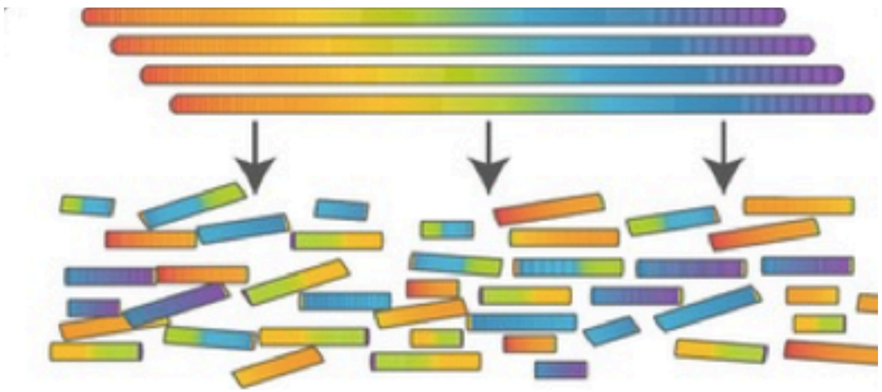
Applications of DNA sequencing

Overview

- DNA Sequencing Applications
  1. Resequencing
  2. Capture/Targeted Resequencing
  3. RAD/ddRAD/GBS
  4. Bisulfite Sequencing
  5. De novo assembly
- Reference Genome availability
- N50

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



App 1: Whole Genome Resequencing

- Sequencing multiple individuals from the same species
- Reference genome is already available
- Discover variations in the genomes between and within samples
  - mutations
  - insertions
  - deletions
  - rearrangements
  - copy number changes

How long do the reads need to be?

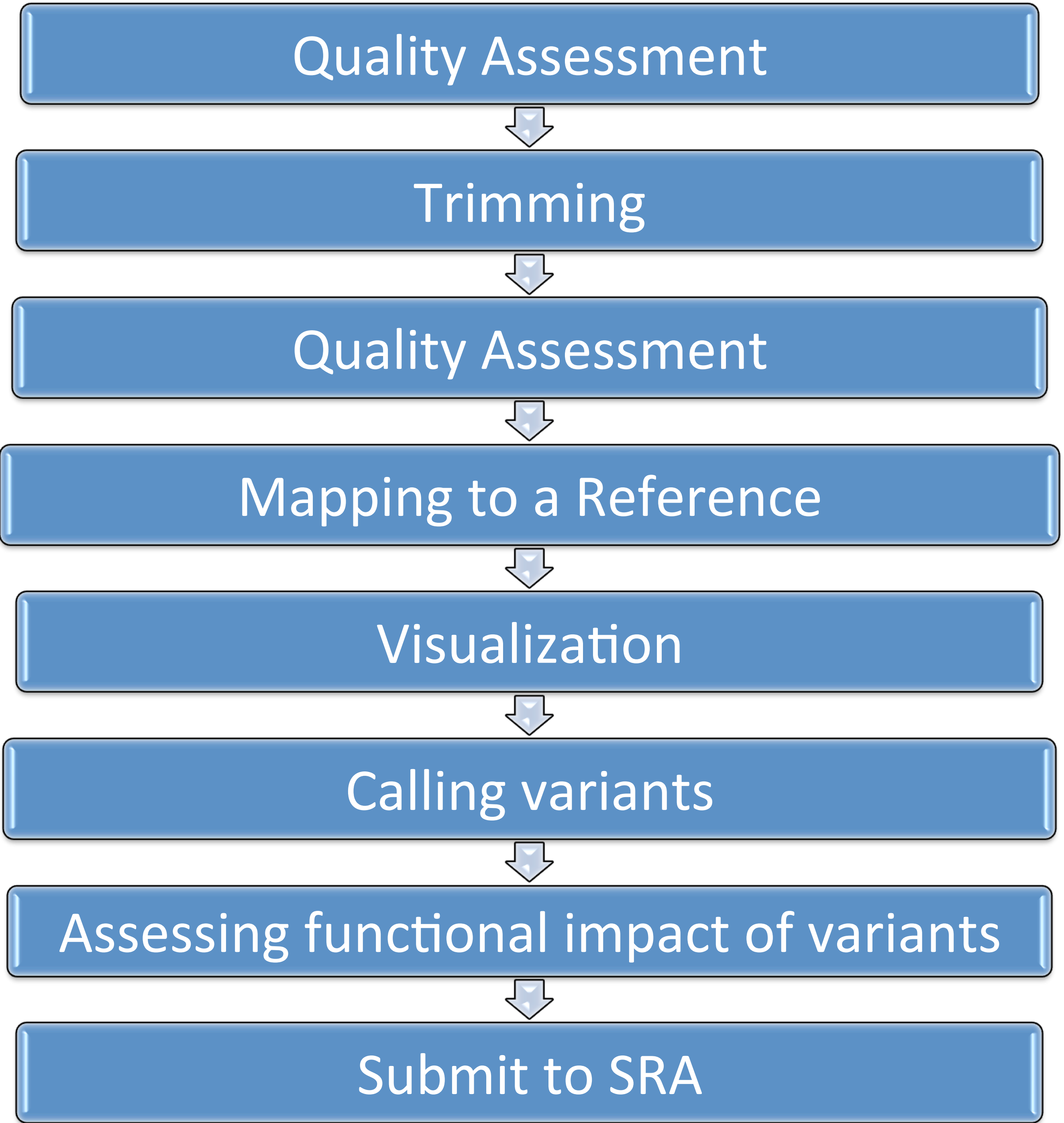
For the human genome, estimates are:

25mers = 80% unique coverage

43mers = 90% unique coverage

But longer is better for repeats, more complex genomes.

Workflow



Quality Assessment

Trimming

Quality Assessment

Mapping to a Reference

Visualization

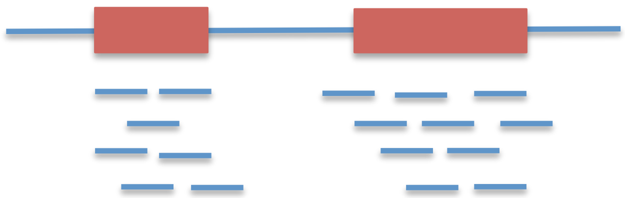
Calling variants

Assessing functional impact of variants

Submit to SRA

App 2: Sequence Capture/Targeted Resequencing

- Reduce costs by only sequencing regions of interest
- Increase multiplexing while maintaining deep coverage
- Different target enrichment strategies
  - PCR amplification of many regions (QIAGEN)
  - Array-based Hybridization (NimbleGen)
  - In solution Hybridization* (Agilent and NimbleGen)



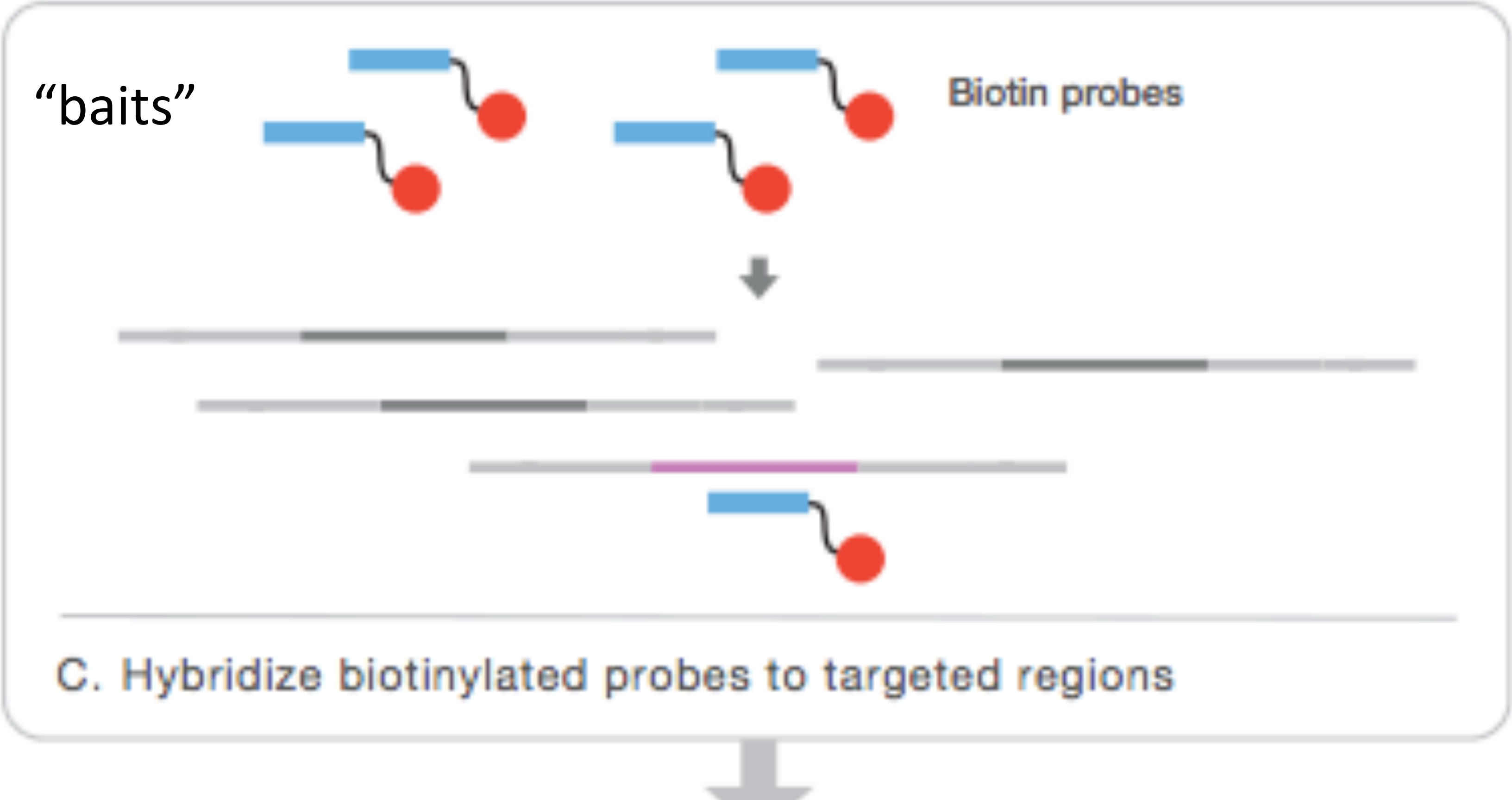
- Most common use is exome capture.
- Can also use to capture exomes plus UTRs and miRNAS
- Can also target large continuous regions, such as QTLs (quantitative trait loci)

Gnirke et al., 2009, Solution Hybrid Selection with Ultra-long Oligonucleotides for Massively Parallel Targeted Sequencing. Nat Biotech.

Exome sequencing with Human Nextera Rapid Capture Exomes Kit

- >340,000 95mer probes
- Genomic footprint of 62Mb (~2% of the genome)

Start with a normal DNA library.

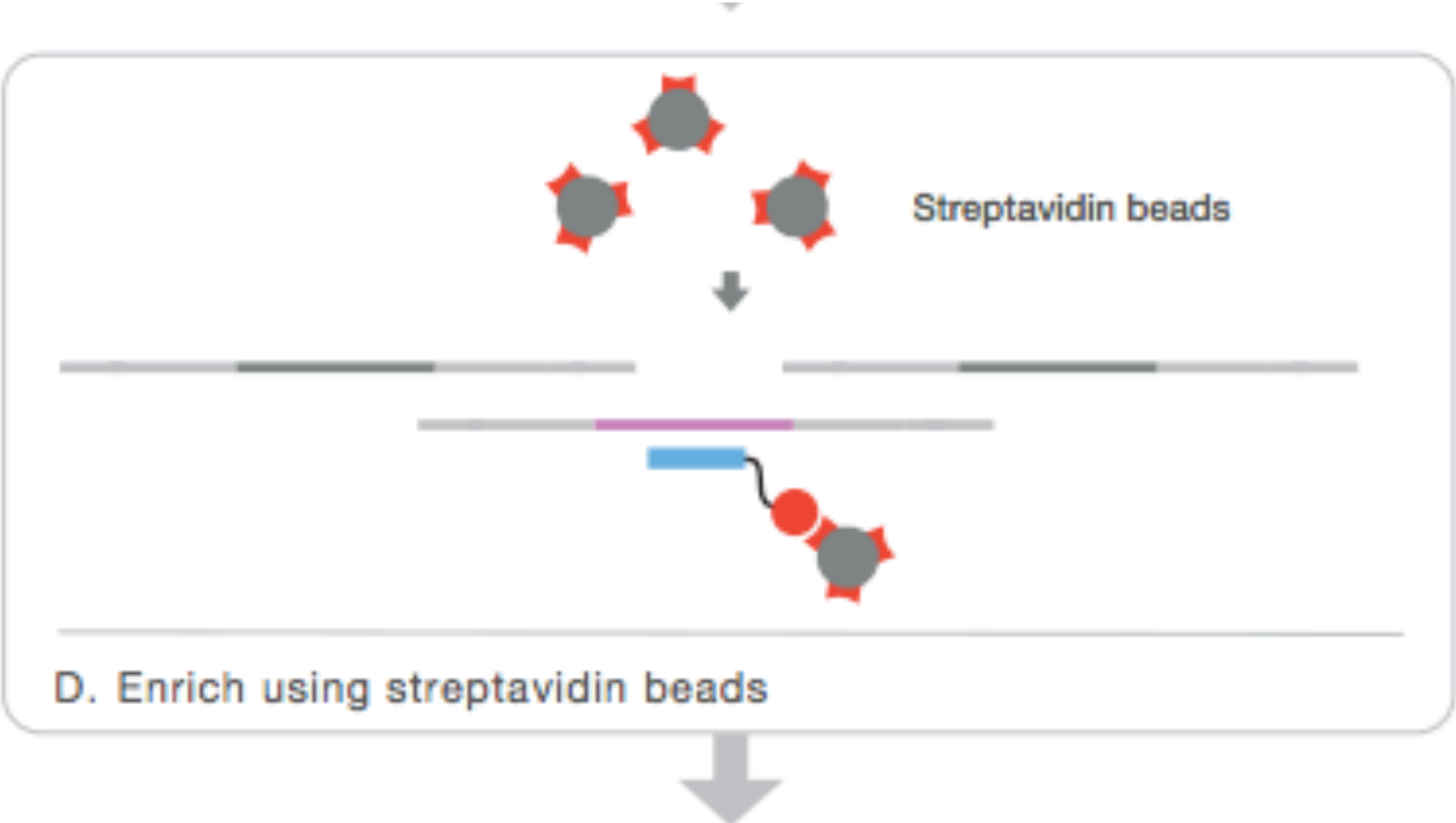


“baits”

C. Hybridize biotinylated probes to targeted regions

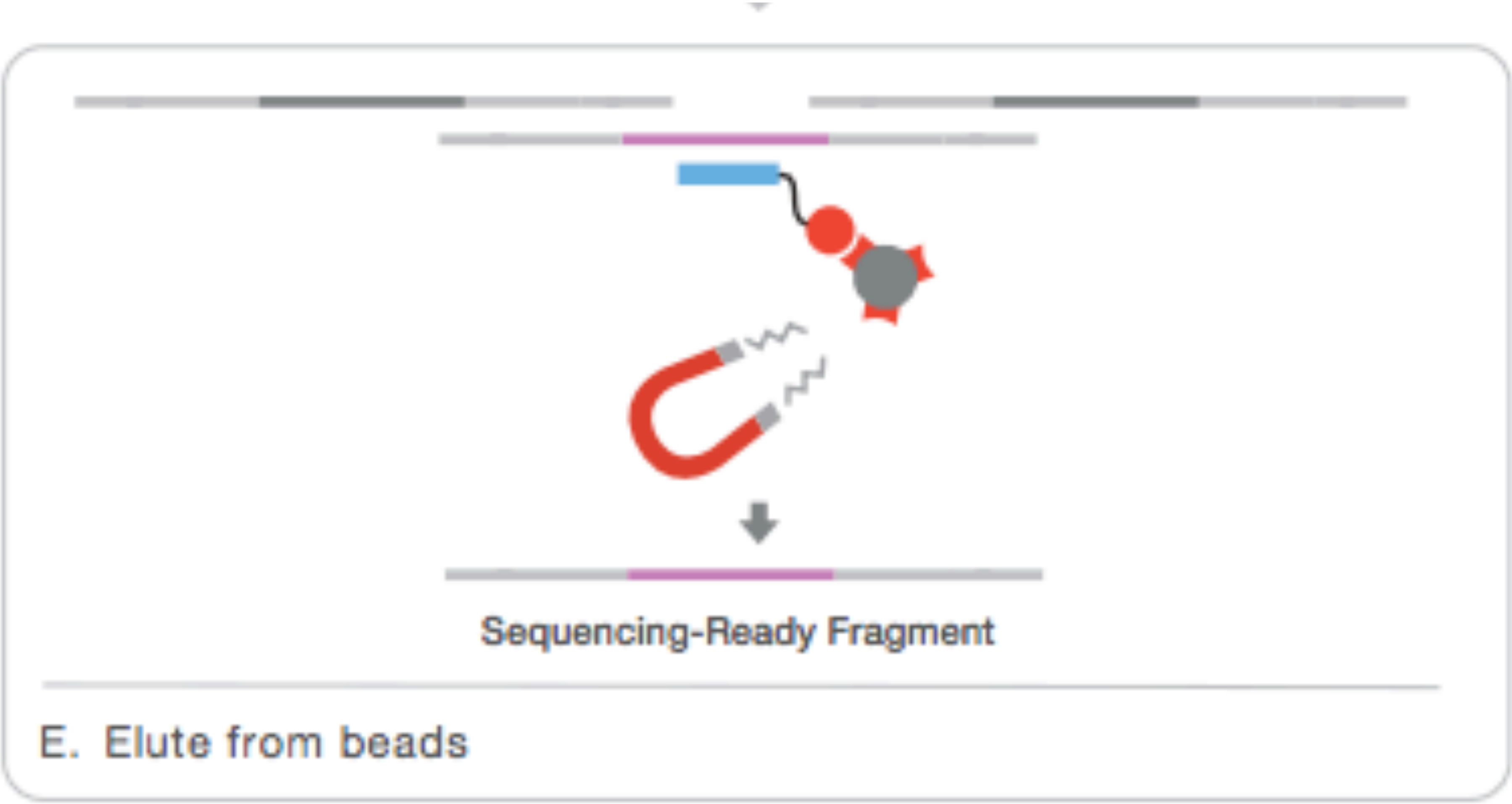
Exome sequencing with Human Nextera Rapid Capture Exomes Kit

- >340,000 95mer probes
- Genomic footprint of 62Mb (~2% of the genome)



Exome sequencing with Human Nextera Rapid Capture Exomes Kit

- >340,000 95mer probes
- Genomic footprint of 62Mb (~2% of the genome)

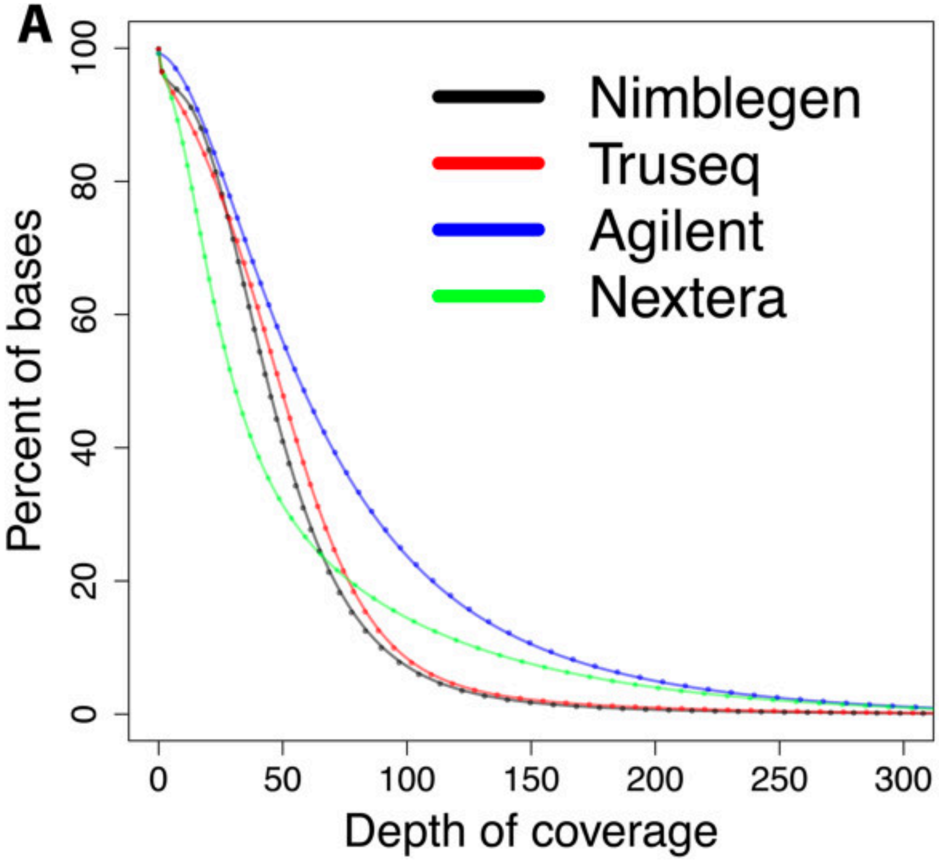


Problems

- Off target fragments get sequenced
- Coverage is not uniform
- Genetic heterogeneity – do the probes work in all individuals?

Current human kits:

- 80% of bases are covered at 10X with 4Gb of sequence



Performance comparison of four exome capture systems for deep sequencing. Chilamakuri et al 2014.

“For all the technologies, 25 million reads were sufficient to cover about 80% of target bases with at least 10× depth, with the exception of the Nextera technology, which covered only about 60% of target bases with the same number of reads”

Workflow

- Essentially the same as a whole genome resequence project
- Depending on variants of interest, may need to assess where coverage failed

App 3: GBS/RADSeq

Polymorphic marker- 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
- How to efficiently use NGS for discovering markers?
- How to efficiently use NGS to do the genotyping?

Restriction Site-associated DNA sequencing (RADSeq)

- Developed Baird et al 2008
- Identify and score thousands of genetic markers
- Randomly distributed across the target genome
- From many individuals using Illumina technology

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.

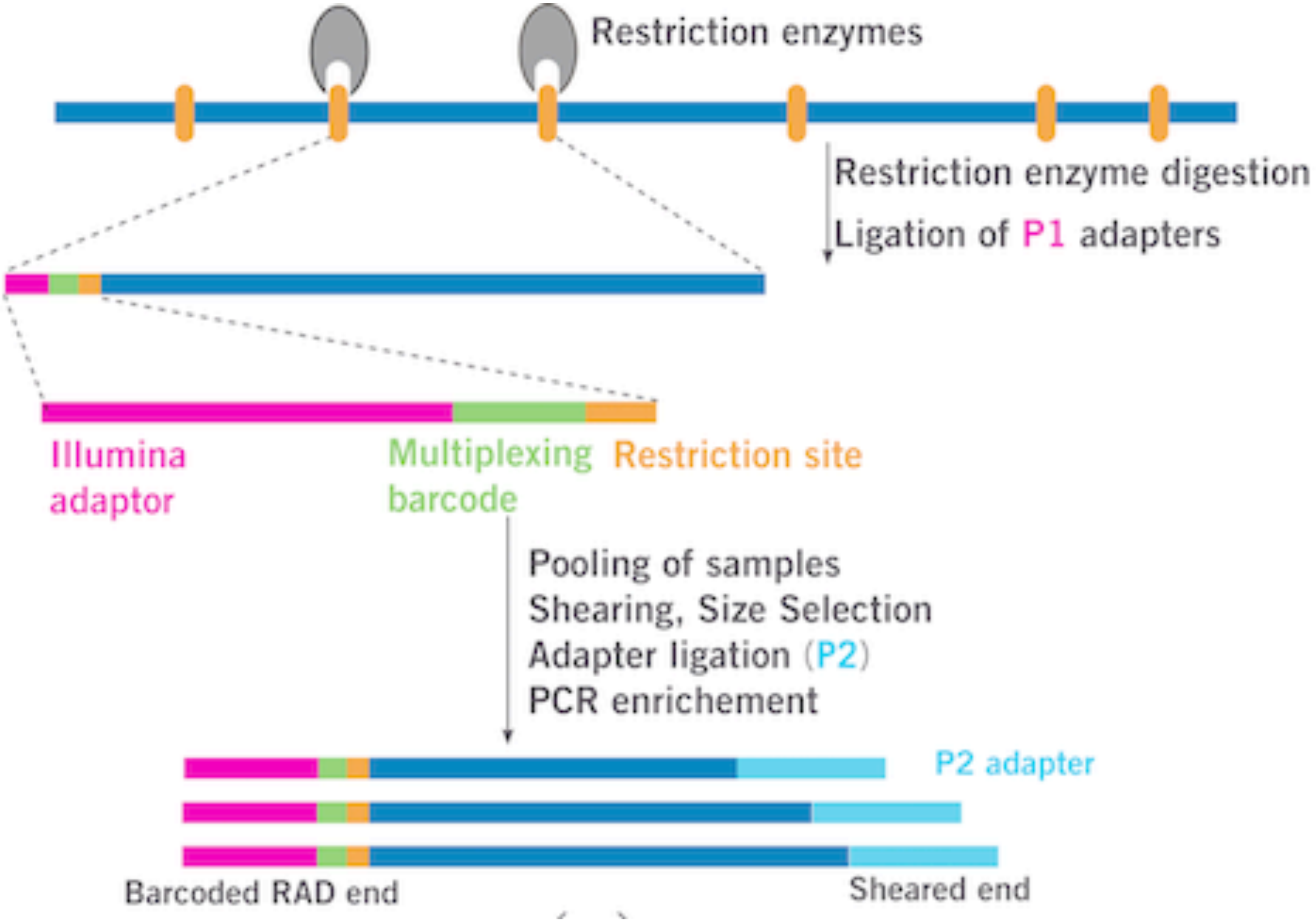
- Subsampling only at specific sites defined by restriction enzyme

Patented by University of Oregon, licensed by:

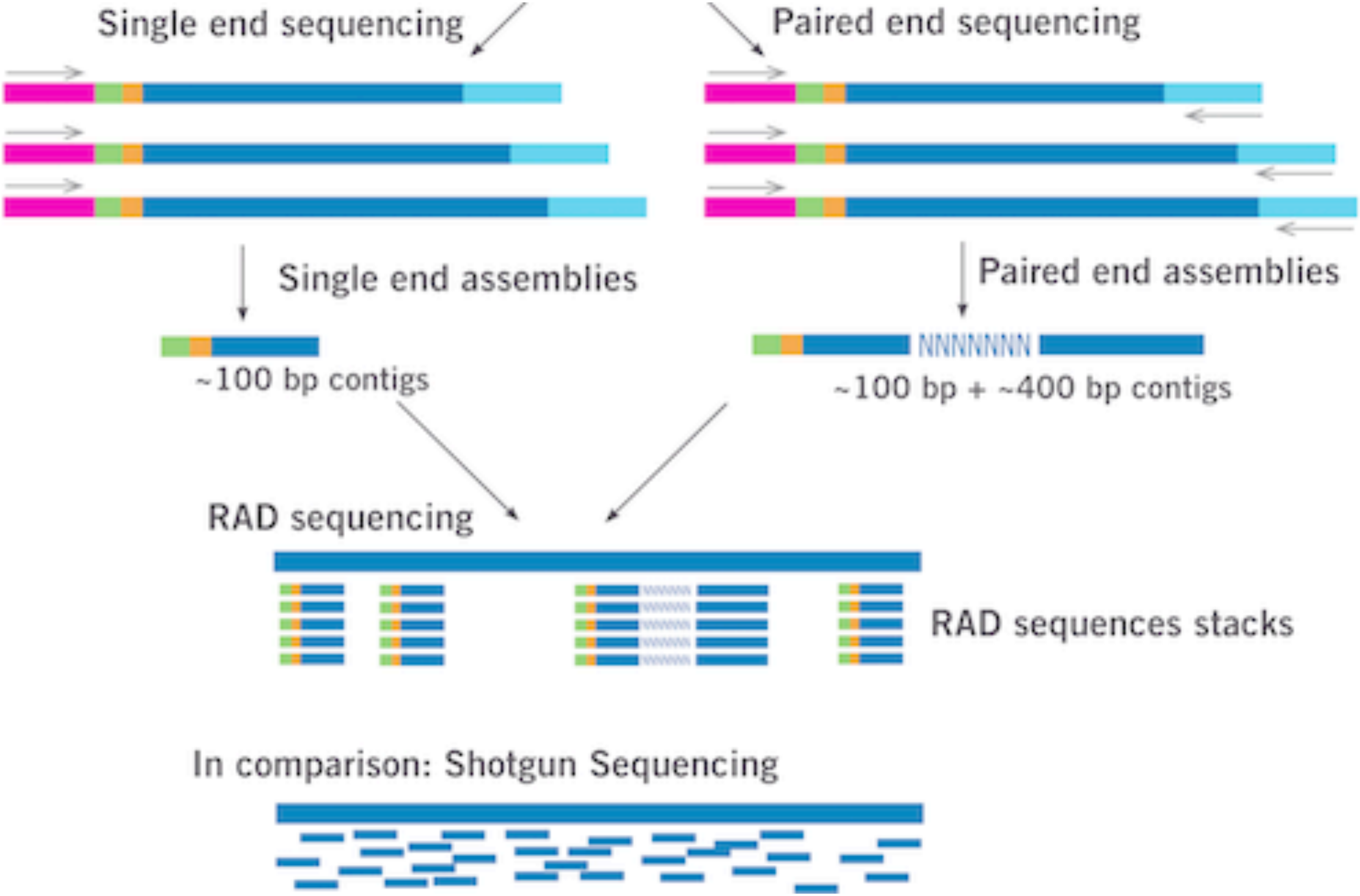
KeyGene

FLORAGENEX

Non profit use at universities can be licensed for free



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



Single end sequencing

Paired end sequencing

Single end assemblies

Paired end assemblies

NNNNNNN

~100 bp contigs

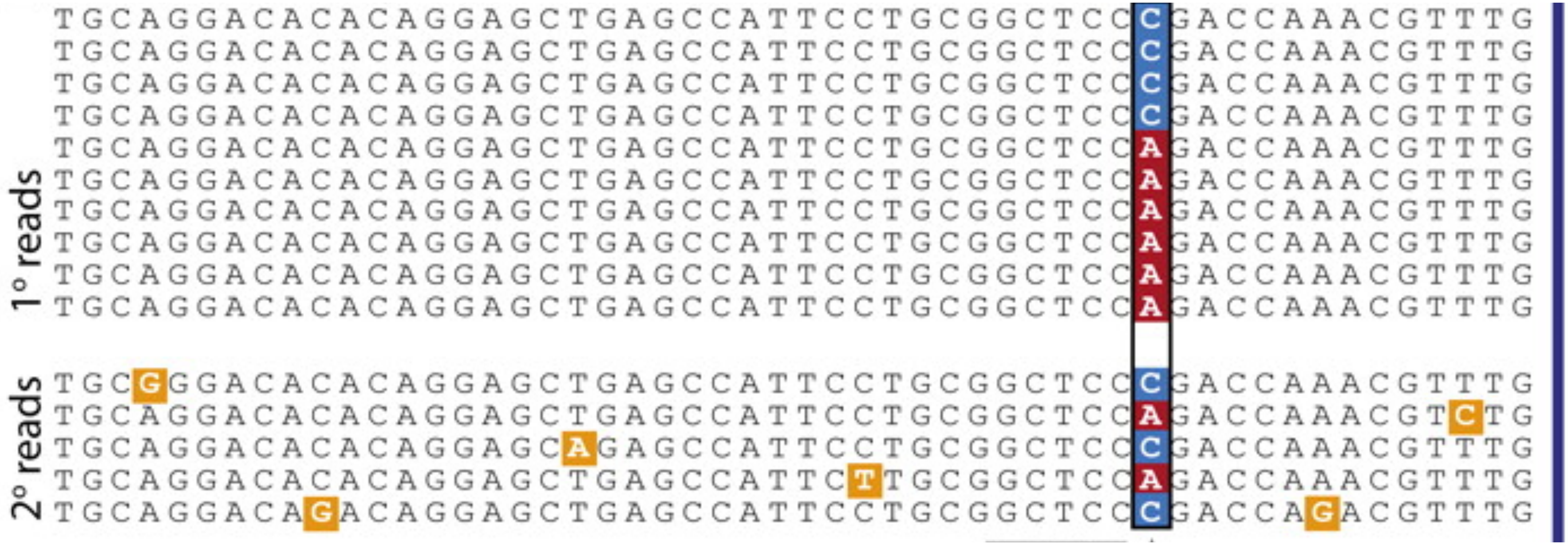
~100 bp + ~400 bp contigs

RAD sequencing

RAD sequences stacks

In comparison: Shotgun Sequencing

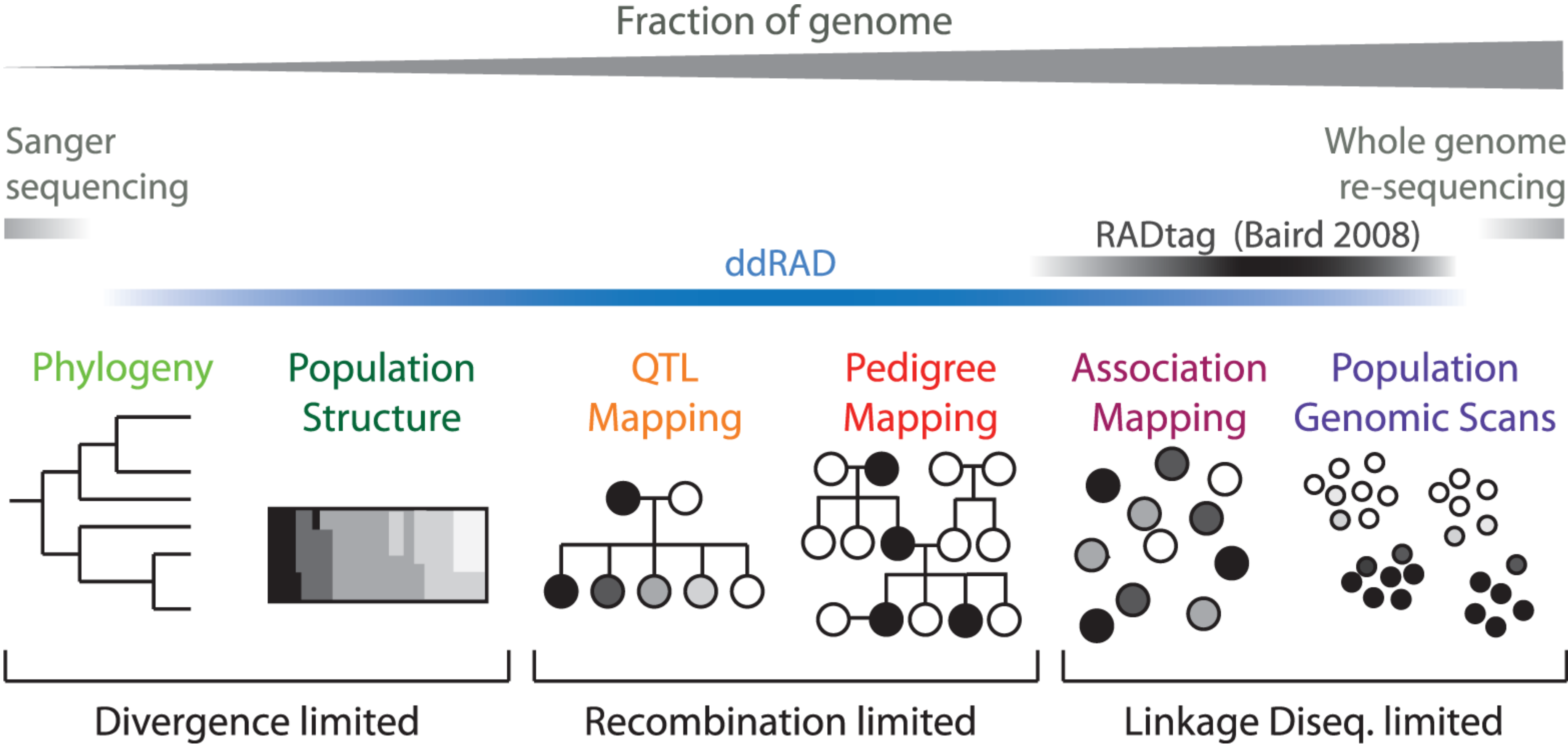
http://www.floragenex.com/rad-seq/


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

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

ddRAD

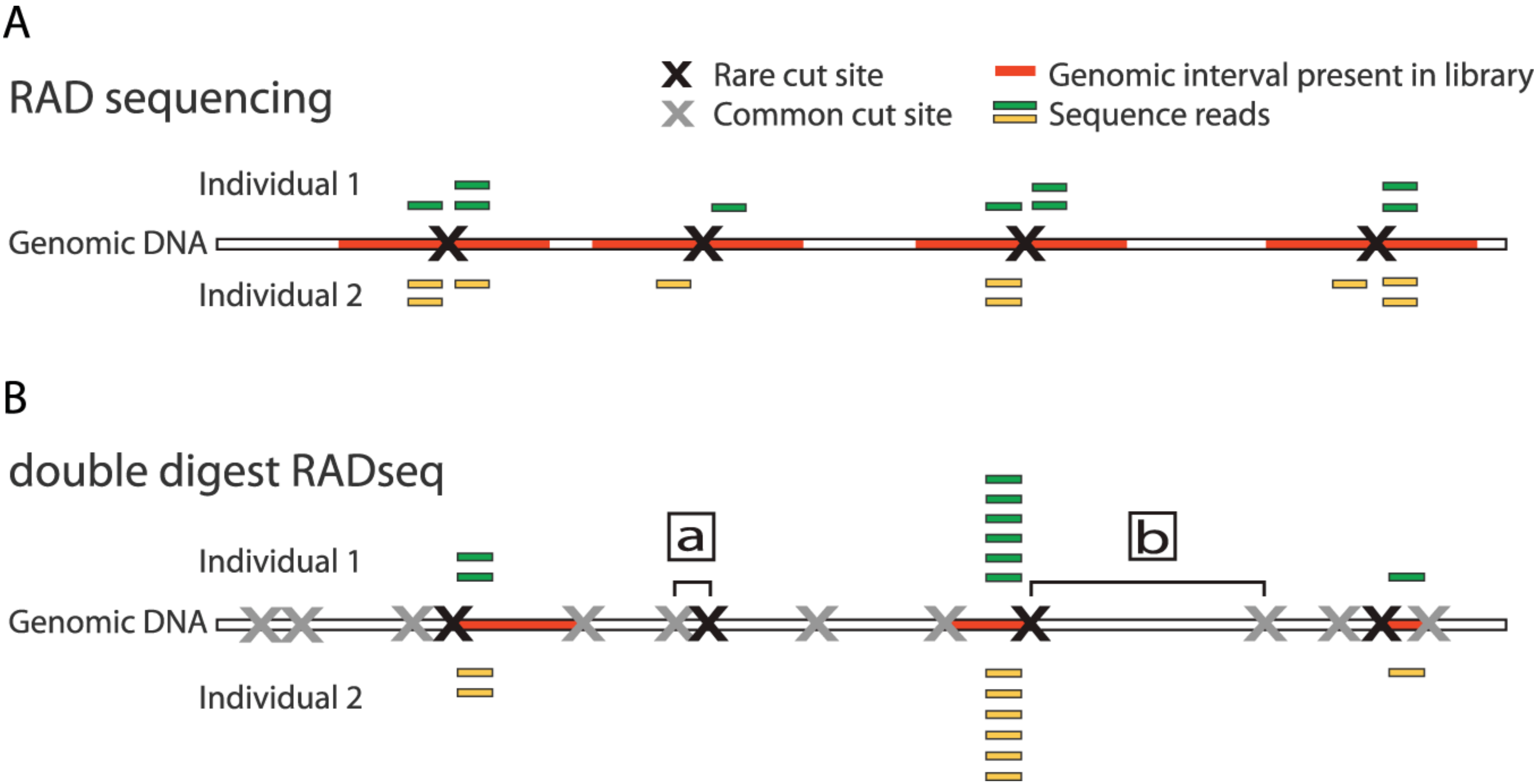
- May have the problem of too many sites across the genome (even if you use a rare cutter)
- Need a way to more accurately control the number of loci sequenced
- Double digest RAD
- Peterson et al 2012
- Also patented



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.

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

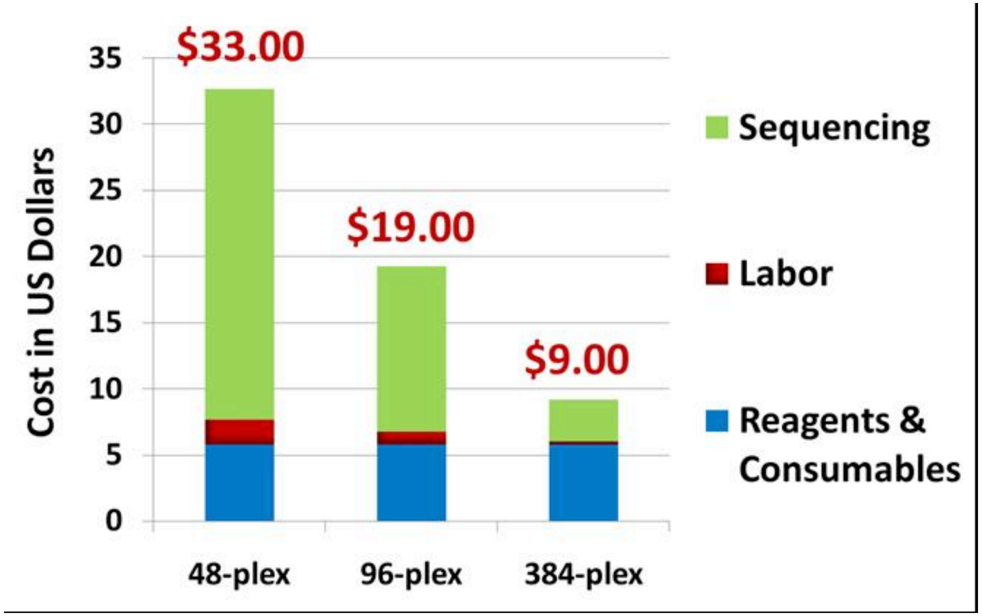


Peterson et al (2012) Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. PLoS ONE 7(5): e37135.

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)

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



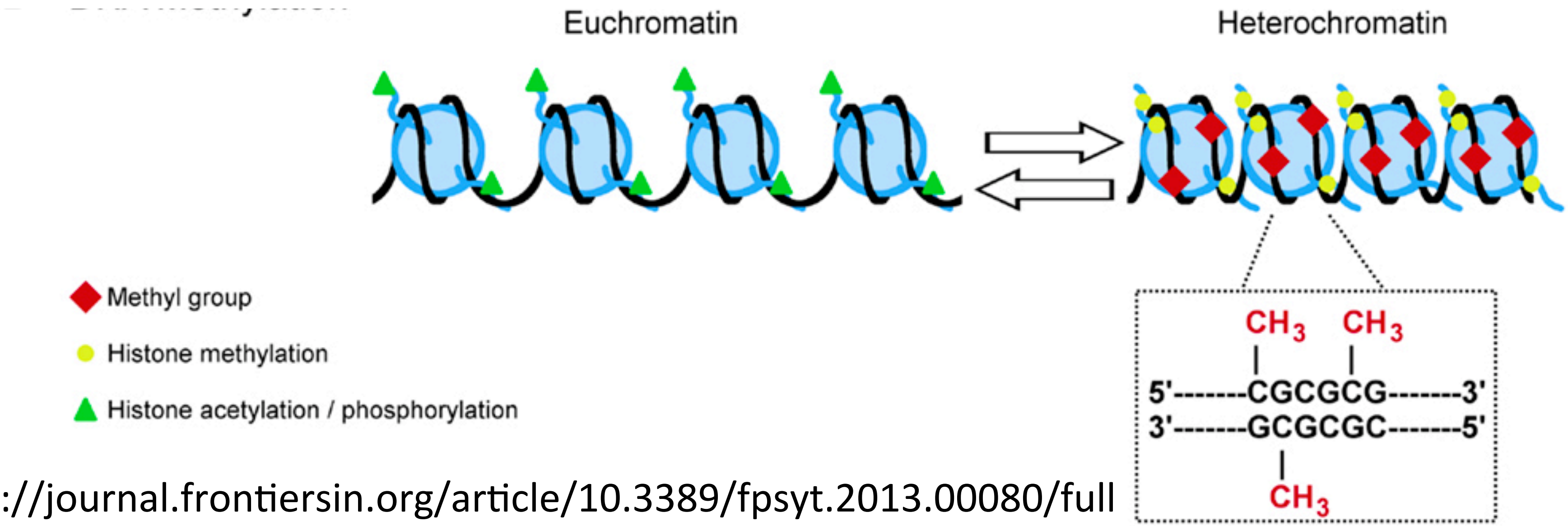
http://www.maizegenetics.net/#!genotyping-by-sequencing-gbs/c9c6

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 4: Bisulfite Sequencing

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



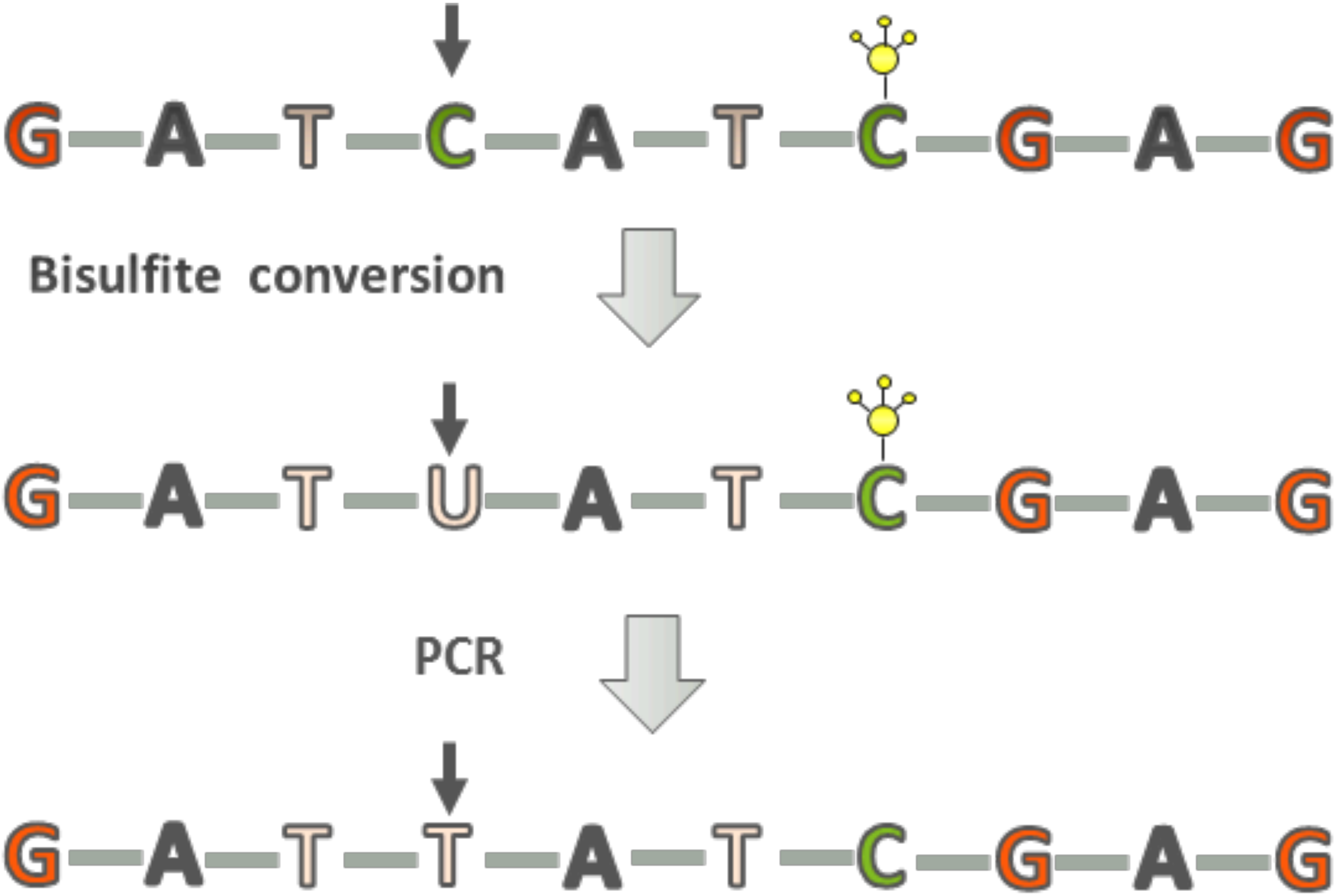
http://journal.frontiersin.org/article/10.3389/fpsyt.2013.00080/full

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

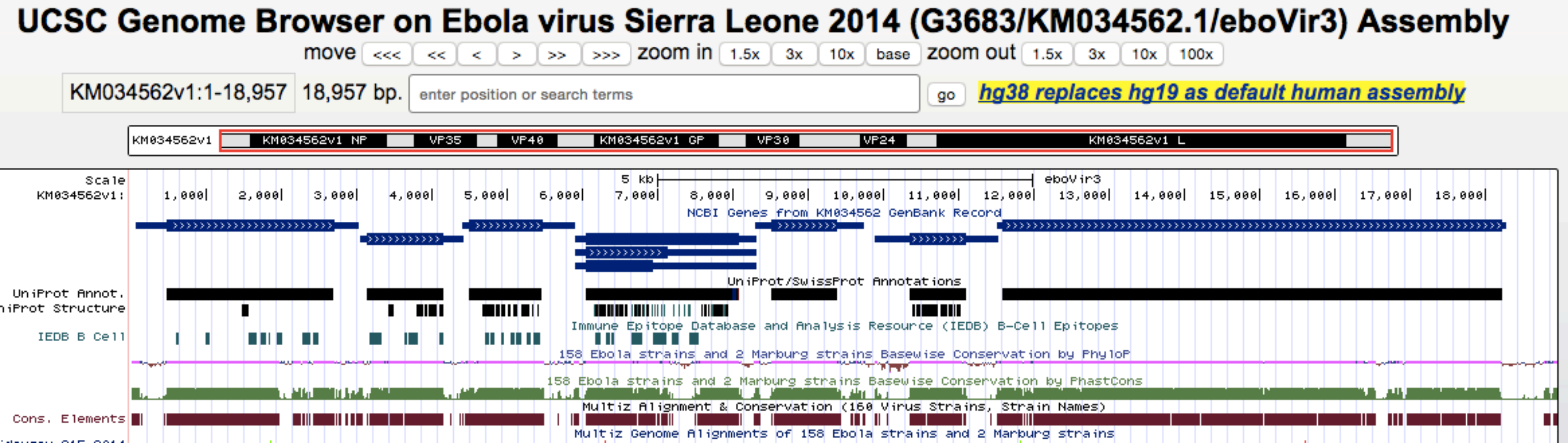
Reference Genomes

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 (49)
- Other Vertebrate (24)
- Deuterostome (3)
- Insect (13)
- Nematode (6)
- Other (2) – yeast and sea hare
- Viruses (1) - Ebola

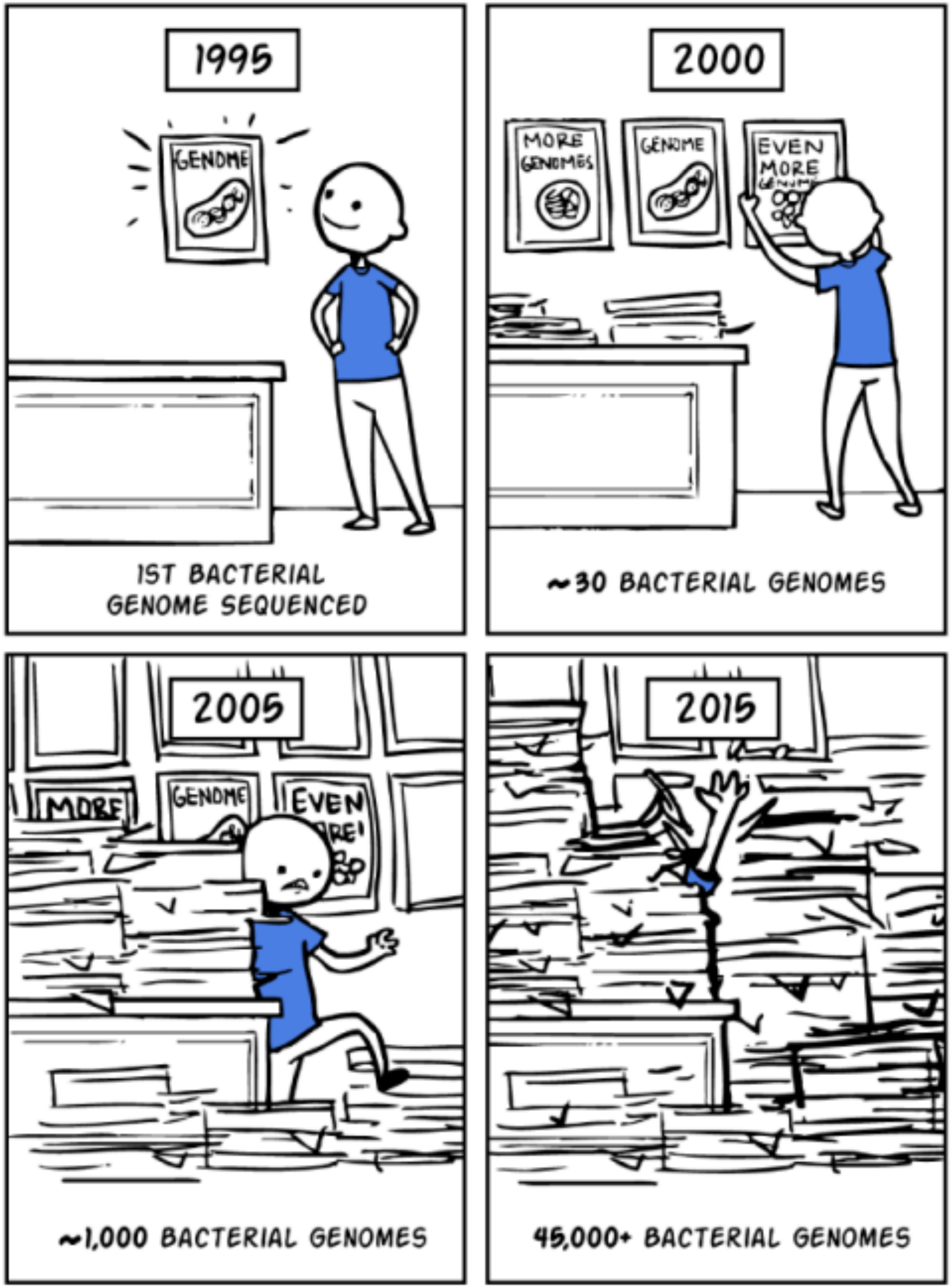


UCSC Genome Browser on Ebola virus Sierra Leone 2014 (G3683/KM034562.1/eboVir3) Assembly

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.
UniProt 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
  - 5,673 genomes
  - 5,639 complete (> 99%)
- Prokaryotes
  - 73,708 genomes
  - 5894 complete (8.0%)
  - 1024 chromosome level (1.4%)
- Eukaryotes
  - 3,494 genomes
  - 22 complete (< 1%)
  - 440 chromosome level (12.6%)

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

Sequencing for *de novo* Assembly

- Reconstructing the original full DNA molecules from (short) read fragments
- Jigsaw puzzle
- How do the pieces fit together? (overlap)
- Missing pieces (sequencing bias)
- Dirty pieces (sequencing error, real biological variation)

ULYSSES
JAMES JOYCE
ULYSSES
JAMES JOYCE
ULYSSES
JAMES JOYCE

An example

A small “genome”:

Friends,
Romans,
countrymen
lend me your ears;

Reads:

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
cryman, lend me
```

Overlaps:

```
Friends, Rom
     ds, Romans, count
             ns, countrymen, le
                     crymen, lend me
                             send me your ears
```

Consensus:

```
Friends, Romans, countrymen, lend me your ears;
```

Sequencing for *de novo* Assembly

- Strategy differs significantly from resequencing
- Spectrum of difficulty:
  - Size
  - Repetitiveness
  - Polyploidy
  - Heterozygosity

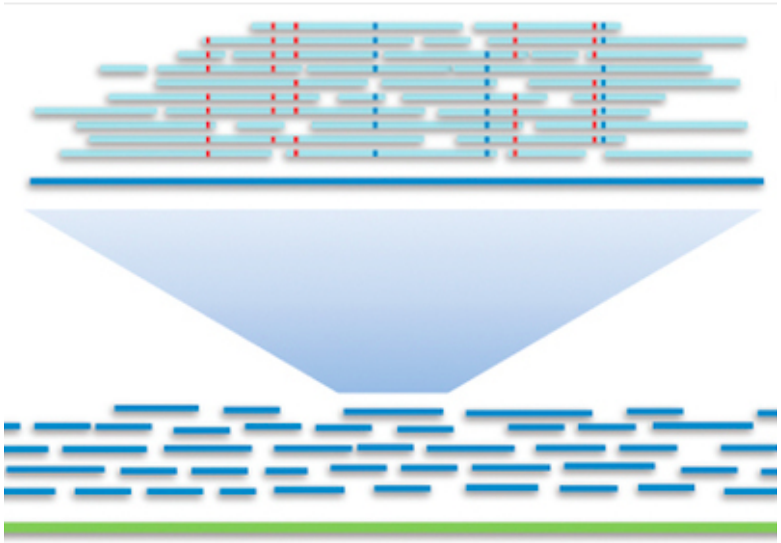
From 1990s-2005, genomes were sequenced with Sanger technology (7-10X coverage) and BAC-based physical maps

Why don't we just do it the way we used to do it?

- Too much money!

NGS Strategies

- New Strategy 1: Ultra high throughput Illumina + Variety of sequence libraries
  - Mate pairs at a range of distances: 5kb, 10kb, 20kb, 40kb
- New Strategy 2: PacBio + Illumina
  - Error correct the long PacBio reads with Illumina
- New Strategy 3: PacBio only
  - More expensive.
- Challenges:
  - Initial contig build is computationally intensive
  - Many assembly algorithms require 100s of Gb of RAM to Tbs of RAM



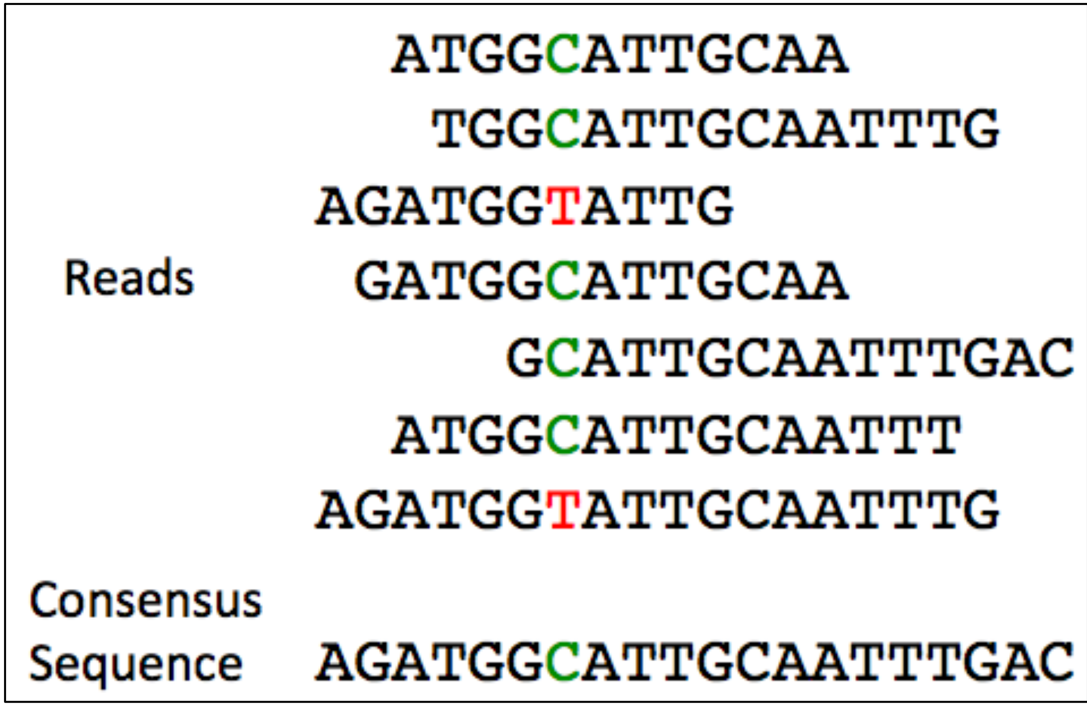
Strategy 2.

Chin et al., Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. 2013 Nature methods

Workflow

1. Clean all contaminants out of reads (much more critical than for resequencing!)
2. Build contigs

(There are many strategies for building contigs – next lesson we will learn about those and about why contigs don't span the whole chromosome)



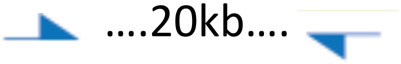
```
                     ATGGCATTGCAA
                      TGGCATTGCAATTTG
                   AGATGGTATTG
Reads               GATGGCATTGCAA
                        GCATTGCAATTTGAC
                     ATGGCATTGCAATTT
                   AGATGGTATTGCAATTTG

Consensus
Sequence           AGATGGCATTGCAATTTGAC
```

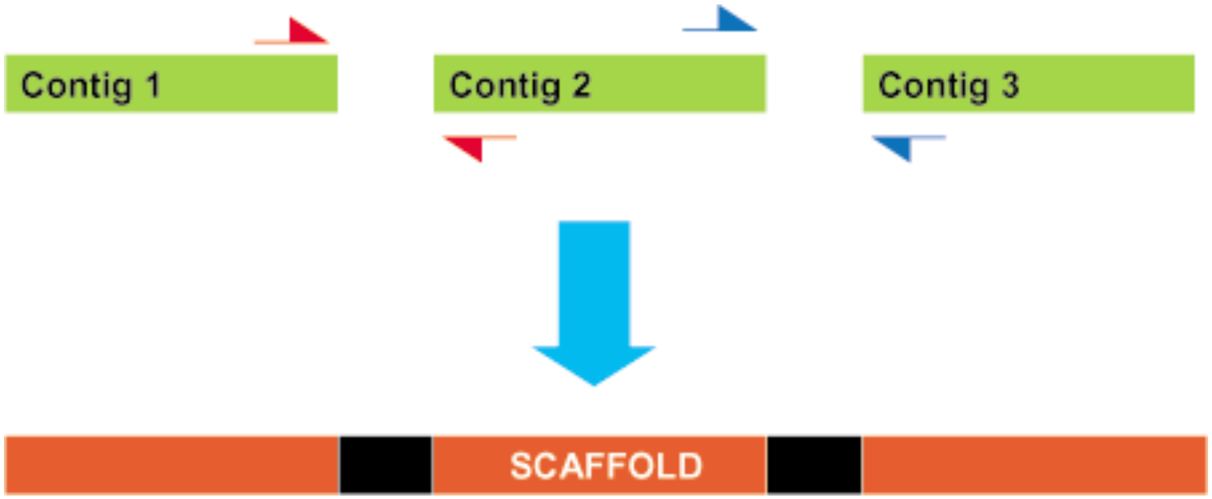
http://gcat.davidson.edu/phast/

Workflow (cont)

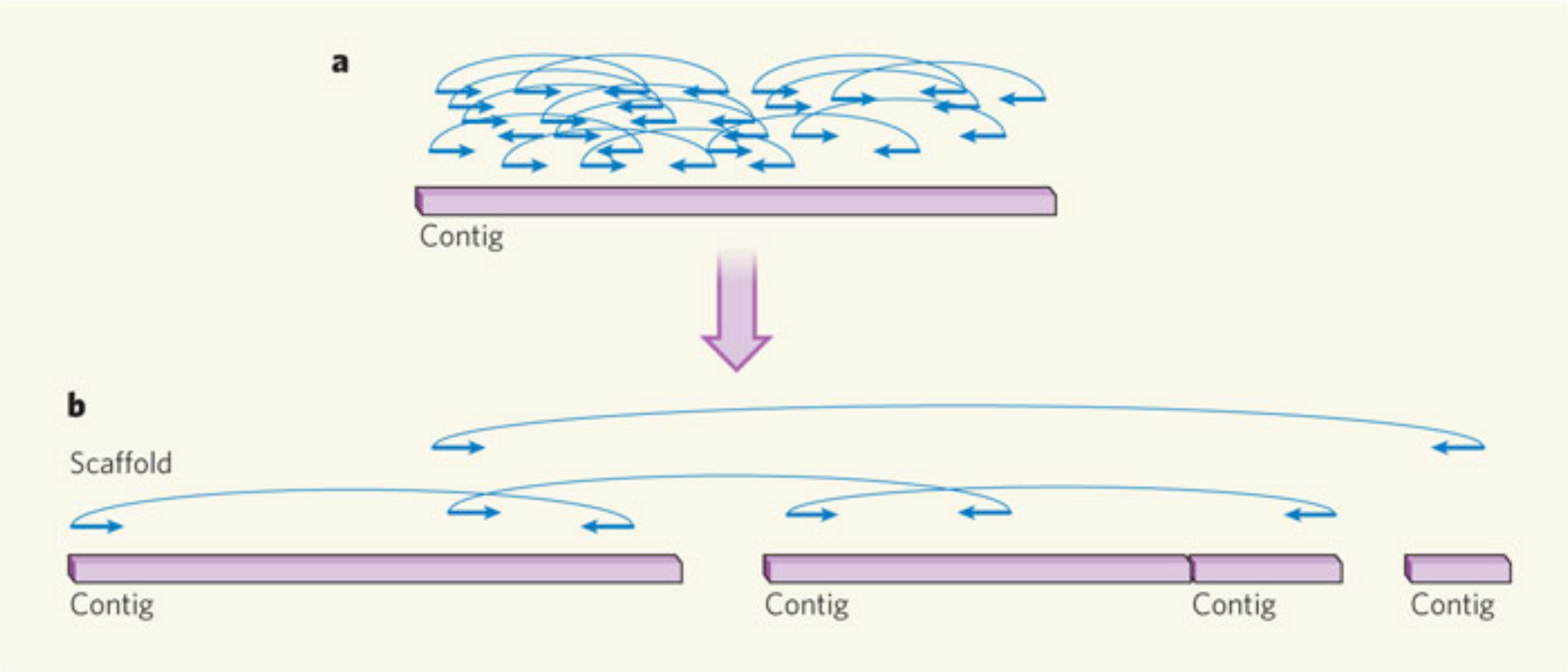
3. Scaffold

Mate pairs





http://scienceblogs.com/evolgen/2007/02/07/shotgun-sequencing-a-eukaryoti/



Mate pairs order and orient the contigs against each other.
Still leave gaps.

Genetics: Decoding a national treasure
Kim C. Worley & Richard A. Gibbs
Nature 463, 303-304(21 January 2010)

Workflow (cont)

4. Gap fill
   - Use your existing reads
   - Sanger sequencing
5. Assess quality and internal consistency
   - Expected size?
   - How many reads map consistently?

N50 statistic

- the length for which the collection of all contigs of that length or longer contains at least 50%
- similar to a mean or median of contig lengths
- used widely in genome assembly, especially in reference to contig lengths within a draft assembly.

N50 Example

- Assembly A contains six contigs of lengths:
  - 80 kbp, 70 kbp, 50 kbp, 40 kbp, 30 kbp, and 20 kbp
  - Sum size of assembly A is 290 kbp
  - N50 contig length is 70 kbp
  - "Half of the assembly is contained in contigs of 70kbp or greater"
  - If you randomly selected a location, 50% of the time it would be in a contig of 70kbp or greater
- Assembly B contains eight contigs of lengths:
  - 80 kbp, 70 kbp, 50 kbp, 40 kbp, 30 kbp, 20 kbp, 10kbp, 5kbp
  - Sum size of assembly B is 305 kbp
  - N50 contig length is 50 kbp

Overview

- DNA Sequencing Applications
  1. Resequencing
  2. Capture/Targeted Resequence
  3. RAD/ddRAD/GBS
  4. Bisulfite Sequencing
  5. De novo assembly
- Reference Genome availability
- N50

Python functions!