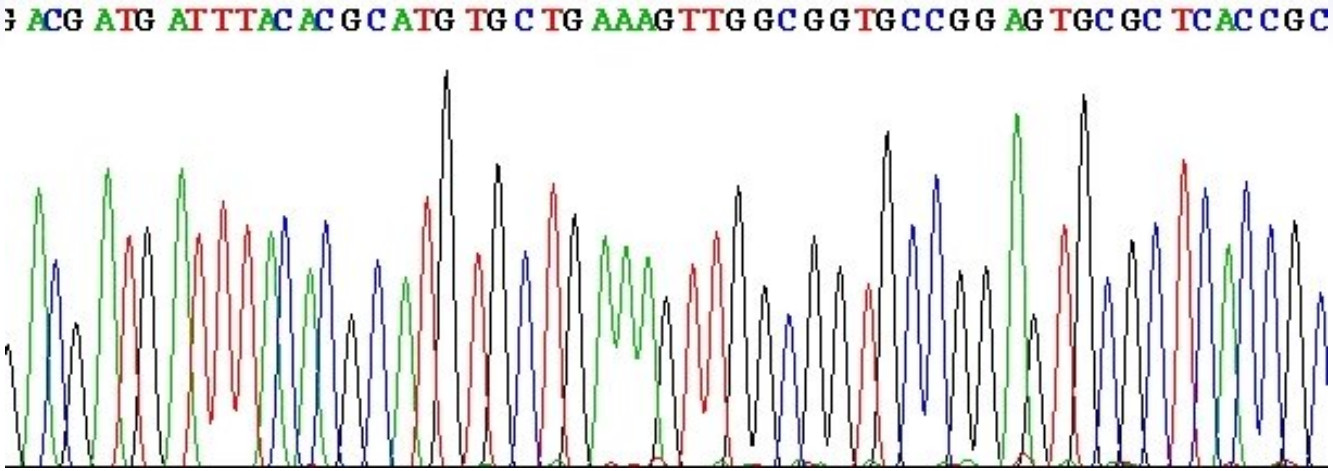




Sequencing Principle




Radioactive Fluorescence Labeling

# Processing of Sequencing Raw Data




- Assign quality score to each peak
- The frequently used Phred scores provide log(10)-transformed error probability values:
  - score = 20 corresponds to a 1% error rate
  - score = 30 corresponds to a 0.1% error rate
  - score = 40 corresponds to a 0.01% error rate
- The base calling (A, T, G or C) is performed based on Phred scores.
- Ambiguous positions with Phred scores $\leq 20$ are labeled with N.

# Outline

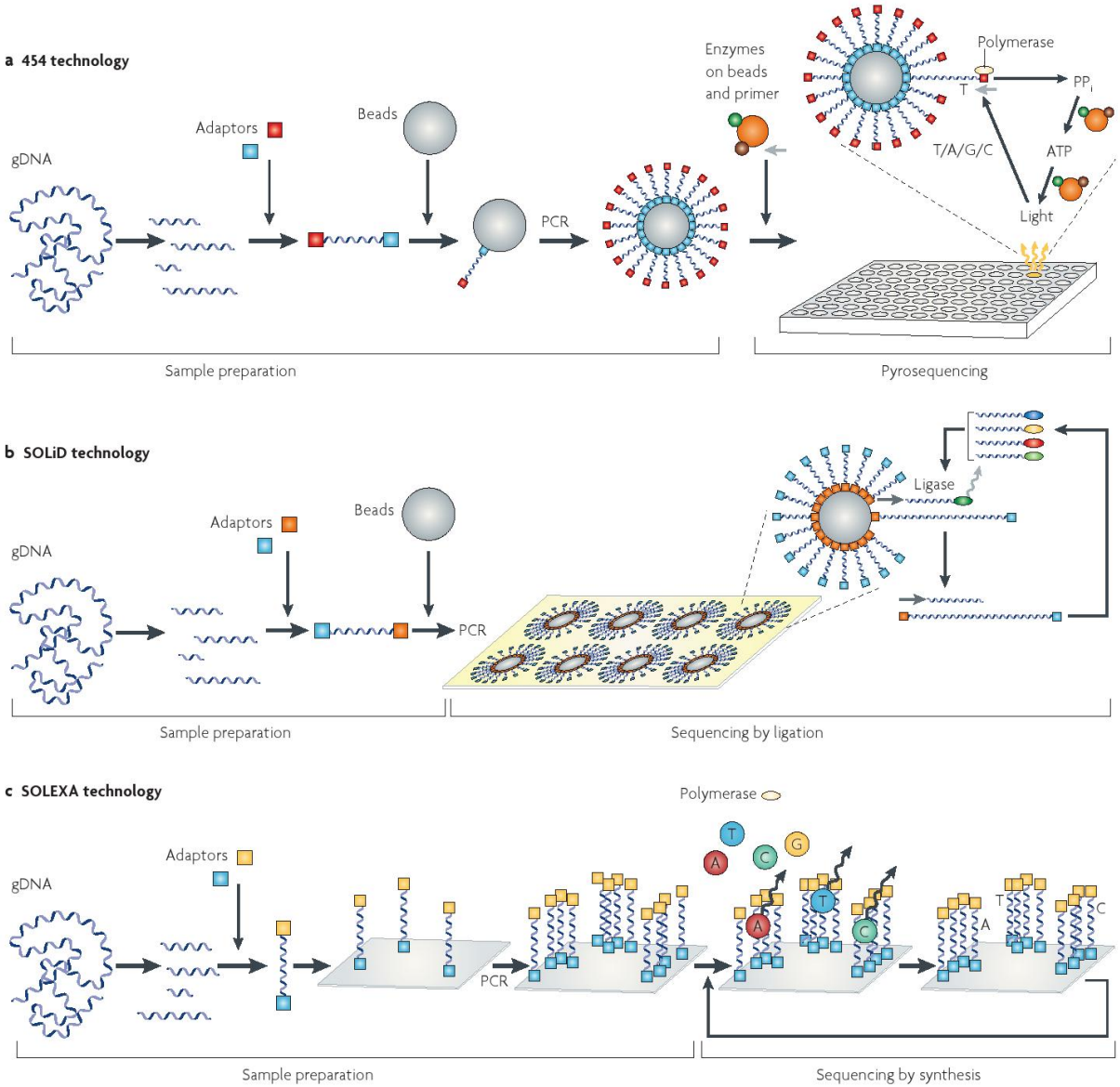
# Common Synonyms

- High-throughput sequencing: HTS or HT-Seq
- Flow cell sequencing (FCS)
- Massively parallel sequencing (MPS)
- Next/this generation sequencing (NGS/TGS)
- Deep sequencing
- Sequencing by synthesis
- Many other synonyms

Review article: [Holt et al 2008]

# Overview: 454, SOLiD and Illumina

From review article: [Medini et al 2008]



# Similarities and Differences of HT-Seq Technologies

## Common components

- Flow cells as reaction chambers
- Iterative sequencing process
- Massive parallelization
- Clonally amplified or single molecule templates

## Differences

- Template preparation
- Sequencing chemistry
- Flow cell configuration

# HT Sequencing Methods

Reversible Terminator Methods (*e.g.* Illumina/Solexa)

- Use reversible versions of dye-terminator reactions.
- Principle steps: adding one nucleotide at a time, detecting fluorescence corresponding to that position, then removing the blocking group to allow the polymerization of another nucleotide.

Single Molecule Methods (*e.g.* Helicos)

- Sequences one of the four nucleotides per cycle.

Pyrosequencing Methods (*e.g.* 454)

- Also use DNA polymerization to add nucleotides.
- Principle steps: adding one type of nucleotide at a time, then detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates.

Supported Oligonucleotide Ligation Methods (*e.g.* SOLiD, Complete Genomics)

- Uses ligation-based approach
- Principle steps: stepwise ligation of labeled random octamers to obtain sequence of attached dinucleotides; the ligated dinucleotides of each ligation round are spaced by several nucleotides; continuous sequence information is obtained by offsetting the sequencing primer.

# Outline

# Example: Illumina/Solexa Technology

Illumina Sequencer

Flow Cell

# Basic Steps of Illumina/Solexa Sequencing Technology

Compare with illustration on next three slides!

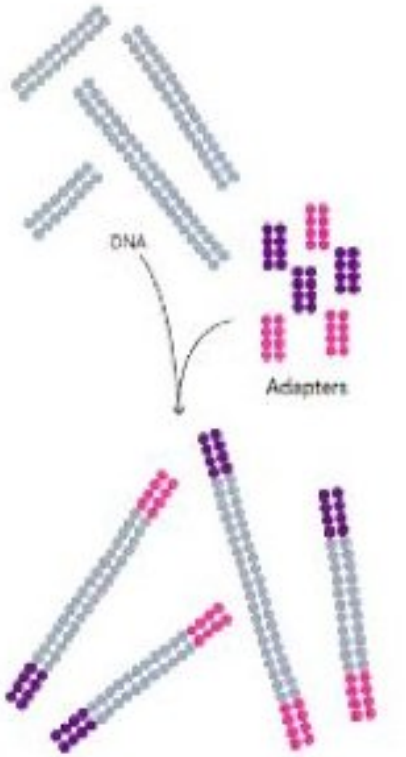
## Flow Cell Loading

1. Generate DNA library (genomic- or cDNA-based) with insert length of ∼200 bp.
2. Load library onto flow cell (nano device for liquid handling).
3. PCR-based bridge amplification of loaded fragments to obtain DNA clusters (serves signal amplification)

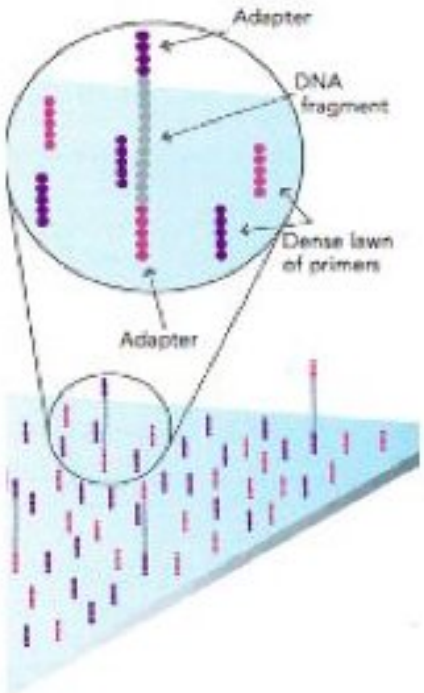
## Sequencing Cycles

4. Start reversible dye-terminator reaction containing primer and labeled dNTPs among other components.
5. Image scan to detect the identity of first base of each cluster via the characteristic fluorescence signal for each labeled nucleotide.
6. De-protection step removes the blocking group and fluorescence group of the incorporated nucleotide.
7. Repeat steps 4-6 about 30-60 times.

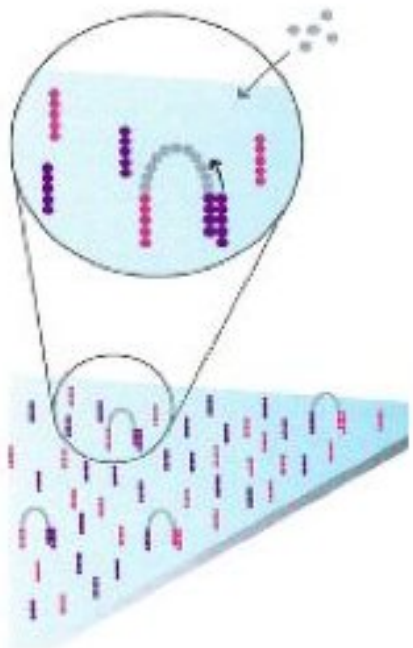
# Loading of Flow Cell



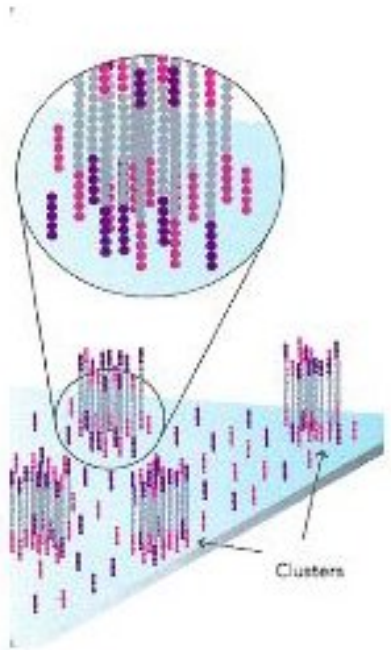

DNA Fragments
With Adaptors




Bind DNA to Surface
of Flow Cell



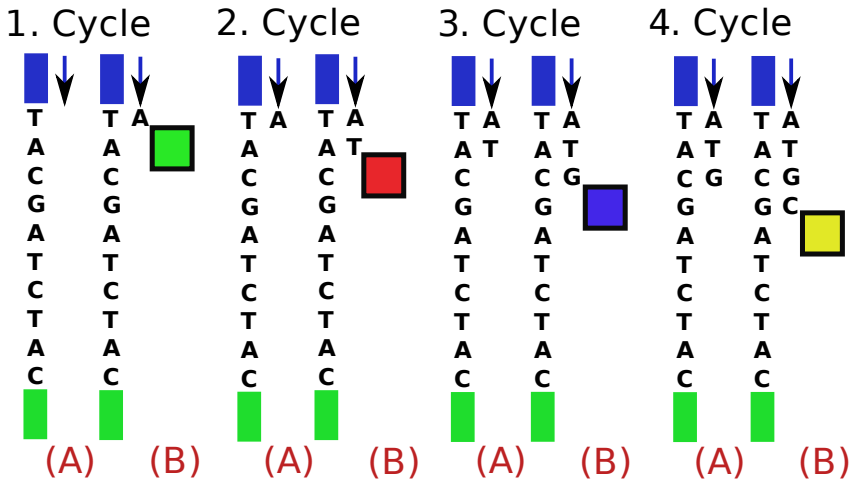
Bridge Amplification




X Amplification Rounds
10 Mio molecule clusters
each diameter of 1um
1000 molecules per cluster

# Details of Sequencing Reaction

## Sequencing Reaction




(A): Polymerase attaches labeled and blocked nucleotides
(B): Deblocking and removal of label

## Data Collection


**Legend**

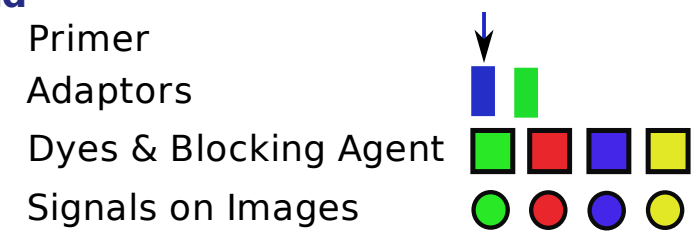
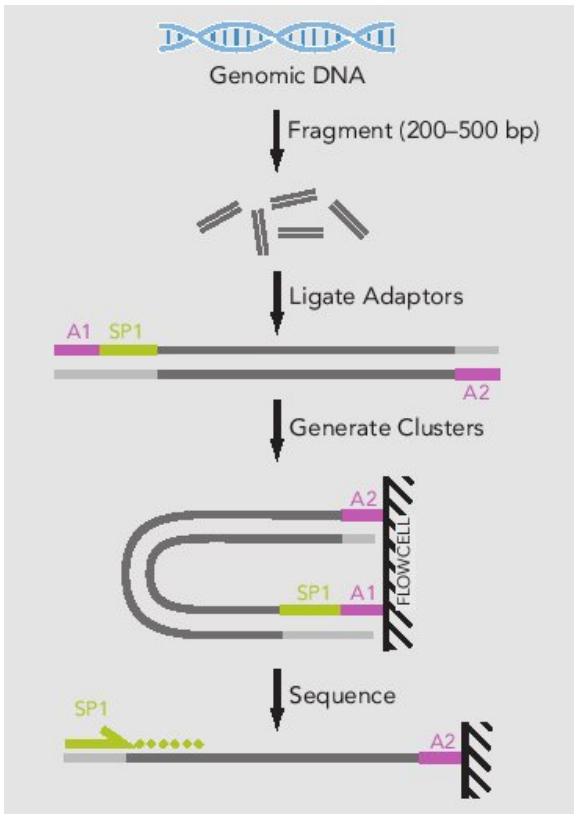


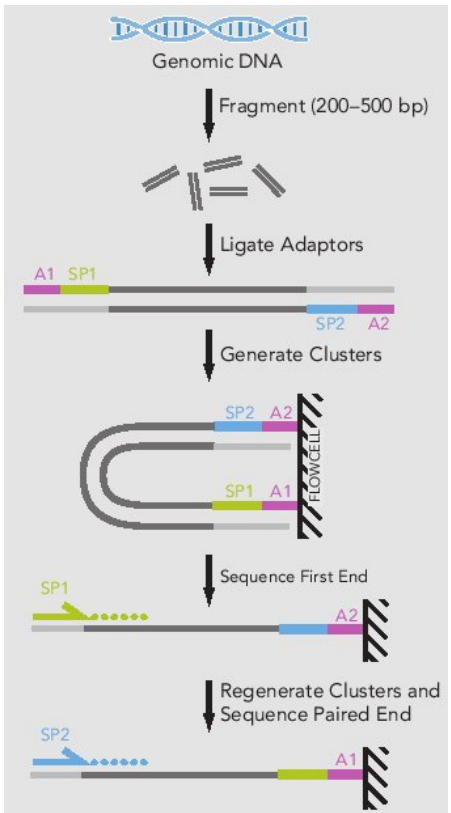

Illustration shows the sequencing cycles for a single template molecule!

# Single End, Paired End and Mate Pair Sequencing

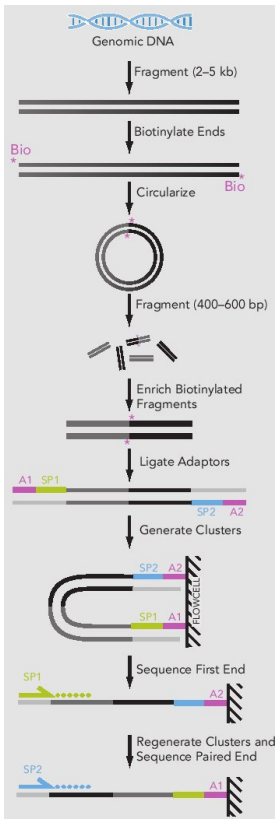
Single End




Paired End




Mate Pair

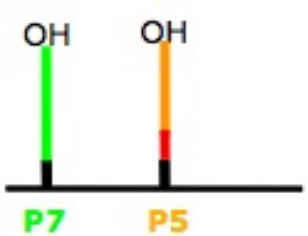



AP1/AP2: flow cell adapators; SP1/SP2: sequencing primers

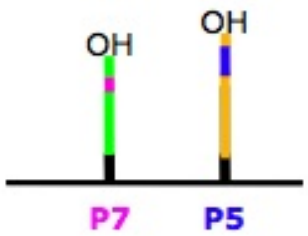
# Paired End Chemistry: Step I

## Grafted Flow Cell

Single End

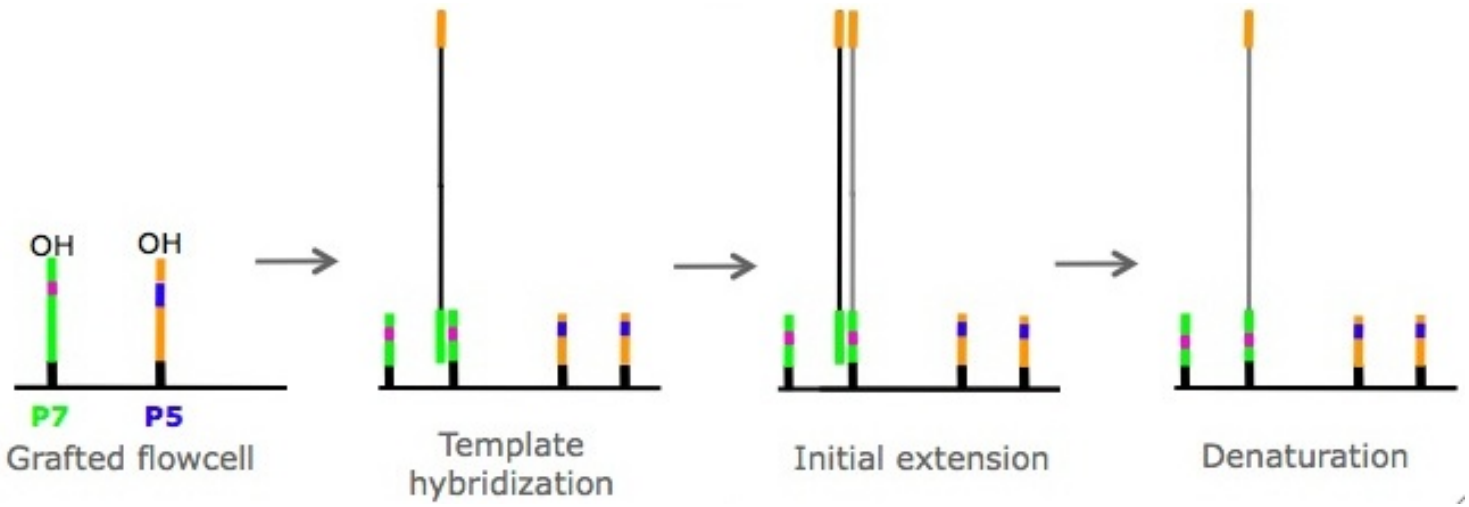



Paired End



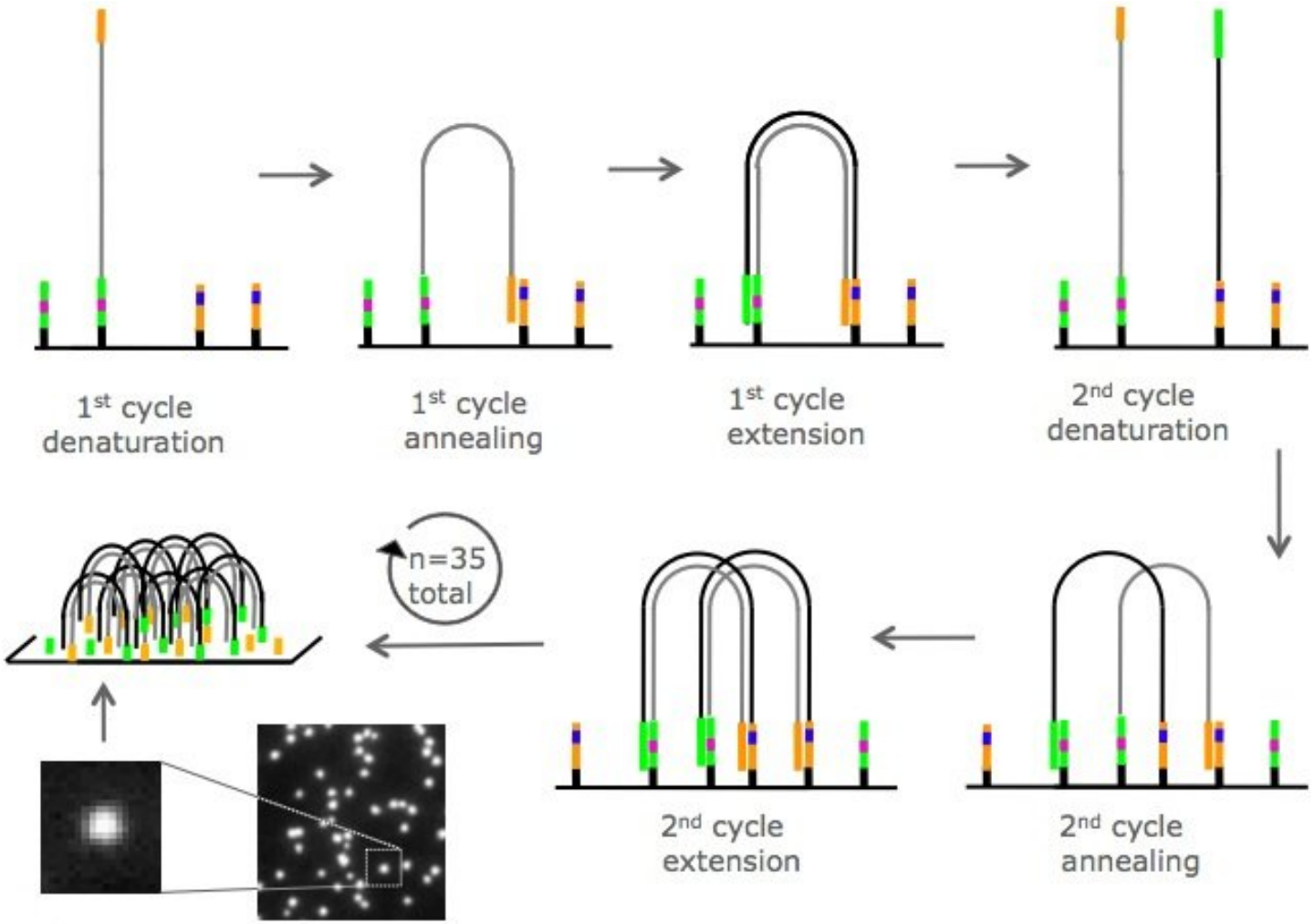

Linearization: periodate two different enzymes

## Cluster Generation: Initial Extension



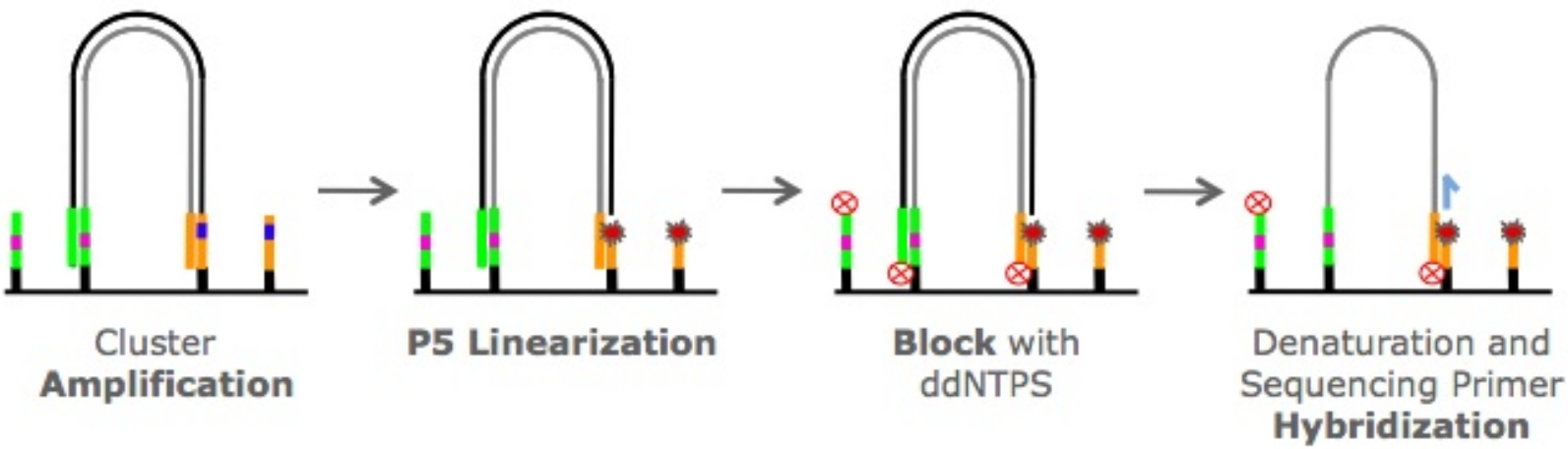
# Paired End Chemistry: Step II

## Cluster Generation: Amplification



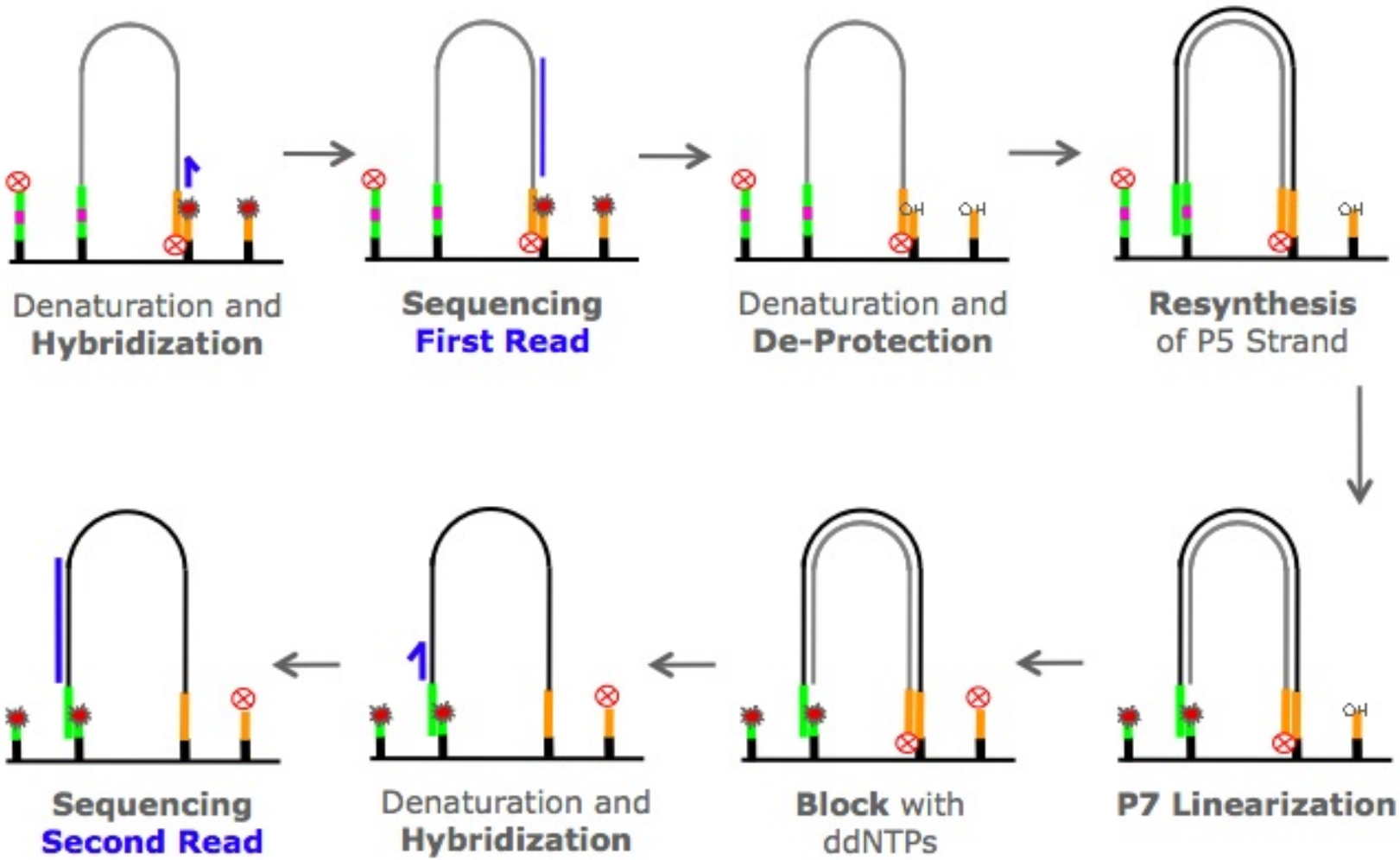
# Paired End Chemistry: Step III

## Cluster Generation: Linearization



# Paired End Chemistry: Step IV

Sequencing



# Processing of Illumina Sequencing Data

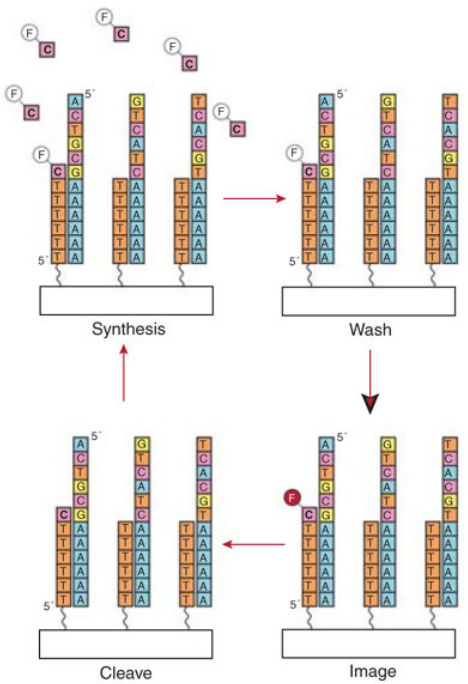
- Convert cluster images to intensity values.
- Base calling based on intensity for each fluorescence dye.
- Generate quality scores similar to Phred scores.
- The length of each sequence corresponds to the number of cycles, e.g. 36 cycles $\rightarrow$ 36 bp.
- Remove sequences with low quality reads.

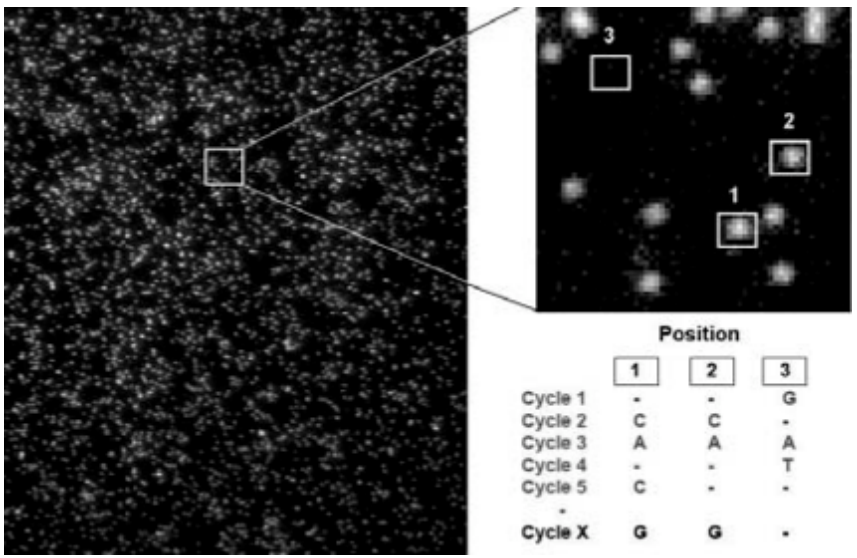
In 2008 a single sequencing run with 36 cycles could generate $\sim 1.5$ billion bp of sequence information and $\sim 130,000$ images or 1TB of image data.

# Outline

# Helicos: Single Molecule Sequencing

- Has similarities to Solexa/Illumina technology, but sequences single molecule templates.
- Attaches one of the four nucleotides at a time using proprietary nucleotide-polymerase formulations. This prevents the incorporation of more than one nucleotide in each cycle in homopolymer regions.



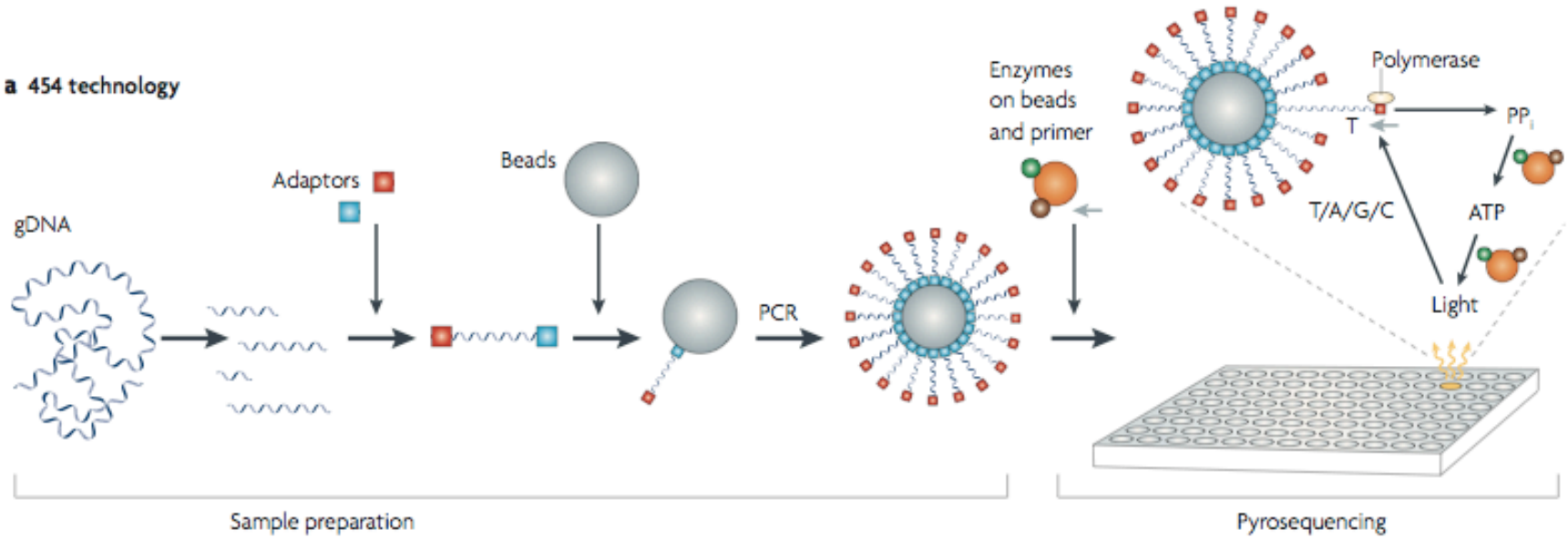
# Outline

# 454 Sequencing Steps

Pyrosequencing Methods (*e.g.* 454)

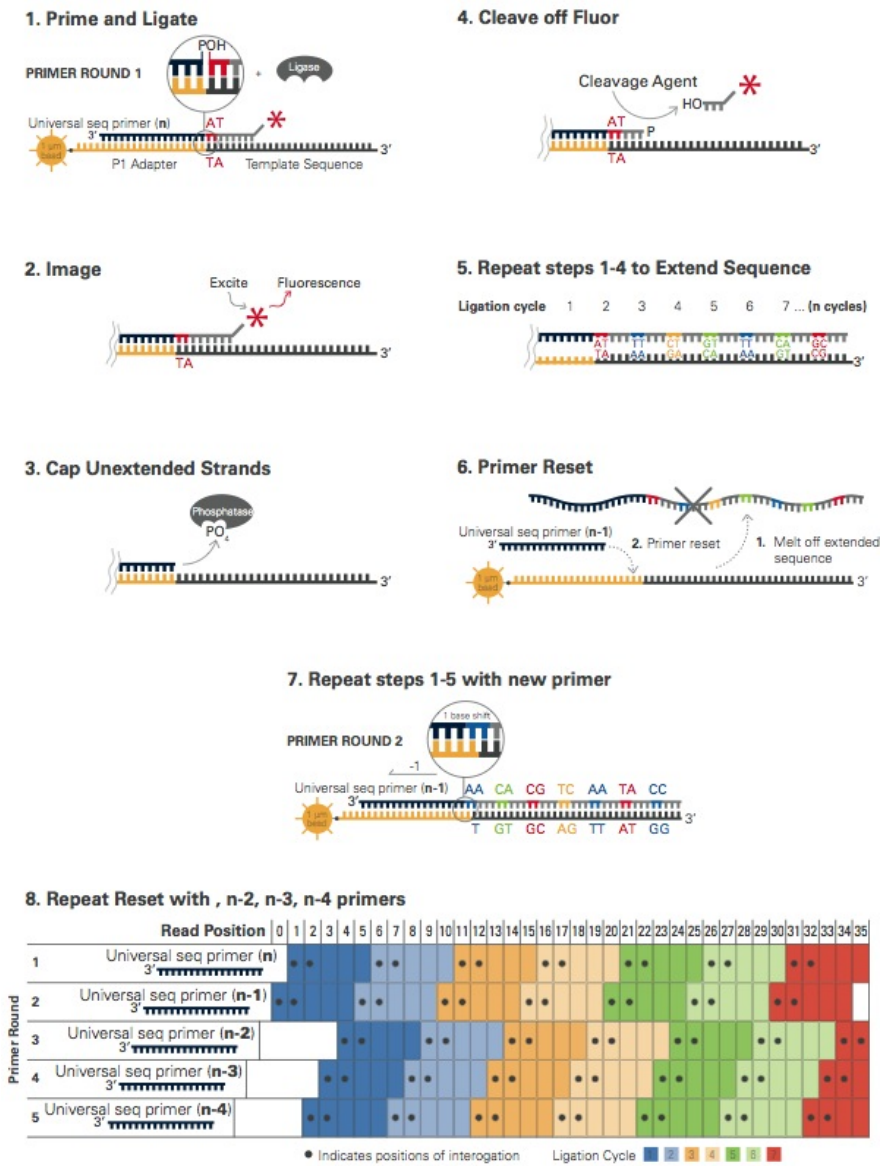
- Also uses DNA polymerization to add nucleotides.
- Principle steps: adding one type of nucleotide at a time, then detecting and quantifying the number of nucleotides added to a given location through the light emitted by the release of attached pyrophosphates.
- For more details see: 454 Web Site (http://www.454.com)

# 454 Pyrosequencing



# Outline

# SOLiD: Sequencing by Supported Oligonucleotide Ligation and Detection



# Comparison of HT-Sequencing Methods

Technology in 2008!!

| | Illumina | 454 | SOLiD | Helicos |
|---|---|---|---|---|
| Method | Rev. Term. | Pyro. Sequ. | Oligo Ligation | Single Mol. |
| Read Length | 36-2x100 | 300-400 | 36 | 25-45 |
| Error Rate | $\sim$1% | $>$1% | $\sim$0.1% | $<$1% |
| Data/Run (Gb) | 1-3 | 0.1 | 2-3 | 8 |
| Cost (per Gb) | \$6,000 | \$84,000 | \$6,000 | \$2,500 |

Data largely from [Holt et al 2008]

# Comparison with Traditional Methods

| Method | Read Length | Sequences per Run | Utility |
|---|---|---|---|
| Dye-Terminator (Sanger) | 500-1500 bp | 384 | *de novo* and low-throughput applications |
| 454/Roche | 120-400 bp | ∼200,000 | *de novo* and medium-throughput applications |
| Illumina/Solexa | 36-60 bp | ∼20,000,000 | high-throughput applications |

- *de novo* applications do not require guide genome sequence, *e.g.* new genomes, etc.
- high/medium-throughput applications require guide genome sequence, *e.g.* SNPs, small RNAs, etc.

All numbers are estimates and apply to the situation in Feb. 2009!

# Outline

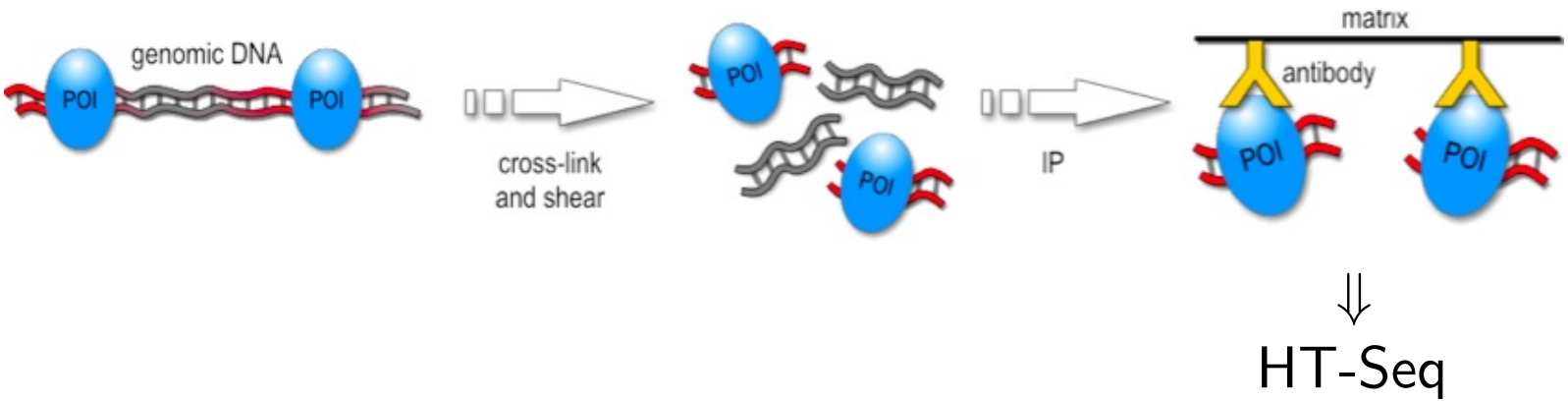
# Applications of HT-Sequencing Methods

HT-Seq technologies provide endless opportunities for genomics, comparative genome biology, medical diagnostics, etc. The following examples provide only a brief overiew.

## Applications

- Genome-wide detection of SNPs and mutations (SNP-seq)
- Methylome profiling by bisulphite sequencing (BS-seq)
- DNA-protein interactions (ChIP-seq)
- Transcriptome sequencing (RNA-seq)
- mRNA expression profiling (DGE)
- Small RNA profiling and discovery

# Application: DNA-Protein Interactions with ChIP-Seq

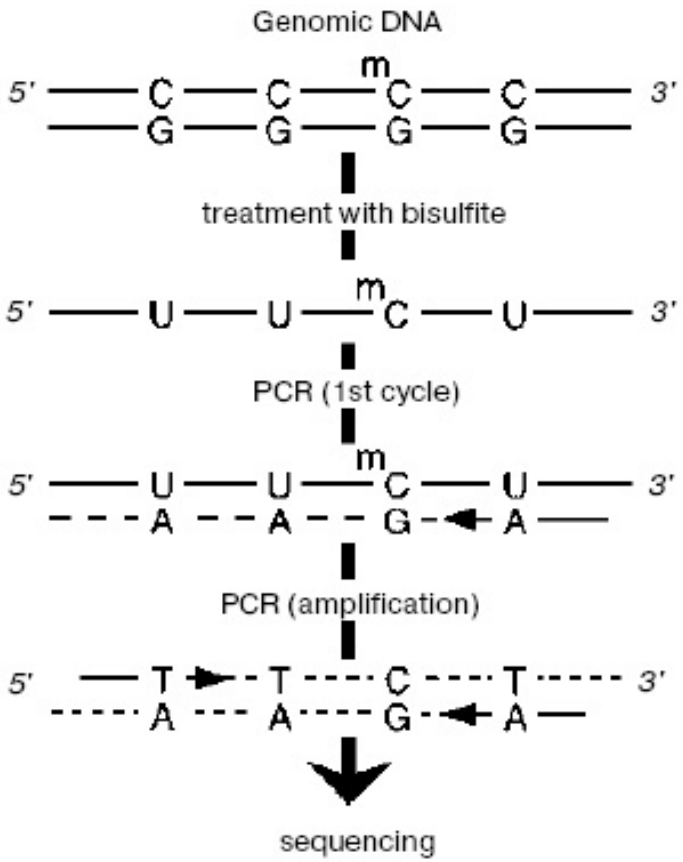



$\Downarrow$

HT-Seq

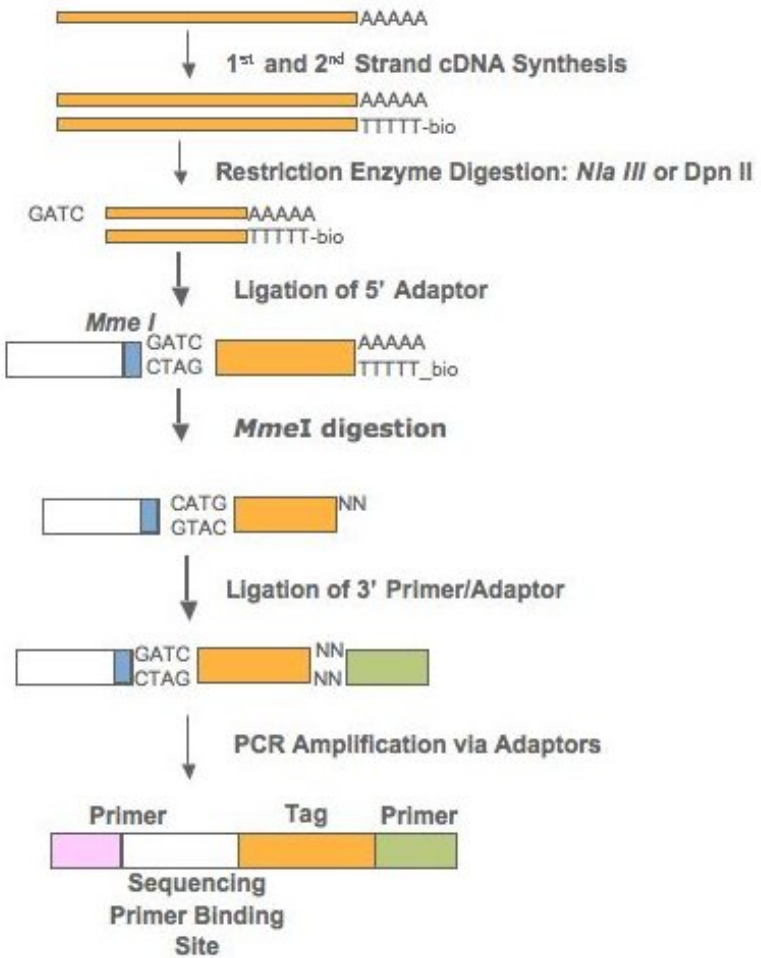
Reference for ChIP-Seq data analysis: [Jothi et al 2008]

# Application: Methylome Profiling with BS-Seq



# Application: Digital Gene Expression (DGE) Profiling

## Illustration for Illumina's DGE

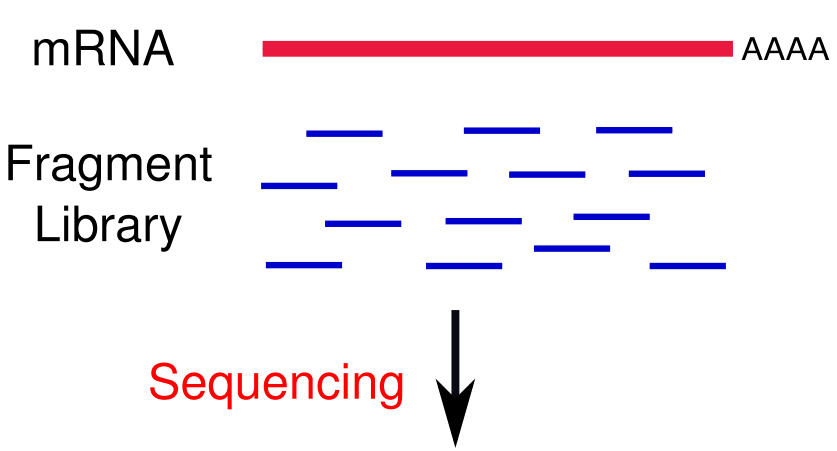



$\Downarrow$

HT-Seq of Tags

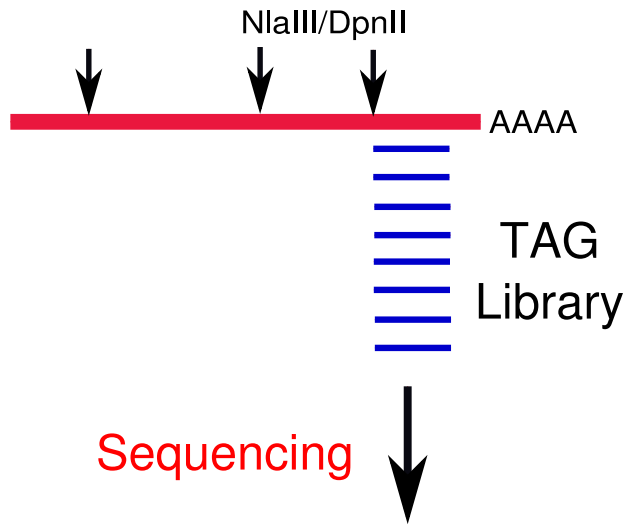
# RNA-Seq versus DGE

## RNA-seq




1. Alternative splicing
2. Limited expression profiling
3. SNP detection
4. Many other applications

## DGE

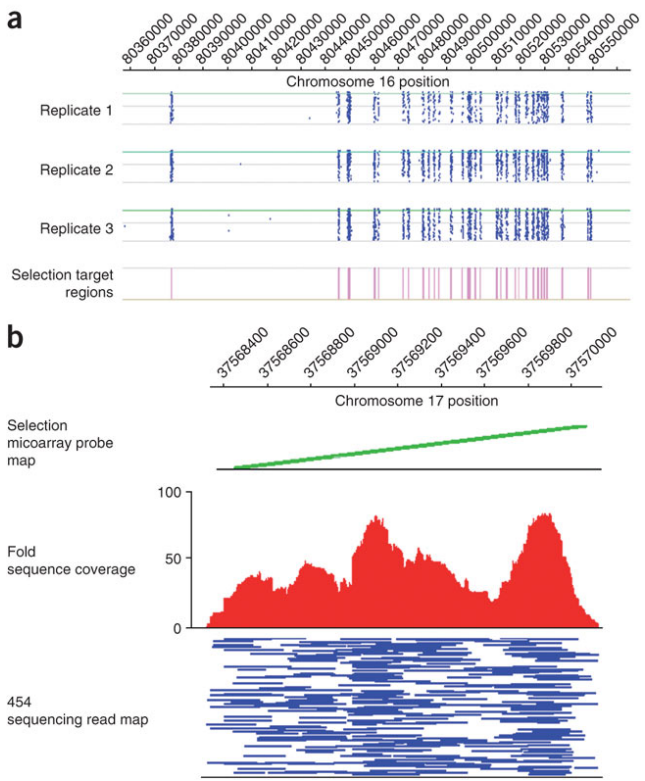



1. Expression profiling
→ more appropriate for many biosamples

# Next Major Technology Improvement

## Targeted Sequencing Using DNA Capture Microarrays

- Powerful approach to make HT-Seq more economic and versatile.




- Example: usage of programmable microarrays (here NimbleGen) to enrich for DNA regions of interest [Albert et al 2007].

# Database: Short Read Archive from NCBI

- SRA: http://www.ncbi.nlm.nih.gov/Traces/sra

# Outline

# Performance Requirements

## Aligning tens of millions of sequences requires:

- Ultra fast search algorithms (100-1000x faster than BLAST)
- Small memory footprint
- Economic data structures and containers

## Alignment requirements

- The requirements for short-read mapping applications are very different from traditional sequence database search approaches for ortholog identification.
- Most short-read alignment algorithms will not work for longer sequences!
- However, most of them are more sensitive for short-reads than BLAST, because they lack its word size limitation.
- Only best hits with almost perfect alignments are required.
- Lower scoring alternative hits (more mismatches) are less interesting.
- Often only perfect matching required, but with the possibility to allow 1-2 mismatches and only sometimes very short gaps.

# List of Short-Read Alignment Tools (not complete)

- Indexing Reference with Suffix Array/Burrows-Wheeler
  - Bowtie [Langmead et al 2009]
  - SOAPv2
- Indexing Reads with Hash Tables
  - ZOOM: uses spaced seeds algorithm [Lin et al 2008]
  - RMAP: simpler spaced seeds algorithm [Smith et al 2008]
  - SHRiMP: employs a combination of spaced seeds and the Smith-Waterman
  - MAQ [Li et al 2008b]
  - Eland (commercial)
- Indexing Reference with Hash Tables
  - SOAPv1 [Li et al 2008]
- Merge Sorting
  - Slider [Malhis et al 2009]

# MAQ: Mapping Quality

## Algorithm [Li et al 2008b]

- Uses a hashing technique that guarantees to find alignments with up to two mismatches in the first 28 bp of the reads.
- It indexes the read sequences in six hash tables and scans the reference genome sequence for seed hits that are subsequenty extended and scored.
- The commercial Eland alignment program uses a very similar approach.

## Performance

- Slower than Bowtie and SOAP. Intermediate rank with regard to memory footprint.

## Features

- Versatile pipeline for SNP detection.
- Can report all hits for queries with multiple ones.
- Allows at most two mismatches.
- Performs on single reads only ungapped alignments. Gaps only possible for paired end reads by applying Smith-Waterman algorithm on small candidate set.

## Limitations

- Read length limit 128bp
- No gapped alignment for single-end reads
- All sequences in one run need to to have the same length.

# MAQ Algorithm - Step 1: Indexing

- It builds six hash tables to index the reads and scans the reference sequence against the hash tables to find the hits. This ensures that sequences with up to two mismatches will be found.
- The six hash tables correspond to six noncontiguous seed templates. For example, the six templates for 8bp reads are 11110000, 00001111, 11000011, 00111100, 11001100 and 00110011.
- Only the nucleotides at 1's will be indexed, but not those at 0's using a 24-bit integer for hashing.
- When all the reads are processed, the 24-bit integers are sorted, such that reads with the same hashing integer are grouped together. Each integer and its corresponding region are then recorded in a hash table with the integer as the key.
- The indices are built for only the first 28bp of the reads, which are typically the most accurate part of the read.

# MAQ: Algorithm - Step 2: Searching

- Each 28-bp subsequence of the reference will be hashed with the first two templates (TMPA in Fig. 1) used for indexing and will be looked up in the corresponding two hash tables.
- It maps the reads to a position that minimizes the sum of quality values of mismatched bases.
- When the scan of the reference is complete, the next two templates (TMPB in Fig. 1) are applied and the reference will be scanned once again (TMPC in Fig. 1) until no more templates are left.
- Using six templates guarantees to find seed hits with no more than two mismatches.
- If a hit is found to a read, MAQ will calculate the sum of qualities of mismatched bases $q$ over the whole length of the read, extending out from the 28bp seed without gaps.

# MAQ: Algorithm - Illustration

## Indexing and Searching

```
Ref4 TACGCGAT    2 continuous mismatches
Ref3 AACCGGAT    2 spaced mismatches
Ref2 AACGCGAT    1 mismatch
Ref1 ATCGCGAT    perfect match
Read ATCGCGAT    Ref1 Ref2 Ref3 Ref4
TMPA 11110000     Y    N    N    N
TMPA 00001111     Y    Y    N    Y
TMPB 11000011     Y    N    N    N
TMPB 00111100     Y    Y    N    Y
TMPC 11001100     Y    N    N    N
TMPC 00110011     Y    Y    Y    Y
```

Fig. 1: Indexing and Search Strategy of MAQ Algorithm

# MAQ: Algorithm - Error Statistics for Genotype Modeling

- Error probabilities are computed for the final genotype calls.
- It uses a Bayesian statistical model.
- This model incorporates:
  - mapping qualities
  - error probabilities from the raw sequence quality scores
  - sampling of the two haplotypes
  - an empirical model for correlated errors at a site

# SOAP

## Algorithm [Li et al 2008]

- Uses a combination of seed and hash look-up table algorithms

## Performance

- 300-1200 faster and higher sensitivity than BLASTN

## Features

- Can report all hits for queries with multiple matches.
- Allows at most two mismatches.
- Performs ungapped and gapped alignments. At most one gap of 1-3 bases in length.
- Mismatches have precedence over gaps.
- Does not allow gaps in flanking regions of gapped alignments.
- Aligns paired end reads simultaneously (only one read can have gap).

## Limitations

- High memory requirements for large genomes: 14GB RAM for human genome

# SOAP: Algorithm

- Uses seed and hash look-up tables to accelerate search and alignment processes.
- Both the reads and the reference sequences are converted by a hashing function to numeric data using 2-bits-per-base for encoding. A search is performed by exclusive-OR comparisons with the reference sequence. Then the value is used as suffix to check the look-up table to know how many bases are different.
- To allow two mismatches, it uses a similar seed template approach as Eland and MAQ by splitting each read into four regions.
- Uses the enumeration algorithm for inserting gaps. This method tries to insert a continuous gap or deletion at each possible position of a read.
- The algorithm obtains the same alignments as dynamic programming but much faster.
- Unlike Eland and Maq, SOAP loads the reference sequences into memory. For each read, it creates seeds and searches the corresponding index table for candidate hits, computes an alignment and reports the results.

# Bowtie

## Algorithm [Langmead et al 2009]

- Burrows-Wheeler index based on the full-text minute-space (FM) index.

## Performance

- Aligns sequences of 4-1,024 bases.
- Handles sequences of variable length in a single run.
- Index requires only 1.3GB memory for the human genome (works on laptop!).
- Aligns 25 million 35bp reads per CPU-hour.
- Fastest of all short-read alignment programs: 35 faster than MAQ and 3 times faster than SOAP.
- With default settings, sensitivity similar to SOAP, but slightly lower than MAQ.

## Limitations

- Requires BWT index, which takes several hours to compute.
- It reports inexact matches, but does not guarantee to find the match with the highest quality alignment.
- With its highest performance settings, it may fail to align a small number of reads with multiple mismatches.
- Increased accuracy options can overcome some of these limitation at the cost of speed performance.

# Bowtie: Burrows-Wheeler Transform (BWT)


(a) The Burrows-Wheeler matrix and transformation for 'acaacg'. Data compression is facilitated by forming stretches of the same characters.

(b) An unpermute step repeatedly applies the last first (LF) mapping to recover the original text (in red on the top line) from the Burrows-Wheeler transform (in black in the rightmost column).

(c) Steps taken by an exact search to identify the range of rows, and thus the set of reference suffixes, prefixed by 'aac'.

# Bowtie: Searching for Inexact Matches

- If the range of an exact search becomes empty, then the algorithm selects an already-matched query position and substitutes its base by different ones (mismatch). Then the exact search resumes from the substituted position.
- If there are multiple candidate substitution positions, then the algorithm greedily selects a position with a minimal quality value.

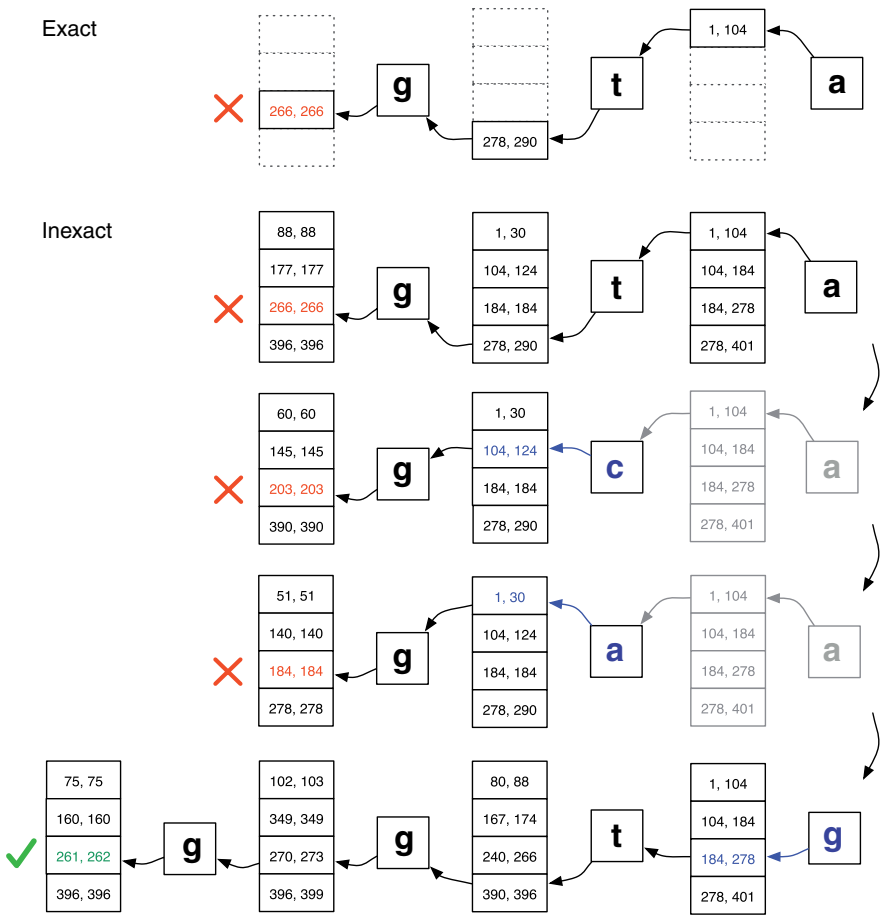



Fig. 2: Inexact match for query 'GGTA'

# Useful Links

- Link Collections for Short-Read Alignment Tools
  - Alignment Tools List from MAQ Developer
  - Alignment Performance Page from BWA Developer

# Outline

# Software for Illumina Sequence Data Analysis

| Type | Software Packages |
|---|---|
| Image Analysis | RTA (Real Time Analysis) |
| Base Calling | OLB (Off-Line Base caller), Alta-Cyclic, Rolexa (R) |
| Alignment | CASAVA Eland, BWA, Soap, Mosaik, Bowtie, 40 more!! |
| Quality Control | Illumina pipeline, SolexaQA, TileQC, BioC |
| RNA-Seq, ChIP-Seq, SNP-Seq | Over 100 tools including BioC |
| Assembly | Velvet, SOAPdenovo, ABySS, Euler, 20 more!! |
| Multipurpose Tools | SamTools, BioC, Perl, Python, ... |

# Data Formats for Illumina Reads

| Format | Description |
|---|---|
| fastaq | sequences and ASCII quality scores |
| sequence text file | sequences and numeric quality scores |
| quality score text file | Detailed quality scores (4 per base) |

# Biosequence Analysis in R and Bioconductor

## R Base

- Some basic string handling utilities. Wide spectrum of numeric data analysis tools.

## BioConductor

- Biostrings: general sequence analysis environment.
- ShortRead: pipeline for short read data.
- IRanges: infrastructure for positional data.
- BSgenome: BioC genome annotation data.
- biomaRt: interface to BioMart annotations.
- rtracklayer: interface to online and other genome browsers.
- chipseq & ChIPpeakAnno: Chip-Seq analysis.

## Non-R Alignment Tools

- SOAP
- MAQ
- Bowtie

# Outline

# Continue on HT-Seq Manual

"HT Sequence Analysis with R and Bioconductor"

# Outline

# References and Books


Albert, T J, Molla, M N, Muzny, D M, Nazareth, L, Wheeler, D, Song, X, Richmond, T A, Middle, C M, Rodesch, M J, Packard, C J, Weinstock, G M, Gibbs, R A (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods, 4: 903-905.
URL http://www.hubmed.org/display.cgi?uids=17934467

Holt, RA, Jones, SJ (2008) The new paradigm of flow cell sequencing. Genome Res, 18: 839-846.
URL http://www.hubmed.org/display.cgi?uids=18519653

Jothi, R, Cuddapah, S, Barski, A, Cui, K, Zhao, K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res, 36: 5221-5231.
URL http://www.hubmed.org/display.cgi?uids=18684996

Langmead, B, Trapnell, C, Pop, M, Salzberg, S L (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome Genome Biol, 10:.
URL http://www.hubmed.org/display.cgi?uids=19261174


Li, R, Li, Y, Kristiansen, K, Wang, J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics, 24: 713-714.
URL http://www.hubmed.org/display.cgi?uids=18227114


Li, H, Ruan, J, Durbin, R (2008b) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 18: 1851-1858.
URL http://www.hubmed.org/display.cgi?uids=18714091


Lin, H, Zhang, Z, Zhang, M Q, Ma, B, Li, M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics, 24: 2431-2437.
URL http://www.hubmed.org/display.cgi?uids=18684737


Malhis, N, Butterfield, Y S, Ester, M, Jones, S J (2009) Slider–maximum use of probability information for alignment of short sequence reads and SNP detection. Bioinformatics, 25: 6-13.
URL http://www.hubmed.org/display.cgi?uids=18974170


Medini, D, Serruto, D, Parkhill, J, Relman, D A, Donati, C, Moxon, R, Falkow, S, Rappuoli, R (2008) Microbiology in the post-genomic era. Nat Rev Microbiol, 6: 419-430.
URL http://www.hubmed.org/display.cgi?uids=18475305


Smith, A D, Xuan, Z, Zhang, M Q (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics, 9: 128-128.
URL http://www.hubmed.org/display.cgi?uids=18307793