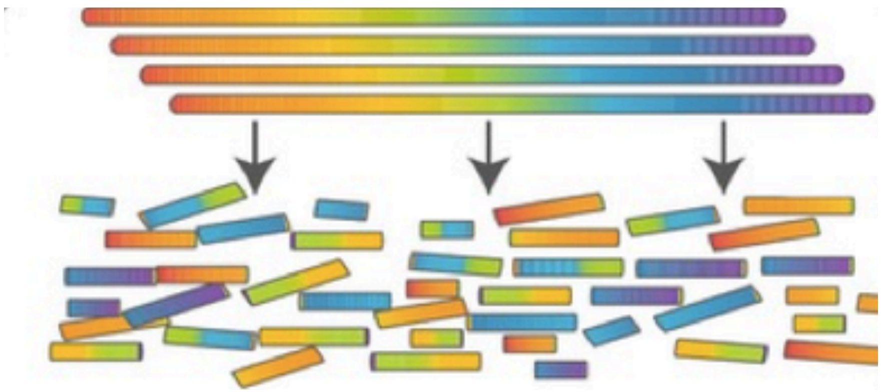
Applications of DNA sequencing

Overview

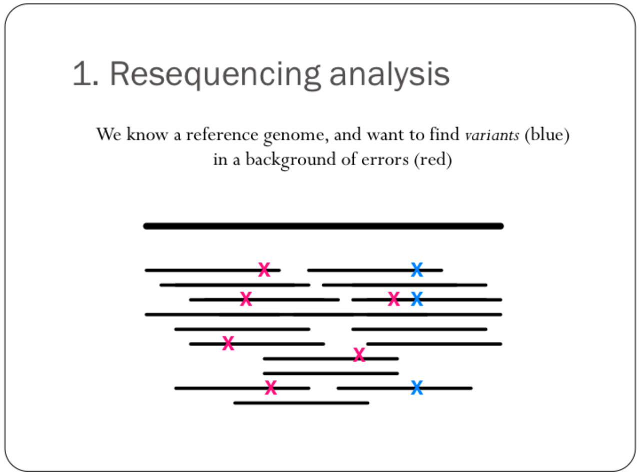
- DNA Sequencing Applications
- Reference Genome availability
- Quality Assessment and Trimming of DNA sequence reads

Whole Genome Shotgun Sequencing

- Start with genomic DNA
- DNA is sheared into fragments
  - Physical
    - Acoustic shearing (Covaris)
    - Sonication (Bioruptor)
    - Hydrodynamic force (Hydroshear)
  - Enzymatic (transposase, DNase I)
  - Chemical
- Ideally, would like a very uniform size selection
  - paired end: depends on kit, from 200-600bp
  - mate pairs: 3-20 Kbp

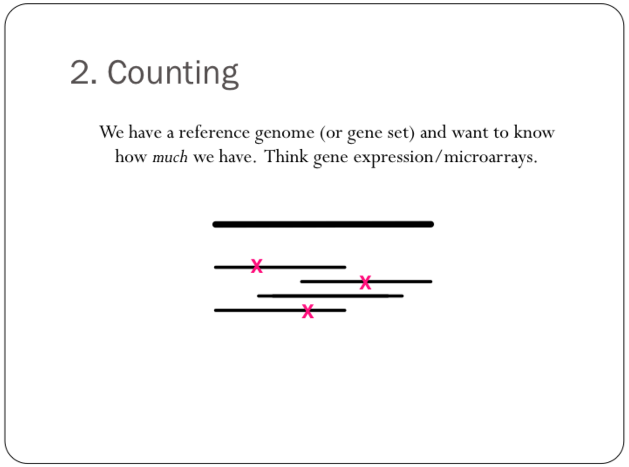


Many Uses



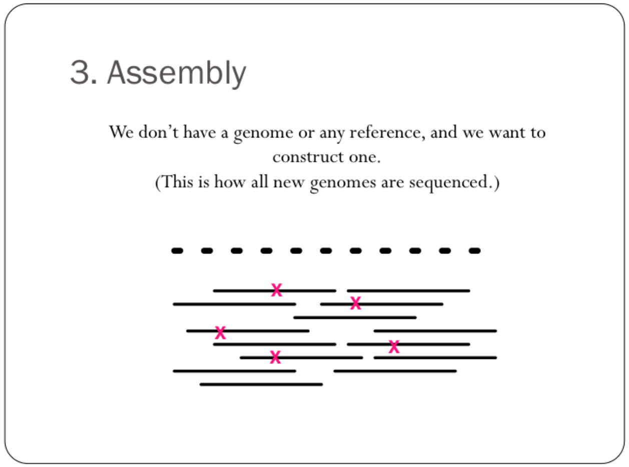
1. Resequencing analysis

We know a reference genome, and want to find *variants* (blue) in a background of errors (red)



2. Counting

We have a reference genome (or gene set) and want to know how *much* we have. Think gene expression/microarrays.



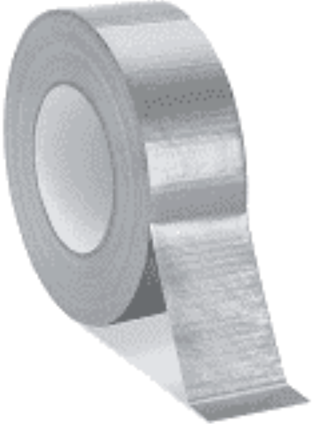
3. Assembly

We don't have a genome or any reference, and we want to construct one.
(This is how all new genomes are sequenced.)

The basic approaches to using sequencing data.

So Many Uses

Biologists are clever and have hijacked shotgun sequencing for many types of new biological assays, mostly having to do with clever ways of treating the DNA and preparing the libraries.



- RNASeq
  - convert RNA to cDNA for sequencing
  - we'll cover this more later
- Targeted DNASeq
  - only sequence regions of interest
- ChIP Sequencing
  - Sequence genome-wide binding sites
- Reduced complexity sequencing (GBS/RADSeq) for genotyping
  - sequence a smaller part of the genome
- Methylation sequencing/ bisulfiite seq
  - Identify DNA bases that have methylation
- Hi-C
  - 3D Chromatin structure
- Etc.

Coverage (Depth of Coverage)

- All applications depend on appropriate coverage
- Depth of Coverage = the average number of reads that align to or “cover” a reference base

- Mean mapped read depth =

# of mapped bases

# of reference bases

- Often displayed as a histogram
- Coverage uniformity matters too

Which SNP is believable?

- Just by eye (there are real statistics to this that we'll learn more about later)

```
Ref: GAAGTGGCATATGGCTGTGAAGAAAAAG

R1:  GAACTGGCATATGGCTGTGA
R2:          ATATGGCAGTGAAGAAA
```

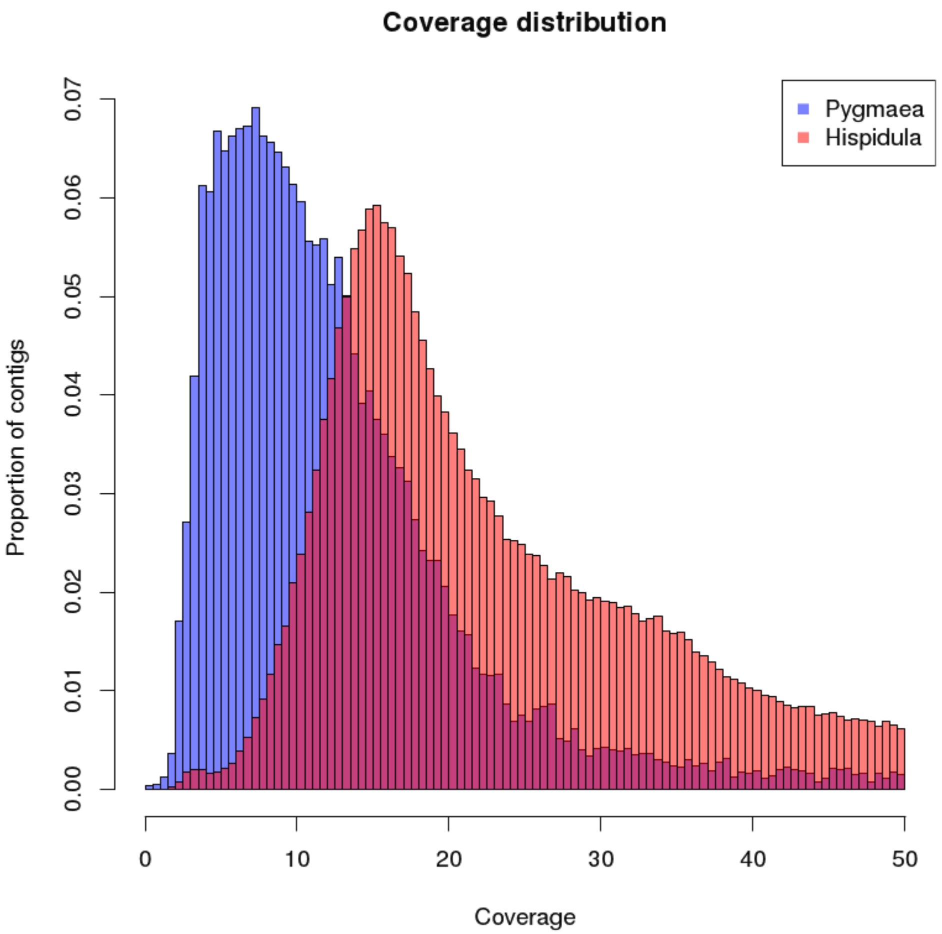
Which SNP is believable?

- Just by eye (there are real statistics to this that we’ll learn more about later)

```
Ref: GAAGTGGCATATGGCTGTGAAGAAAAAG

R1:  GAACTGGCATATGGCAGTGA
R2:   AACTGGCATATGGCAG
R3:          ATATGGCAGTGAAGAAA
R4:            ATGGCAGTGAAGAAAA
R5:               GCAGTGAAGAAAAAG
```

Depth of Coverage as a Histogram

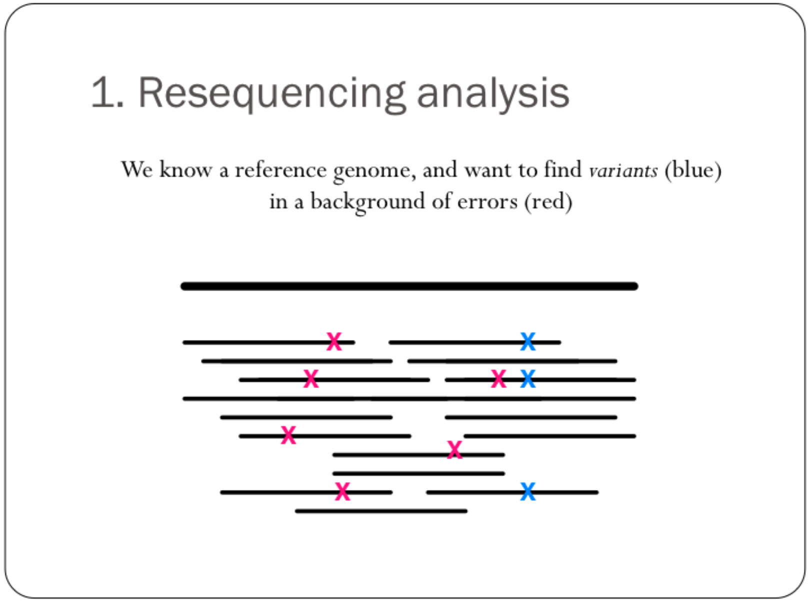


https://www.biostars.org/p/48633/

What do you think causes the long tail?

App 1: Whole Genome Resequencing

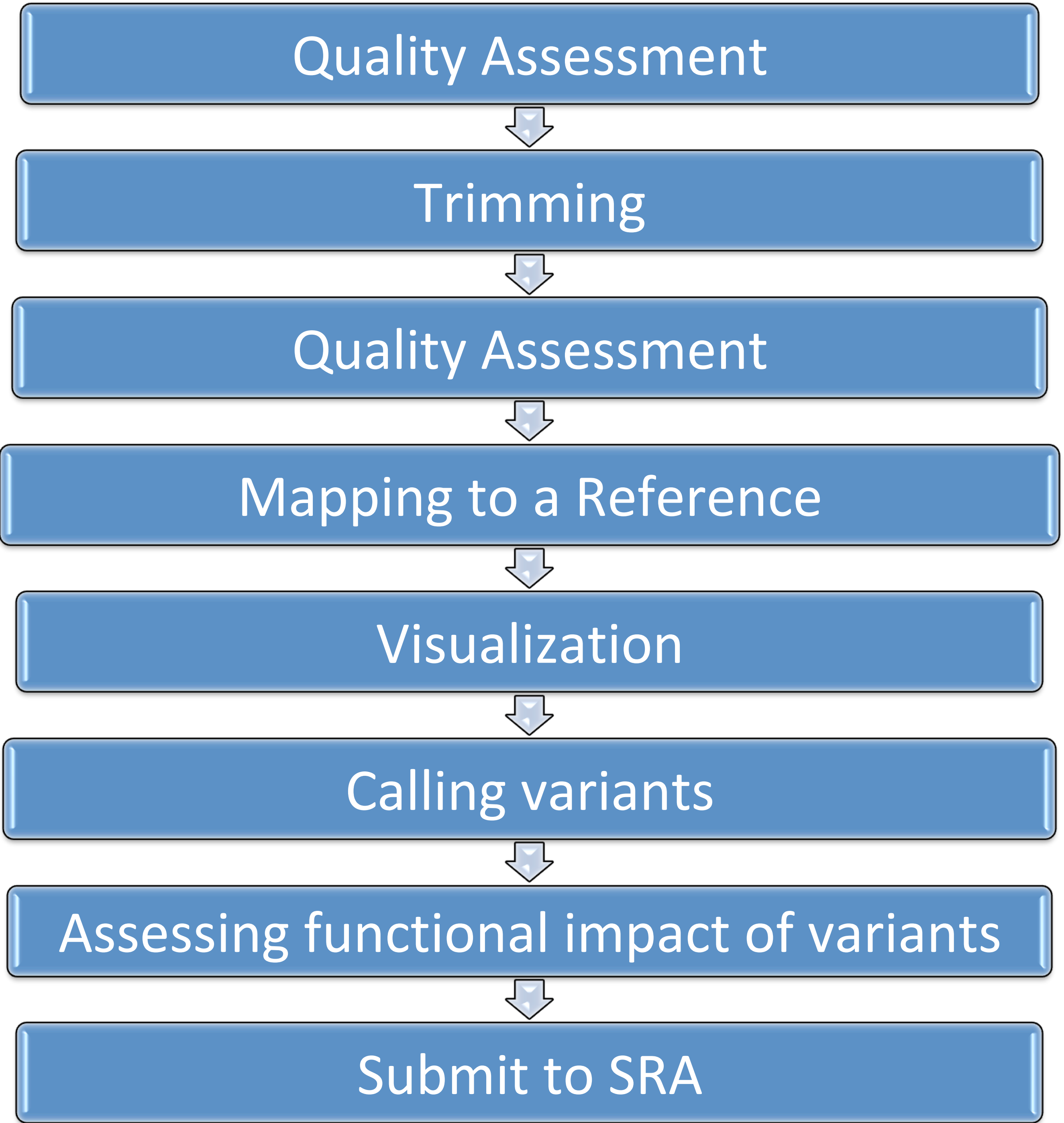
- Reference genome is already available
- Sequencing one or more individuals, usually from the same species
- Discover variations in the genomes between and within samples
  - Straightforward
    - SNPs
    - insertions
    - deletions
  - Harder
    - rearrangements
    - copy number changes



1. Resequencing analysis

We know a reference genome, and want to find *variants* (blue) in a background of errors (red)

Example Workflow (Pipeline)



Quality Assessment

Trimming

Quality Assessment

Mapping to a Reference

Visualization

Calling variants

Assessing functional impact of variants

Submit to SRA

App 2: GBS/RADSeq

Polymorphism discovery

Polymorphisms are an essential genomic tool for:

- Population Structure
- Association mapping
- Pedigree mapping
- QTL mapping
- Phylogeny

- Use the high volume and low cost of sequencing to replace SNP chips and microsatellites
- Whole genome resequencing is expensive!
- How to efficiently use NGS for discovering markers?
- How to efficiently use NGS to do the genotyping?

Restriction Site-associated DNA sequencing (RADSeq)

- Random genome reduction
- Developed Baird et al 2008
- Identify and score thousands of genetic markers
- Subsampling only at specific sites defined by restriction enzyme

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, et al. (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLoS ONE 3(10): e3376.

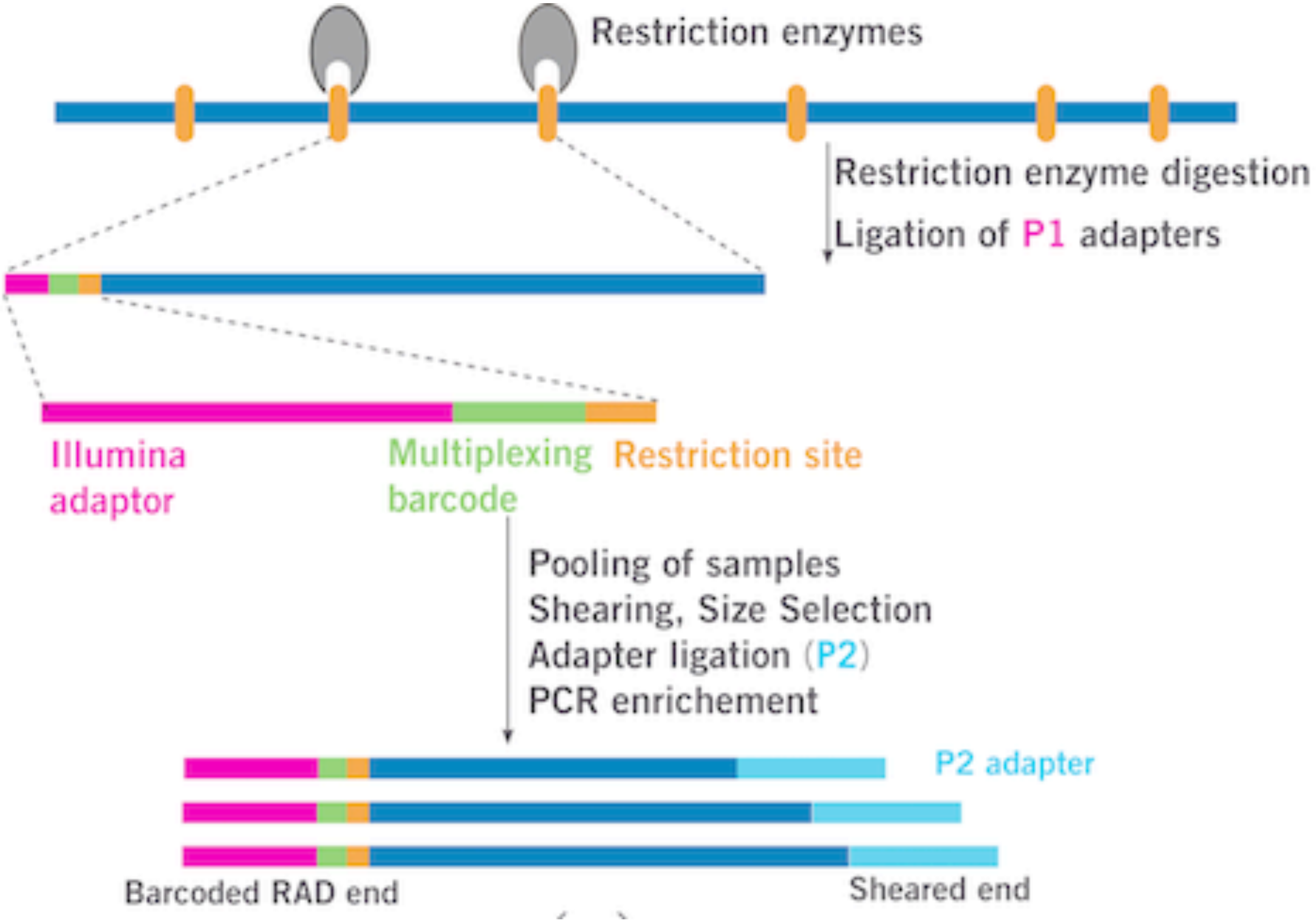
- Randomly distributed across the target genome
- From many individuals using Illumina technology

Patented by University of Oregon, licensed by:

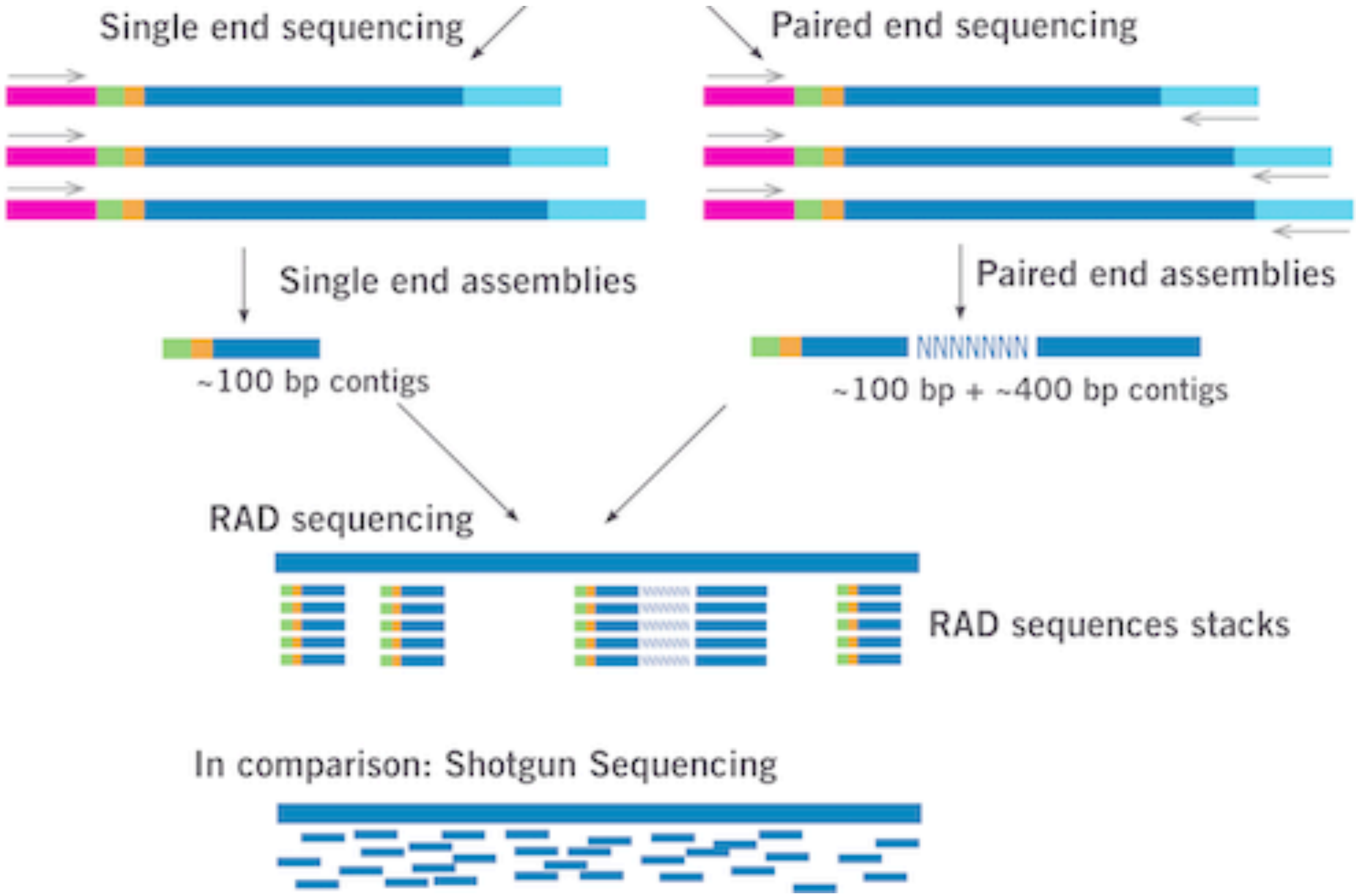
KeyGene

FLORAGENEX

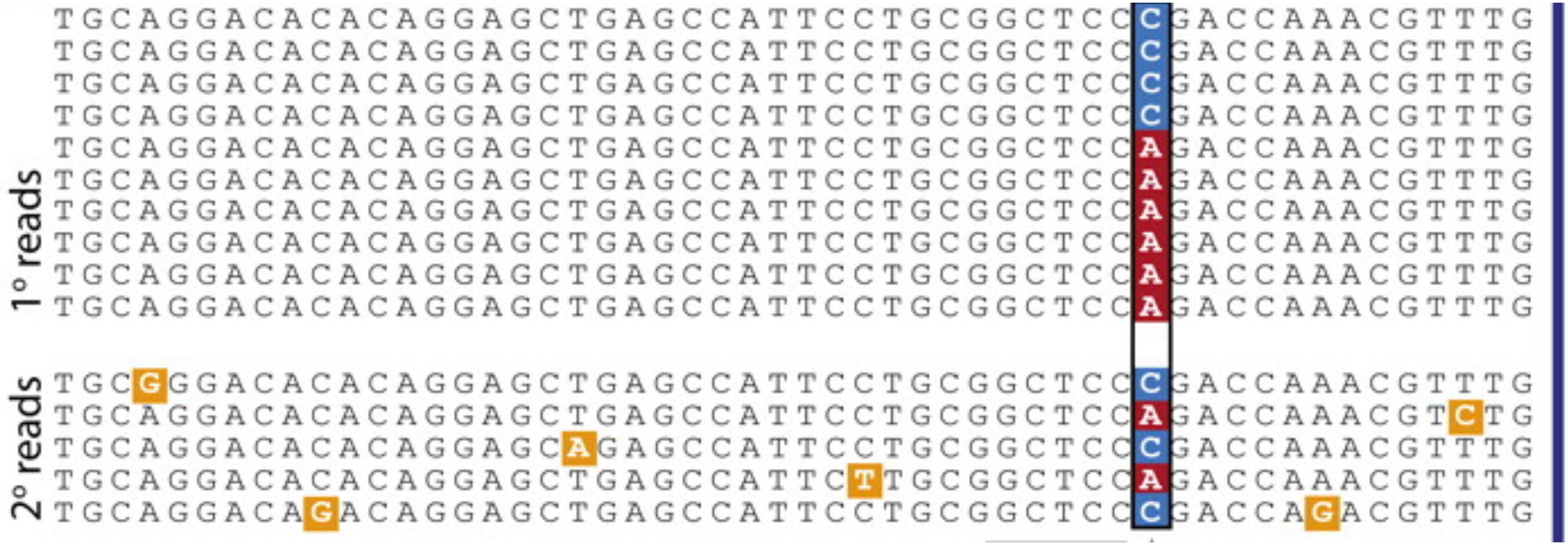
Non profit use at universities can be licensed for free



http://www.floragenex.com/rad-seq/



http://www.floragenex.com/rad-seq/


TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTTTG
1° reads
TGCGGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCCTGCGGCTCCAGACCAAACGTCTG
TGCAGGACACACAGGAGCAGAGCCATTCCTGCGGCTCCCGACCAAACGTTTG
TGCAGGACACACAGGAGCTGAGCCATTCTTGCGGCTCCAGACCAAACGTTTG
TGCAGGACAGACAGGAGCTGAGCCATTCCTGCGGCTCCCGACCAGACGTTTG
2° reads

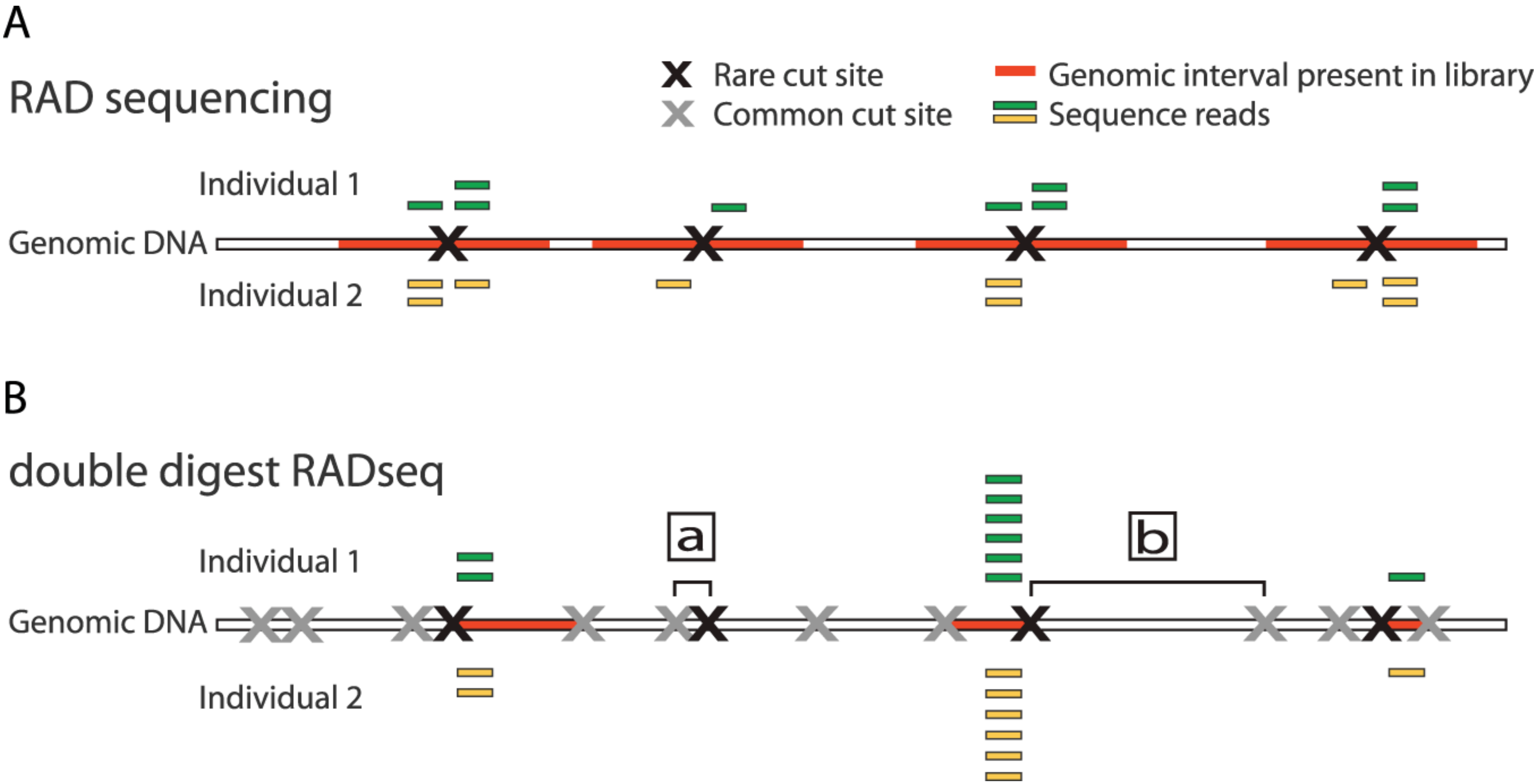
Julian M. Catchen et al. G3 2011;1:171-182

ddRAD

- May have the problem of too many sites across the genome (even if you use a rare cutter)
- Need a way to more accurately control the number of loci sequenced
- Double digest RAD
- Peterson et al 2012
- Also patented

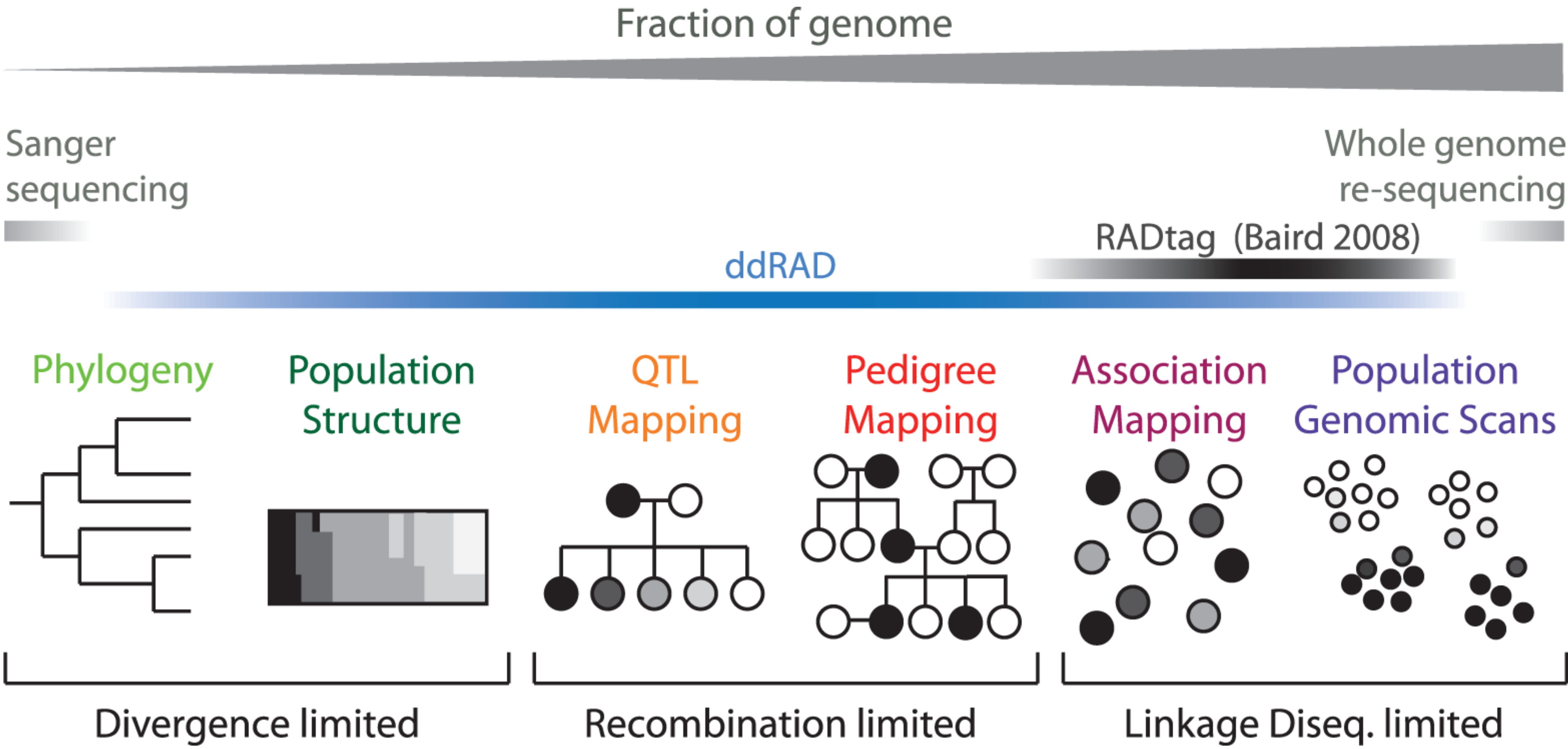
ddRAD

- Double digest RAD
- (Peterson et al 2012)
- Simpler and cheaper library construction
- restriction digest with two enzymes simultaneously
- eliminate random shearing and end repair
- explicitly use size selection
- Sequence fragments generated by cuts with both REs and which fall within the size-selection window



Peterson et al (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135.

Its important to carefully select the technology that meets your goals.

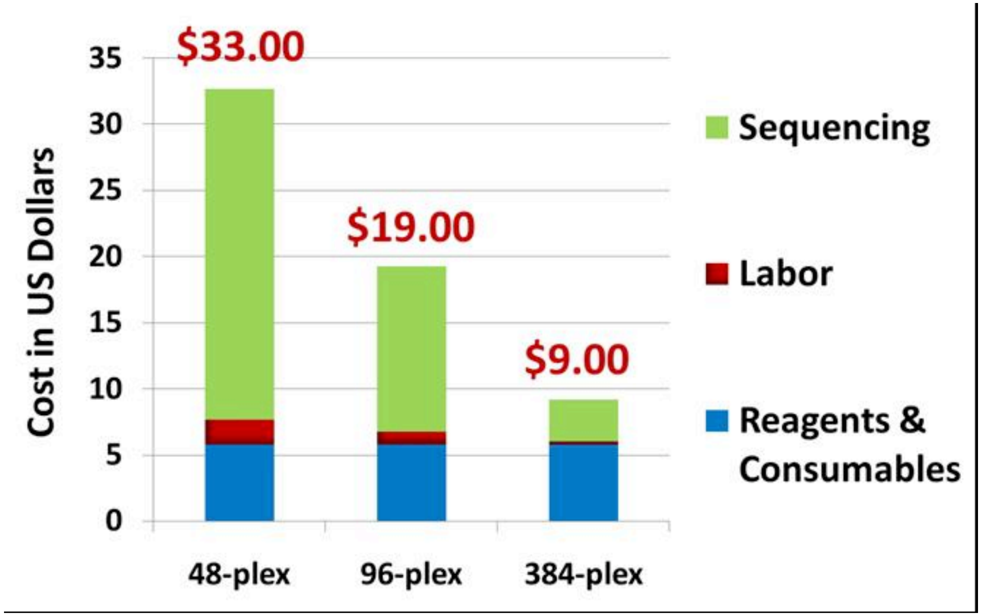


Peterson et al (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135.

GBS: Genotyping by Sequencing

- Elshire et al 2011
- Increased efficiency and cost benefits
- Reduced sample handling
- Methylation-sensitive REs used to filter out the repetitive portion of the genome
- Better barcoding system
- Fewer steps
- Free to use and to sell (ie not licensed)

*In 2015 I paid $46 per sample
96-plex, 1 plate
Including SNP calling
Including enzyme optimization
External rate at Cornell



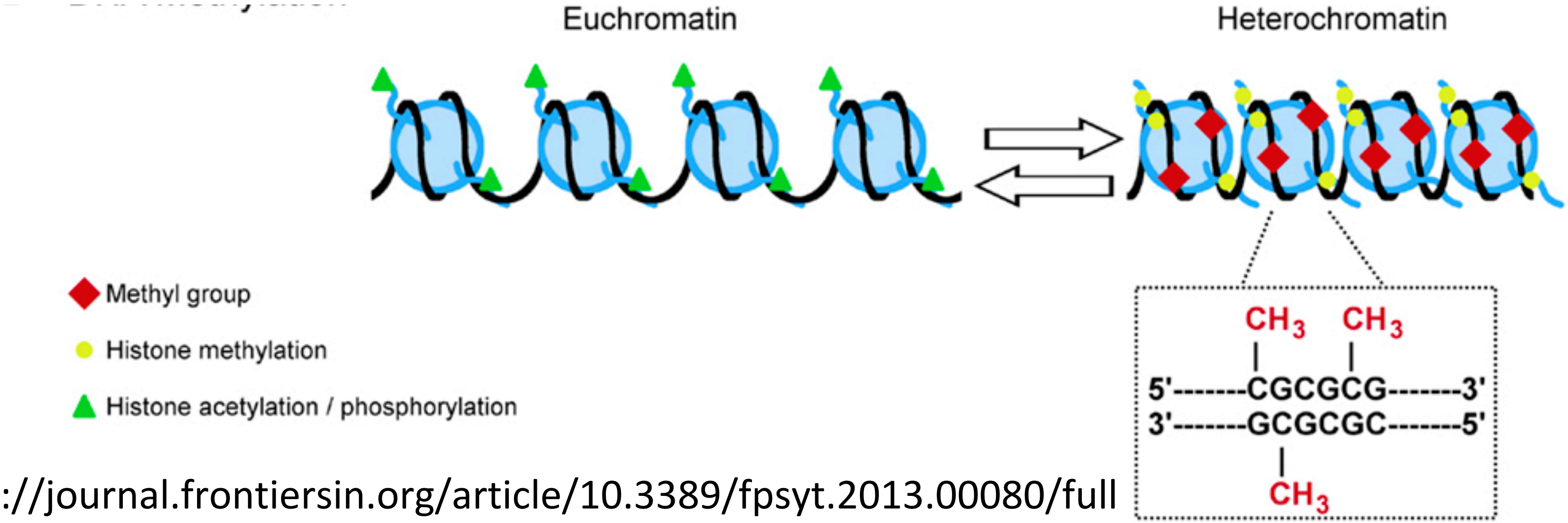
http://www.maizegenetics.net/#!genotyping-by-sequencing-gbs/c9c6

Workflow

- Informatics can be difficult if you don’t have a genome
- Much easier if you have a reference genome – can be similar to whole genome resequencing
- If not, some software packages:
  - Universal Network Enabled Analysis Kit (UNEAK) – part of the TASSEL software suite
  - Stacks

App 3: Bisulfite Sequencing

- DNA methylation
  - First discovered epigenetic mark
  - methyl groups are added to DNA
  - Suppresses transcription when present in promoter
  - Adenine and Cytosine can be methylated in prokaryotes
  - Only cytosine is methylated in eukaryotes



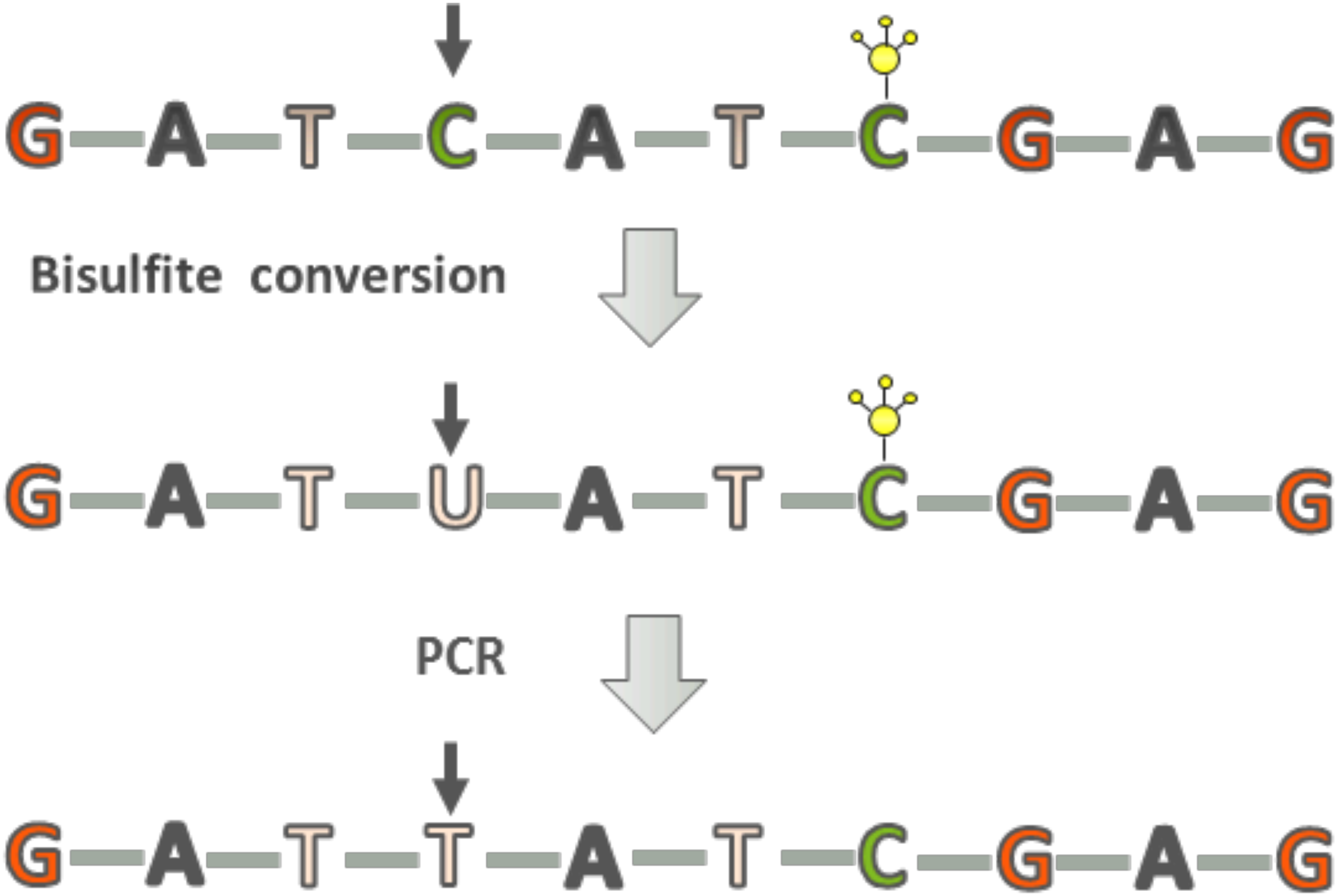
http://journal.frontiersin.org/article/10.3389/fpsyt.2013.00080/full

Bisulfite treatment

- How to figure out where methylation occurs while sequencing?
- Treatment of DNA with bisulphite:
  - Unmethylated cytosine -> uracil
  - 5-methylcytosine stays the same
- Sequencing can yield single- nucleotide resolution of methylation patterns

Problems:

- Incomplete conversion
- DNA degradation during conversion
- 5-methylcytosine and 5-hydroxymethylcytosine both read as a C in bisulphite sequencing



http://www.gatc-biotech.com/en/products/inview-applications/inview-epigenome.html

Workflow

- Need special software for mapping
  - Bismark
  - BSMap
  - BSMapper
- Downstream analylsis
  - Methylkit – statistics, visualization, tiling windows

Overview

- Many types of omics yield different types of information
- Your data analysis protocols must be tailored to fit the data production protocols

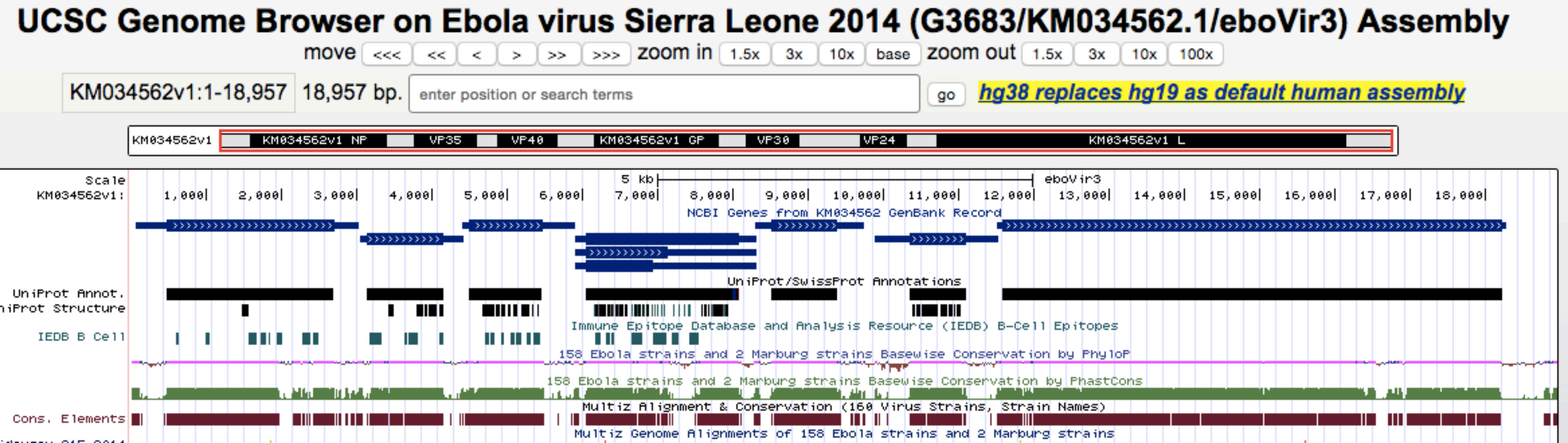
Reference Genomes

All the methods we talked about so far depend on (or are easier with) a reference genome.

How many genomes are out there to use as a base for mapping reads?

UCSC Genome Browser

- Mammal (55)
- Other Vertebrate (27)
- Deuterostome (3)
- Insect (13)
- Nematode (6)
- Other (2) – yeast and sea hare
- Viruses (1) - Ebola

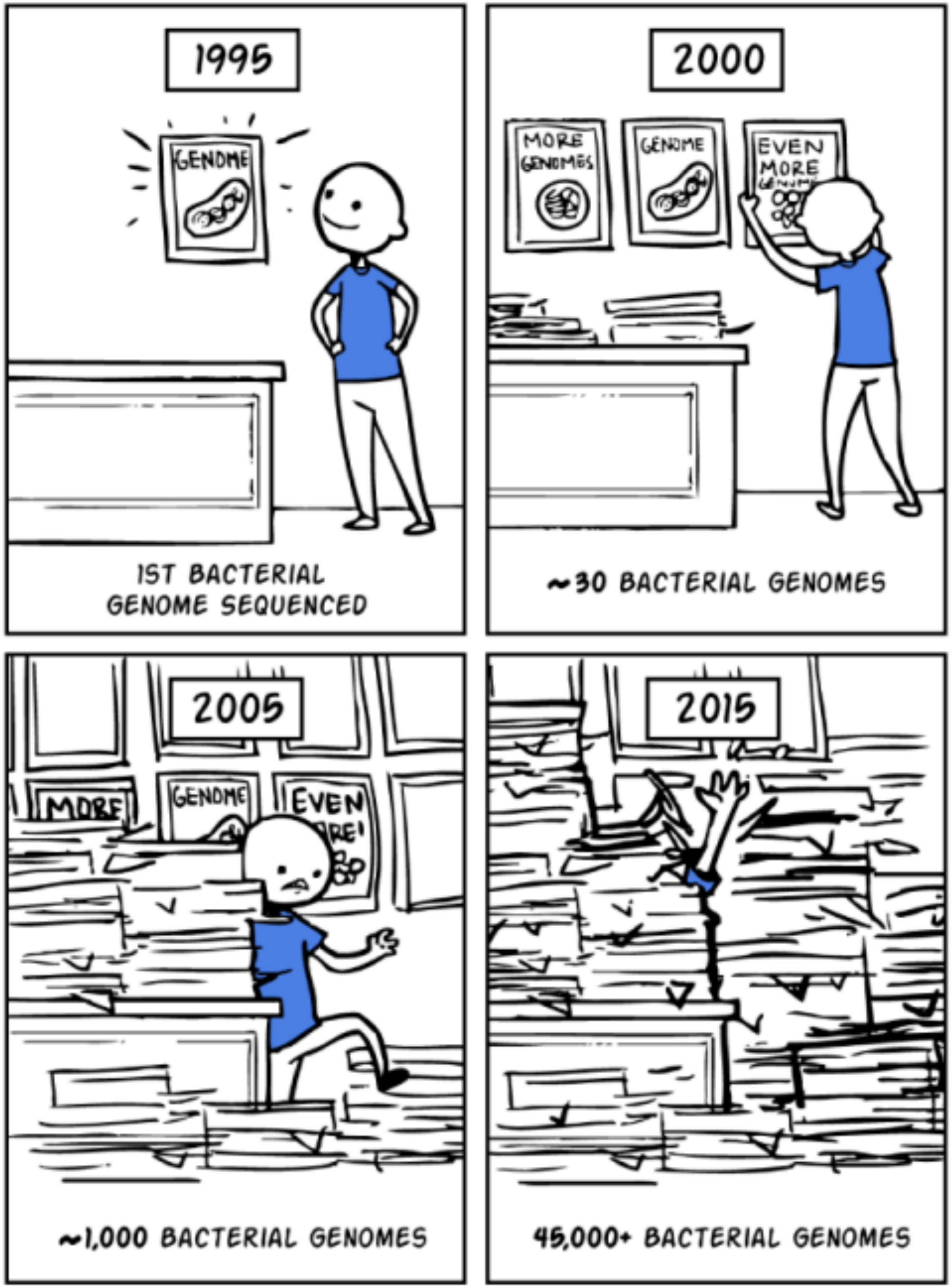


UCSC Genome Browser on Ebola virus Sierra Leone 2014 (G3683/KM034562.1/eboVir3) Assembly

KM034562v1:1-18,957 18,957 bp.

hg38 replaces hg19 as default human assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
enter position or search terms
go
KM034562v1
KM034562v1 NP
VP35
VP40
KM034562v1 GP
VP30
VP24
KM034562v1 L
Scale
KM034562v1:
5 kb
eboVir3
1,000
2,000
3,000
4,000
5,000
6,000
7,000
8,000
9,000
10,000
11,000
12,000
13,000
14,000
15,000
16,000
17,000
18,000
NCBI Genes from KM034562 GenBank Record
UniProt/SwissProt Annotations
UniProt Annot.
niProt Structure
Immune Epitope Database and Analysis Resource (IEDB) B-Cell Epitopes
IEDB B Cell
158 Ebola strains and 2 Marburg strains Basewise Conservation by PhyloP
158 Ebola strains and 2 Marburg strains Basewise Conservation by PhastCons
Multiz Alignment & Conservation (160 Virus Strains, Strain Names)
Cons. Elements
Multiz Genome Alignments of 158 Ebola strains and 2 Marburg strains



TAKEHOMEMESSAGE.COM

Reference Genomes

NCBI Genome

NCBI Genome has 4 levels:

- Complete - all chromosomes are gapless and have no runs of 10 or more ambiguous bases (Ns), there are no unplaced or unlocalized scaffolds, and all the expected chromosomes are present
- Chromosome - there is sequence for one or more chromosomes, gaps OK.
- Scaffold - some sequence contigs have been connected across gaps to create scaffolds, but the scaffolds are all unplaced or unlocalized
- Contig - nothing is assembled beyond the level of sequence contigs

NCBI Genome Records

- Viruses
  - 19,743 genomes
  - 19,084 complete (> 96%)
- Prokaryotes
  - 162,676 genomes
  - 11,576 complete (7.0%)
  - 2,057 chromosome level (1.3%)
- Eukaryotes
  - 6,563 genomes
  - 47 complete (< 1%)
  - 874 chromosome level (13.3%)

http://www.ncbi.nlm.nih.gov/genome/browse/#

What are we doing today?

1. Examining Read Quality
2. Quality/Adapter Trimming

Quality Control

Goals

- Is my data of sufficient quality to use?
- The instrument assigns a confidence value to each base. Are the bases high quality overall?
- Does the complexity look normal?
  - PCR and library prep problems can lead to duplication of the same sequences over and over
- Are there adapters or other over-represented sequences?
- Are there lane batch effects?

FASTQC

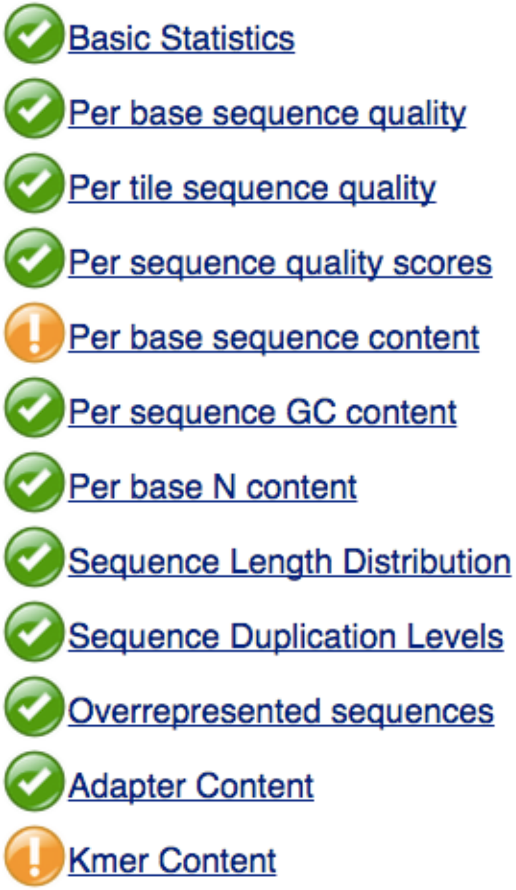
Babraham Bioinformatics

Accepts input formats:

- FastQ (all quality encoding variants)
- GZip compressed FastQ
- SAM
- BAM

Does a 12 point analysis of quality

Generates an html output file

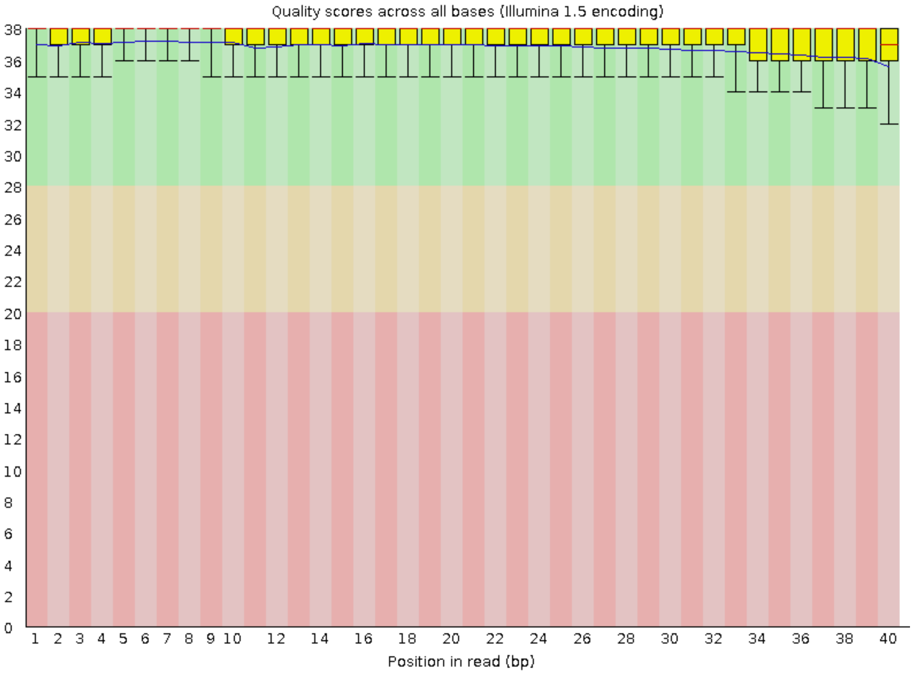


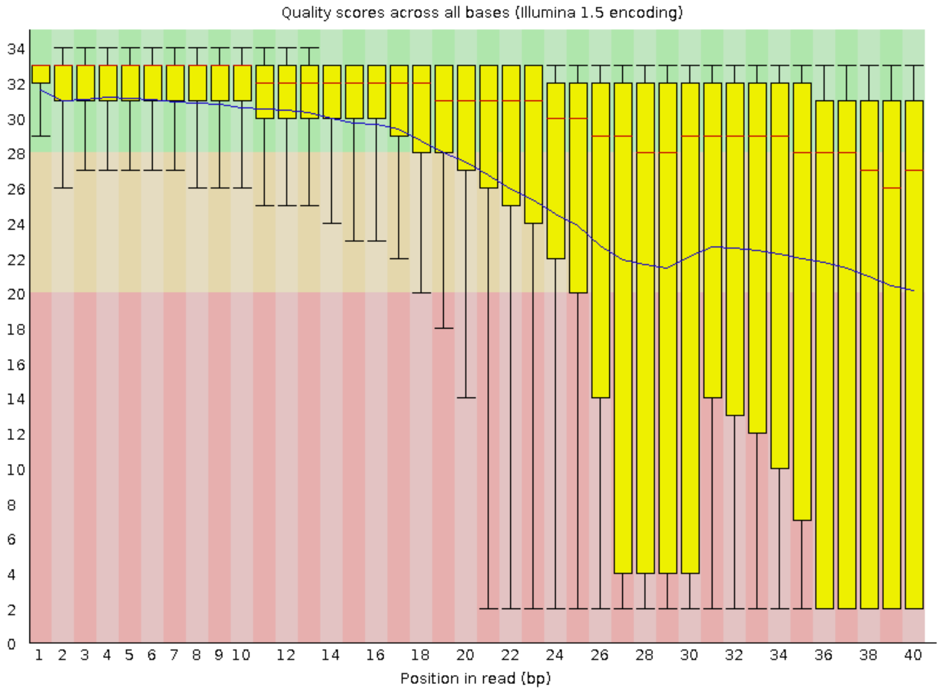
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC

Babraham Bioinformatics


Quality scores across all bases (Illumina 1.5 encoding)


Quality scores across all bases (Illumina 1.5 encoding)

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

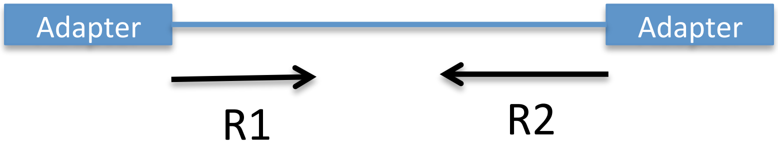
Trimming

- From the quality control step, we know where the problems are
- All illumina reads tend to have degrading quality at the end of the read
- Get rid of the bad data, keep the good data
  - Cut adapter sequences from the read.
  - Trim off low quality bases
  - Drop a read entirely if is too low quality or too short

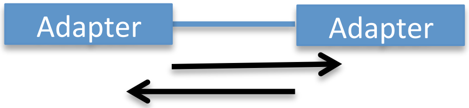
Trimmomatic

- Optimized for Illumina NGS
- Very flexible
- Handles paired end data well
- Threaded

- Detects adapter read through
- No read through:



- Read through:



Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, btu170.

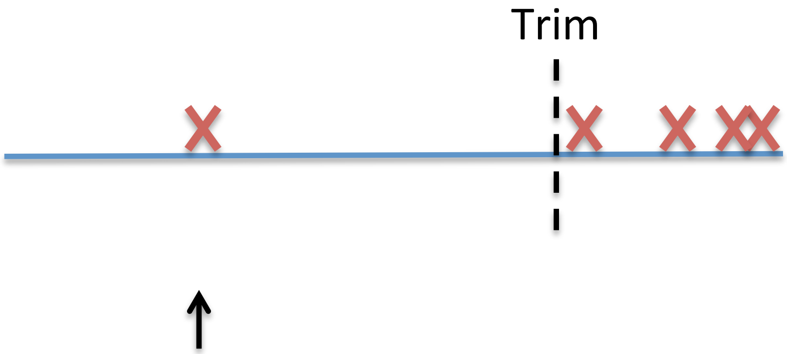
Trimmomatic

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read. Must specify adapter sequence. Comes with basic Illumina adapter files, make sure yours are in there or add yours!
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length after trimming

Trimmomatic

- Maximum Information Quality Filtering:
- Retain low-quality bases early in a read in order to make sure the read is sufficiently long to be informative
- Trimming process becomes more strict later in the read



Leave this one bad base so we don’t lose the usefulness of the read

Skewer

- Faster than trimmomatic
- A bit less flexible
- “Gentle” quality trimming by default

https://github.com/relipmoc/skewer

Jiang, H., Lei, R., Ding, S.W. and Zhu, S. (2014) Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. BMC Bioinformatics, 15, 182.

Trimming

- Current community wisdom:
  - Quality trimming reduces error
  - But also reduces content and contiguity
- Gentle trimming is preferred – many times the defaults are too stringent, you will lose lots of data!
- Application matters
  - For mapping and counting, gentle to no trimming (phred 3 to 5)
  - For assembly and variant calling, a bit more trimming is good (phred 10 to 15)