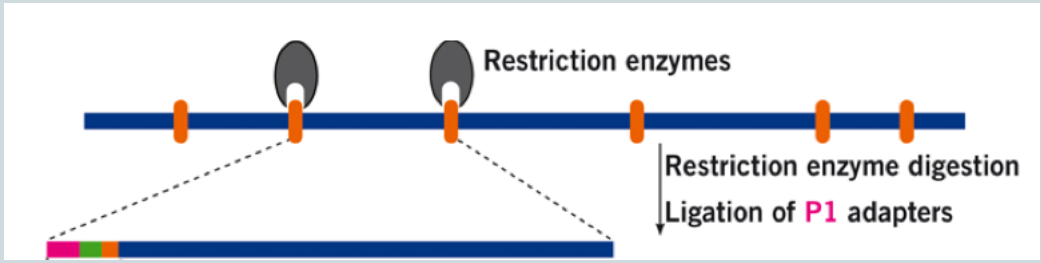
# RADseq: from tissue to data

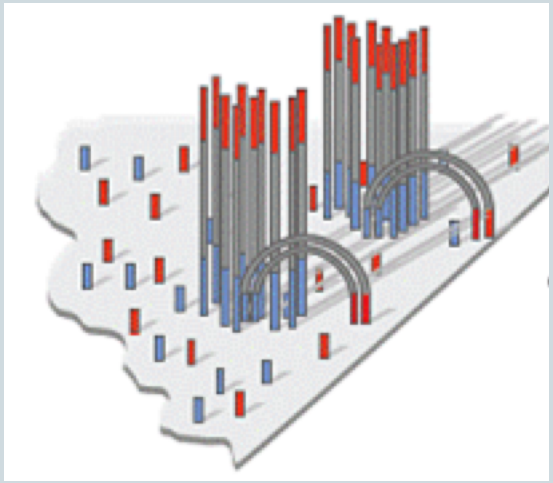
ALEX JANSEN VAN RENSBURG
DEBBIE LEIGH

INSTITUTE OF EVOLUTIONARY BIOLOGY & ENVIRONMENTAL STUDIES

JOSH VAN BUSKIRK

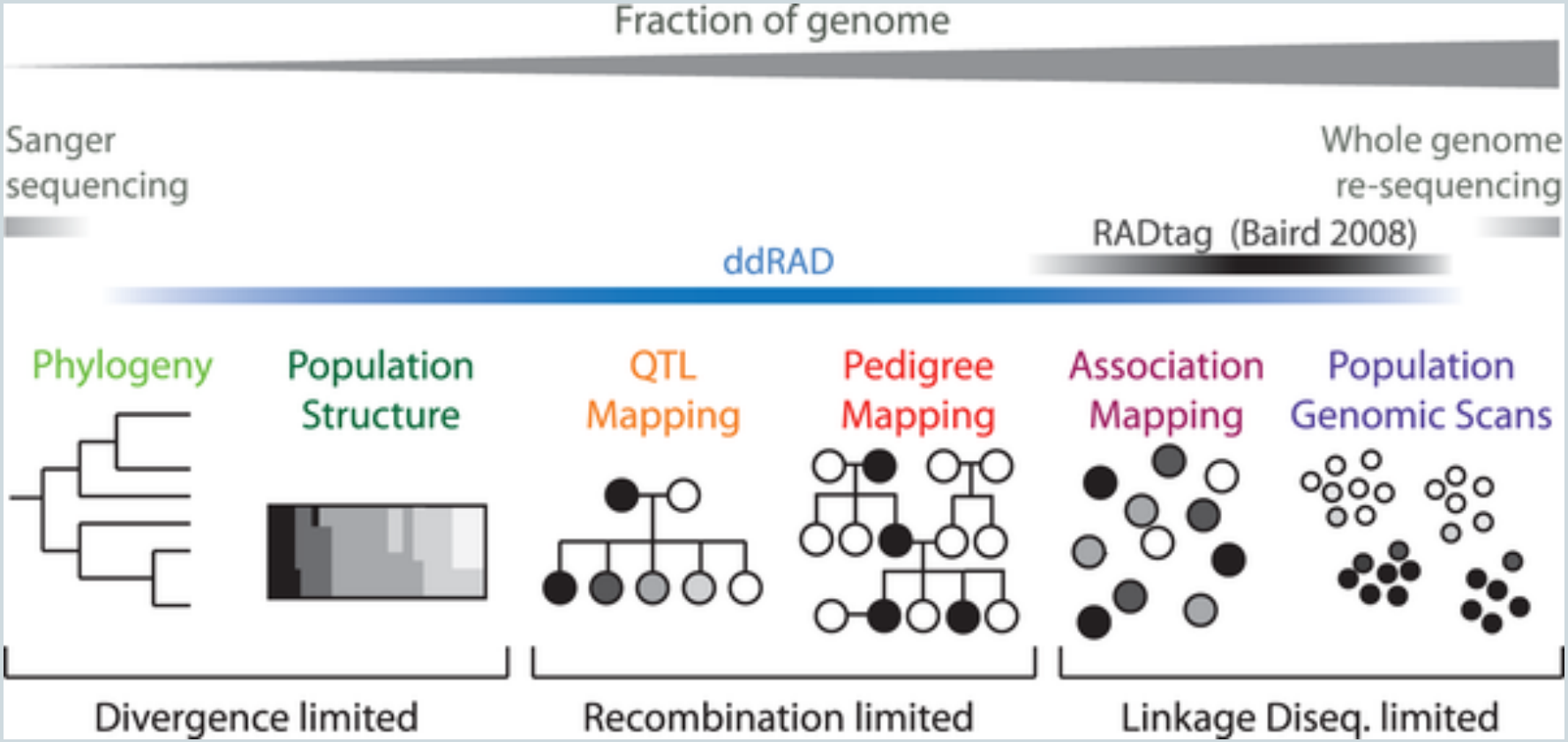
# What is RADseq?



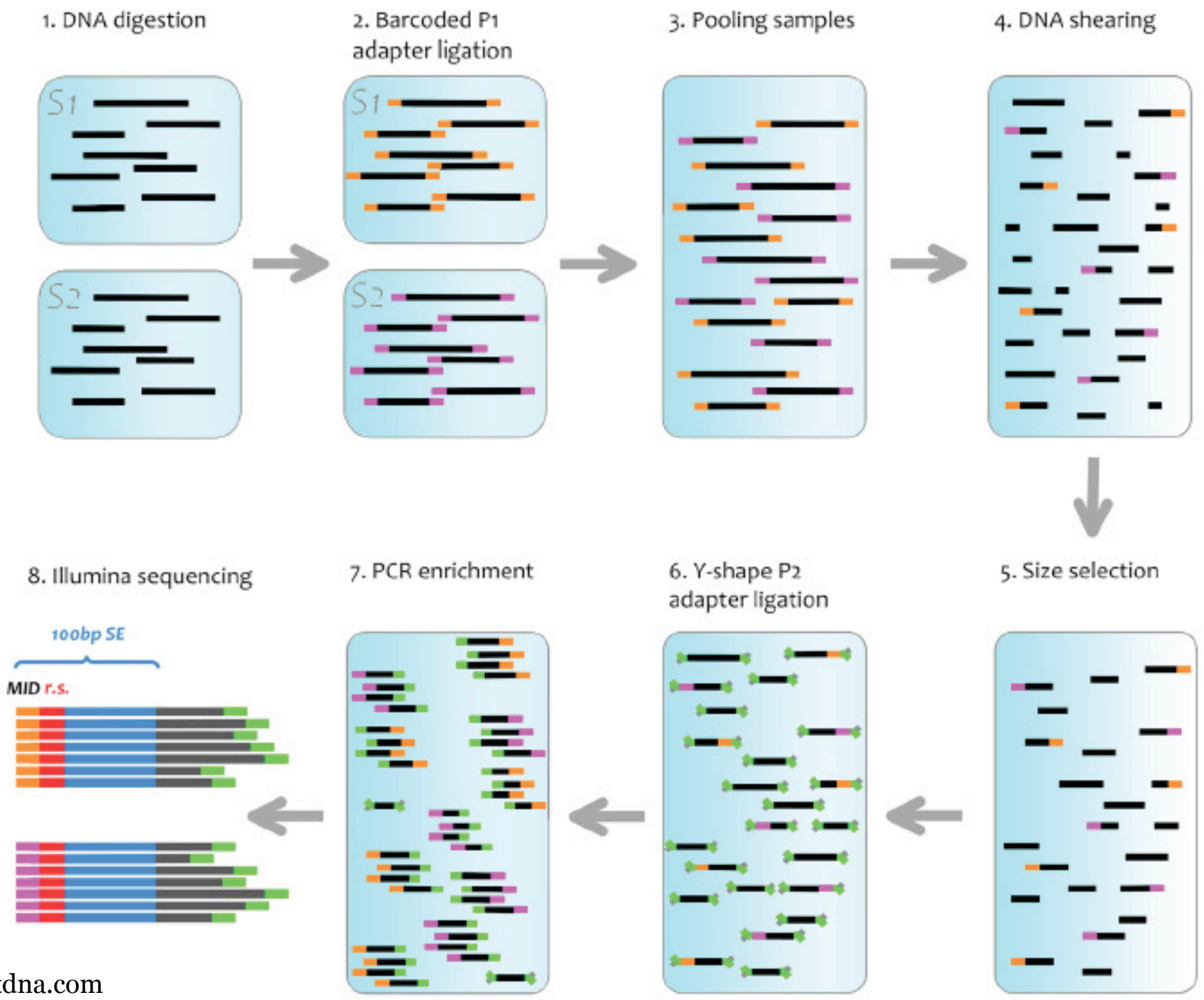



```
ggt ggc tgg ggt cag
ggt ggc tgg ggc cag
ggt ggc tgg ggc cag
ggt ggc tgg ggt cag
ggt ggc tgg gga cag
gga ggc tgg ggt caa
```

# What is RAD good for?




Peterson *et al.* 2012

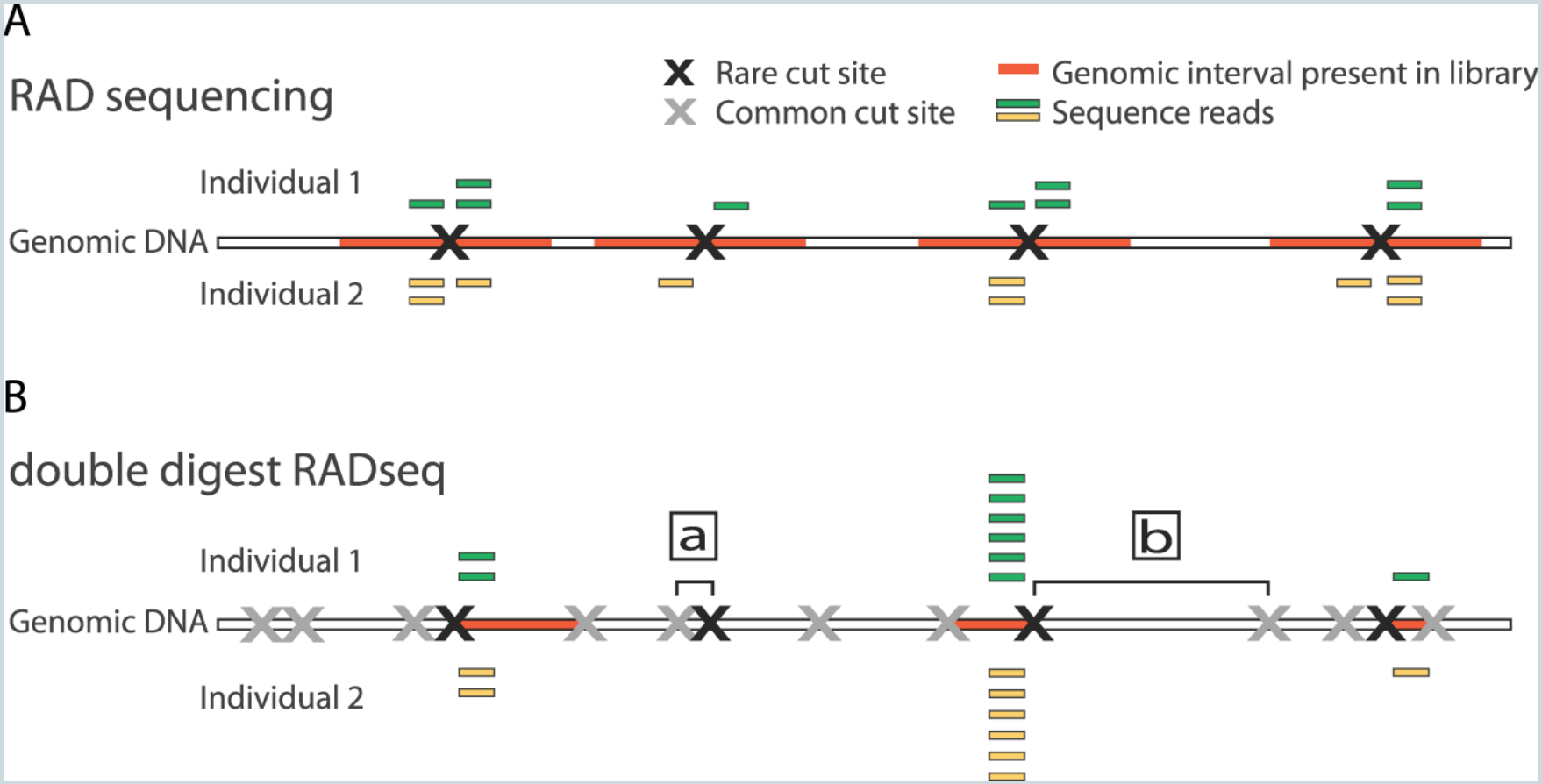

1. DNA digestion
S1
S2
2. Barcoded P1 adapter ligation
S1
S2
3. Pooling samples
4. DNA shearing
5. Size selection
6. Y-shape P2 adapter ligation
7. PCR enrichment
8. Illumina sequencing
100bp SE
MID r.s.

# Many different types…

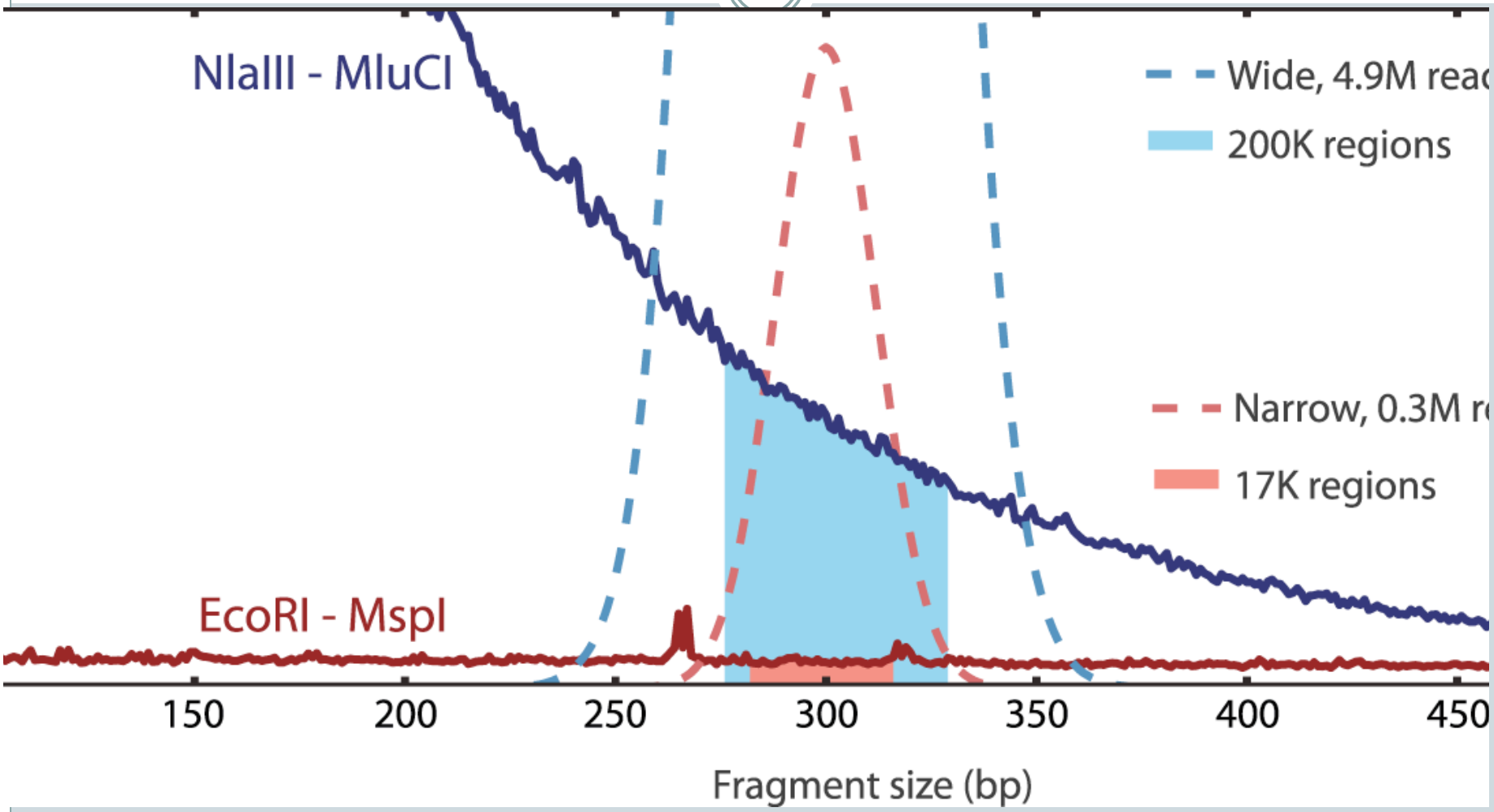
- **mbRAD** (Miller *et al.* 2007; Baird *et al.* 2008)
- **ddRAD** (Peterson *et al.* 2012)
- ezRAD (Toonen *et al.* 2013)
- 2bRAD (Wang *et al.* 2012)
- …GBS


Review: Davey *et al.* 2011
Demystifying the RAD fad: Puritz *et al.* 2014

# mbRAD vs ddRAD

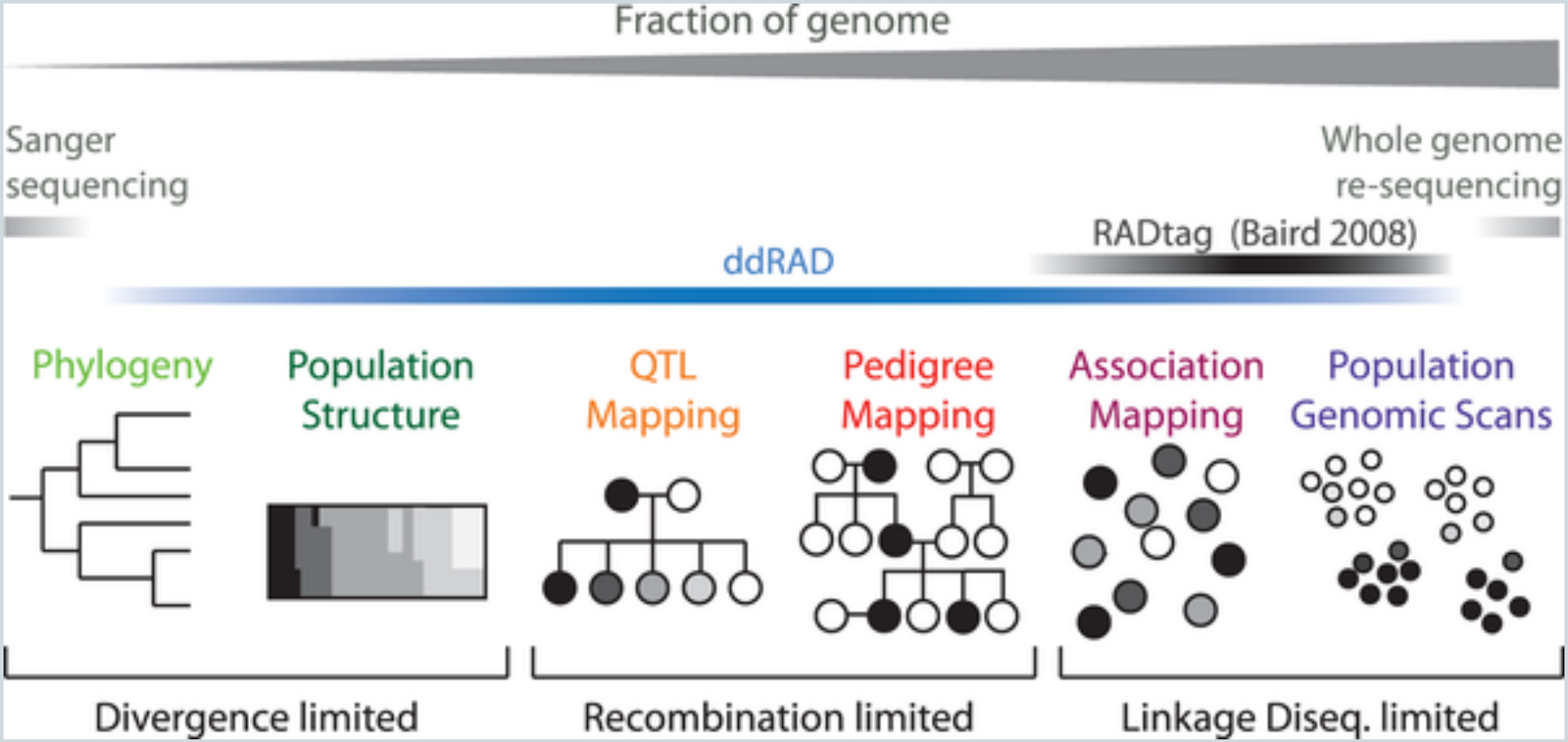


# ddRAD – flexibility in the number of loci




Peterson *et al.* 2012

# What is RAD good for?




Peterson *et al.* 2012

# Many different types…

- **mbRAD** (Miller *et al.* 2007; Baird *et al.* 2008)
- **ddRAD** (Peterson *et al.* 2012)
- ezRAD (Toonen *et al.* 2013)
- 2bRAD (Wang *et al.* 2012)
- …GBS


So… which one suits me best?

Paired-end vs single-end sequencing?

Pooled vs individual labeling?

Experimental design?

Davey *et al.* 2011
Demystifying the RAD fad (Puritz *et al.* 2014)

# Experimental design

- What is the question?
  - Population genetics
  - Phylogenetics
  - Looking for selection
- What is your organism/system?
- Available genomic resources?
- Budget

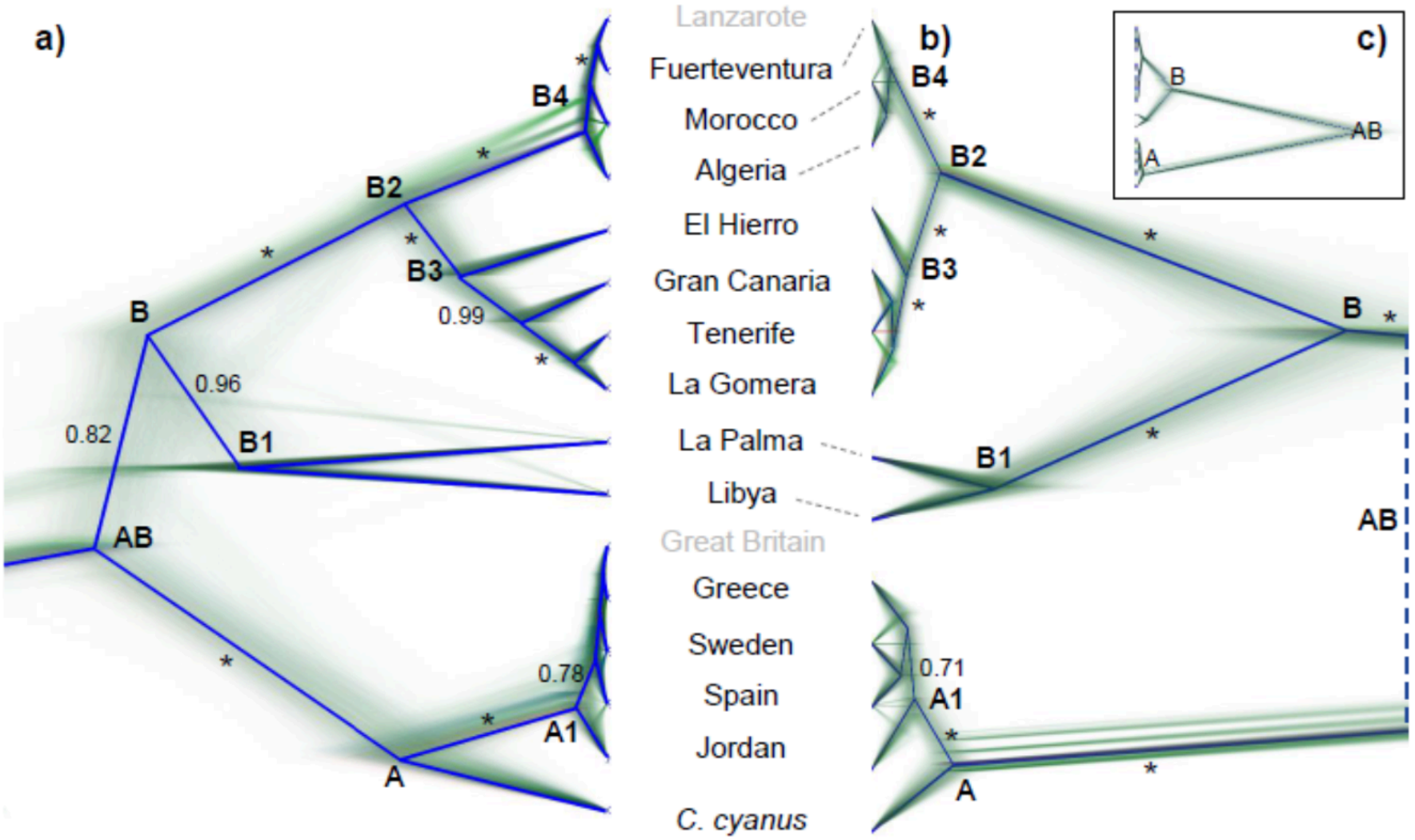
# Experimental design

- What is the question?
  - Population genetics
  - Phylogenetics
  - Looking for selection
- What is your organism/system?
- Available genomic resources?
- Budget

# Experimental design: Question?

- Pop genetics & phylogenetics:
  - **Controversial/discussed points:**
    - Pooled vs individual labels (e.g. Emerson *et al.* 2014)
    - Coverage needed? (e.g. Davey *et al.* 2013)
    - Markers vs individuals

# Experimental design: Question?

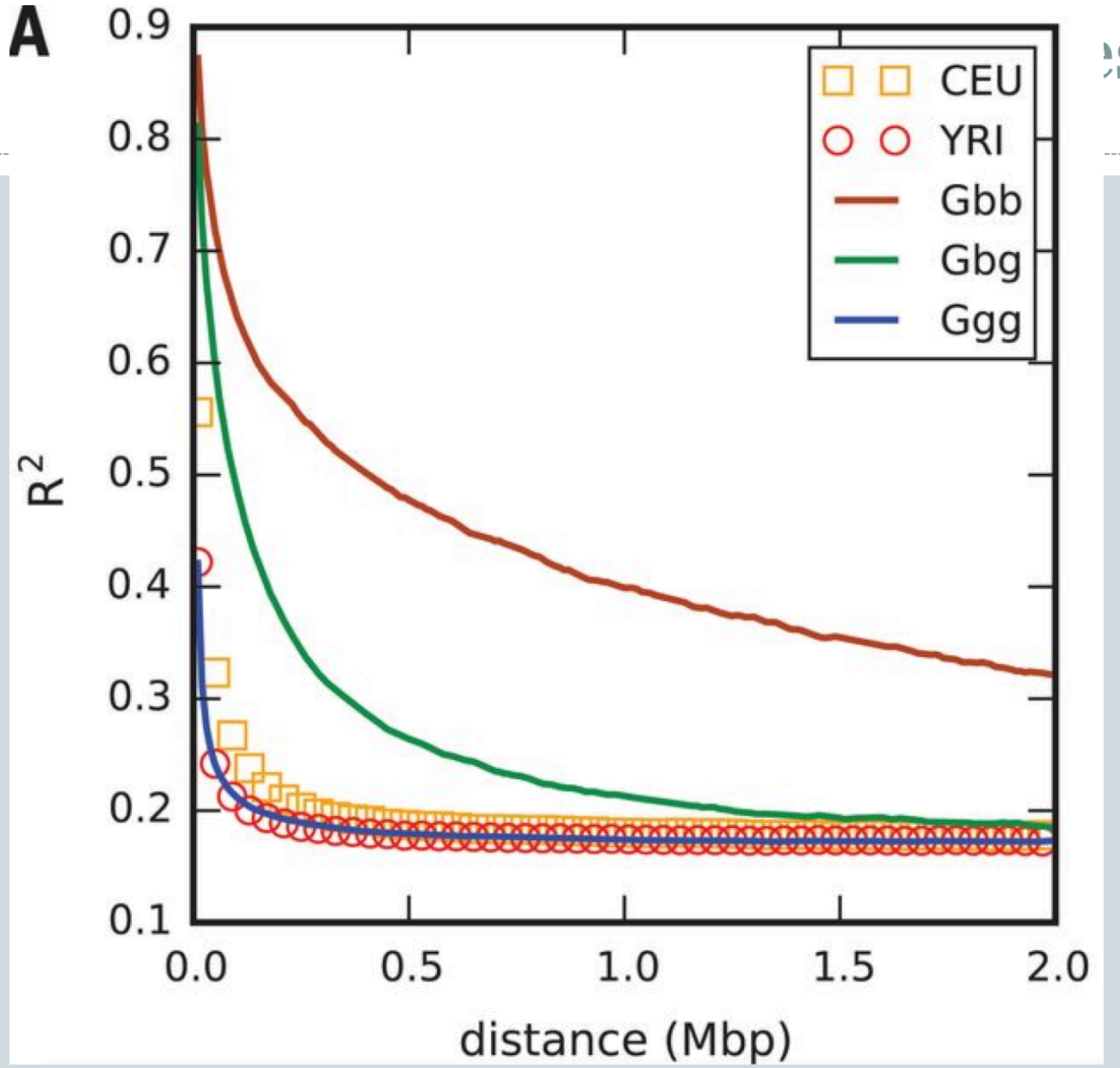



Stervander *et al.* 2015

# Experimental design: Question?

- Marker density:
  - Which restriction enzymes?
- Looking for selection?
  - LD
    - How big are my LD blocks?

# estion?




Mountain Gorilla

Xue *et al.* 2015

# Experimental design

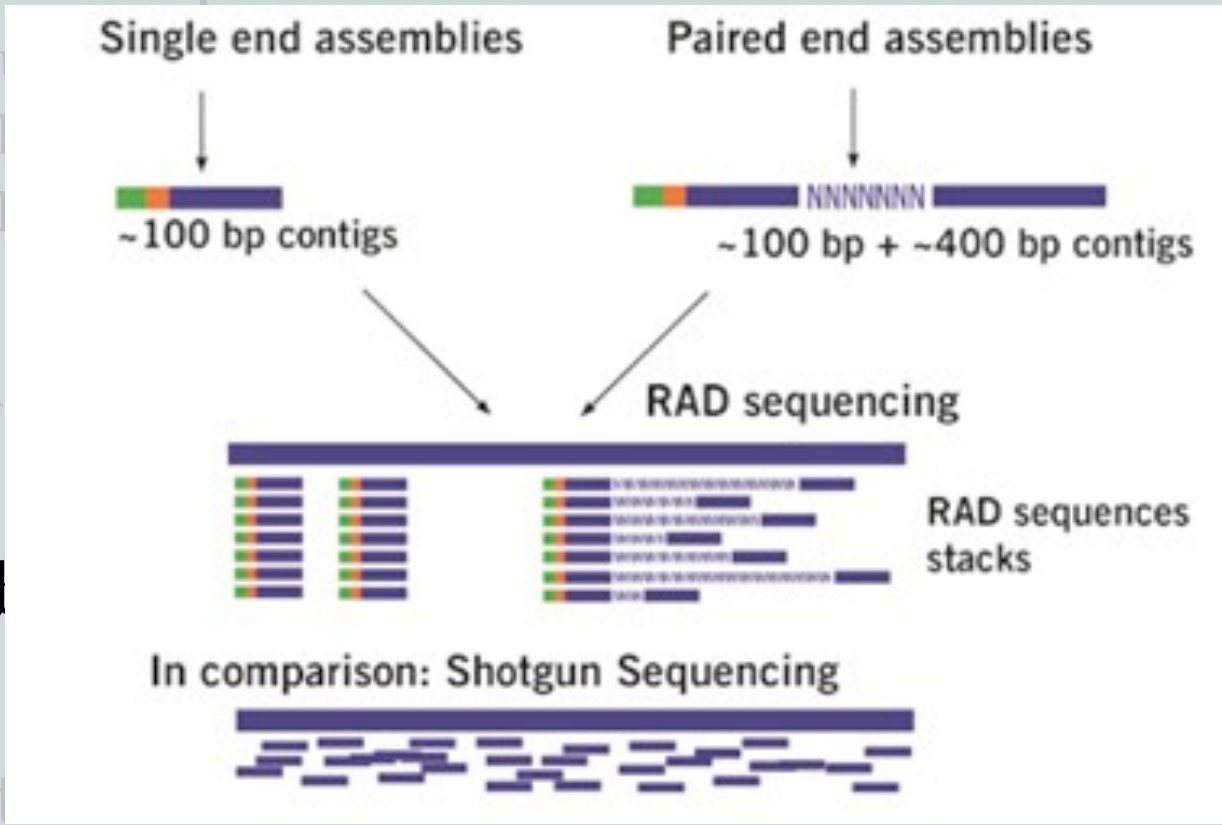
- What is the question?
  - Population genetics
  - Phylogenetics
  - Looking for selection
- What is your organism/system?
- Available genomic resources?
- Budget

# Experimental design

- What is the question?
  - Population genetics
  - Phylogenetics
  - Looking for selection
- What is your organism/system?
- Available genomic resources?
- Budget

# Experimental design

- What is the question?
  - Popu
  - Phyl
  - Look
- What i
- Availab
- Budget



# Experimental design

- What is the question?
  - Population genetics
  - Phylogenetics
  - Looking for selection
- What is your organism/system?
- Available genomic resources?
- Budget

# Lab & Sequencing facilities

- GDC
  - Adapter aliquots (RAD & ddRAD)
  - Covaris
  - Caliper
  - BioAnalyzer
  - Qubit
  - qPCR
  - MiSeq
  - Help & advice!
- FGCZ
  - HiSeq (8 lanes = 1 flowcell; 250Mil reads per lane)
  - MiSeq (1 lane; 44-50Mil reads)

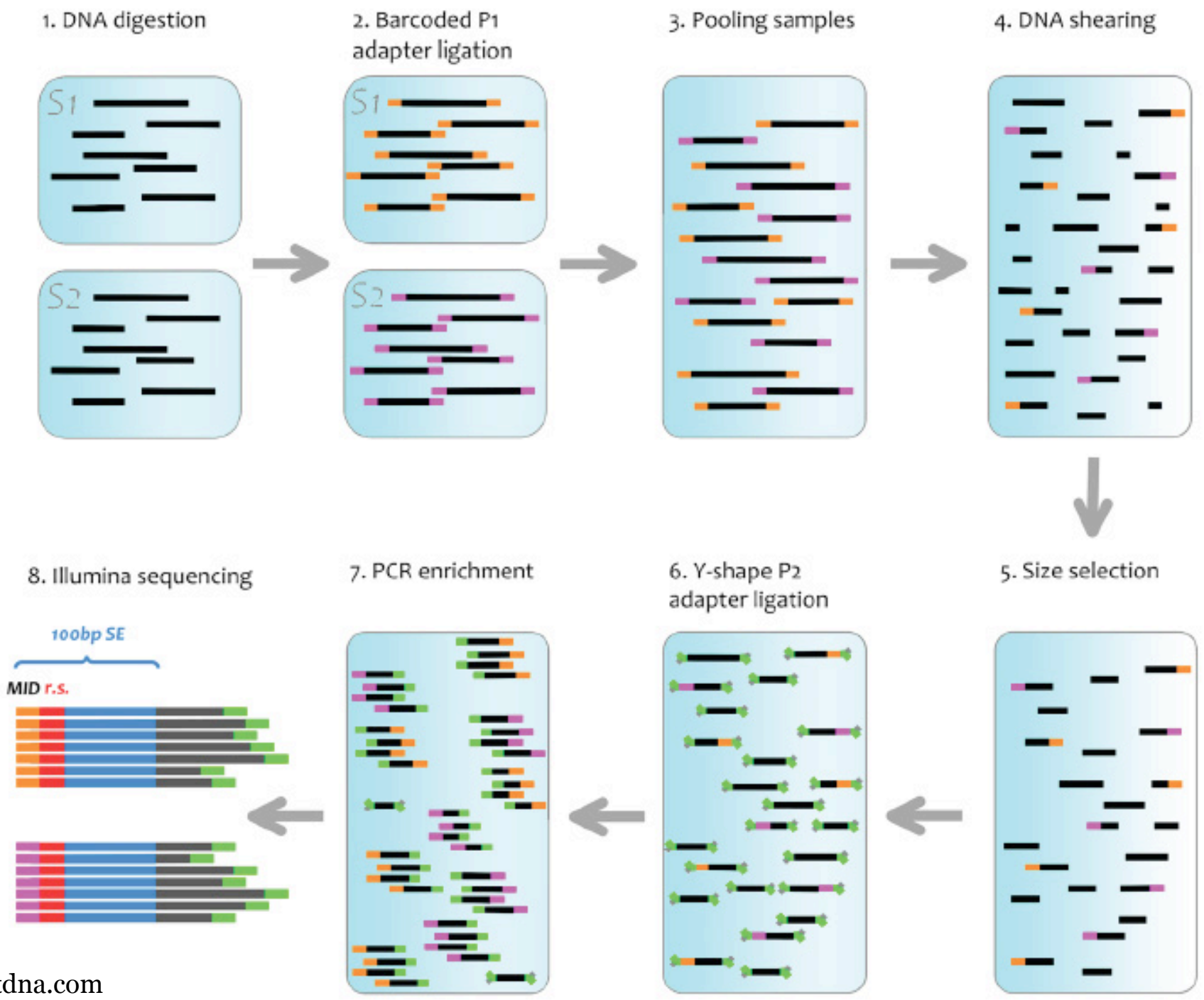
# Wet lab

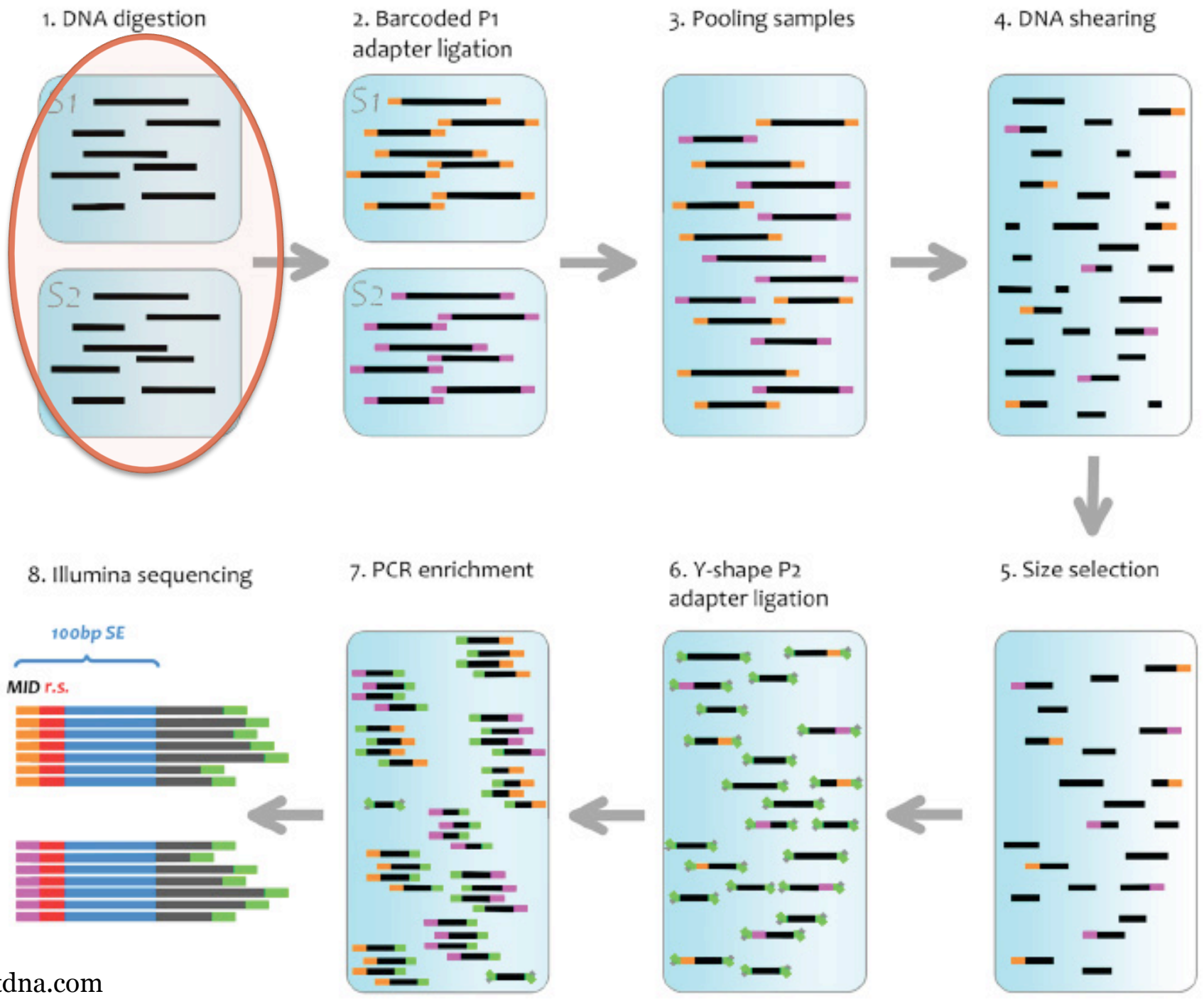
**Lab effort**:

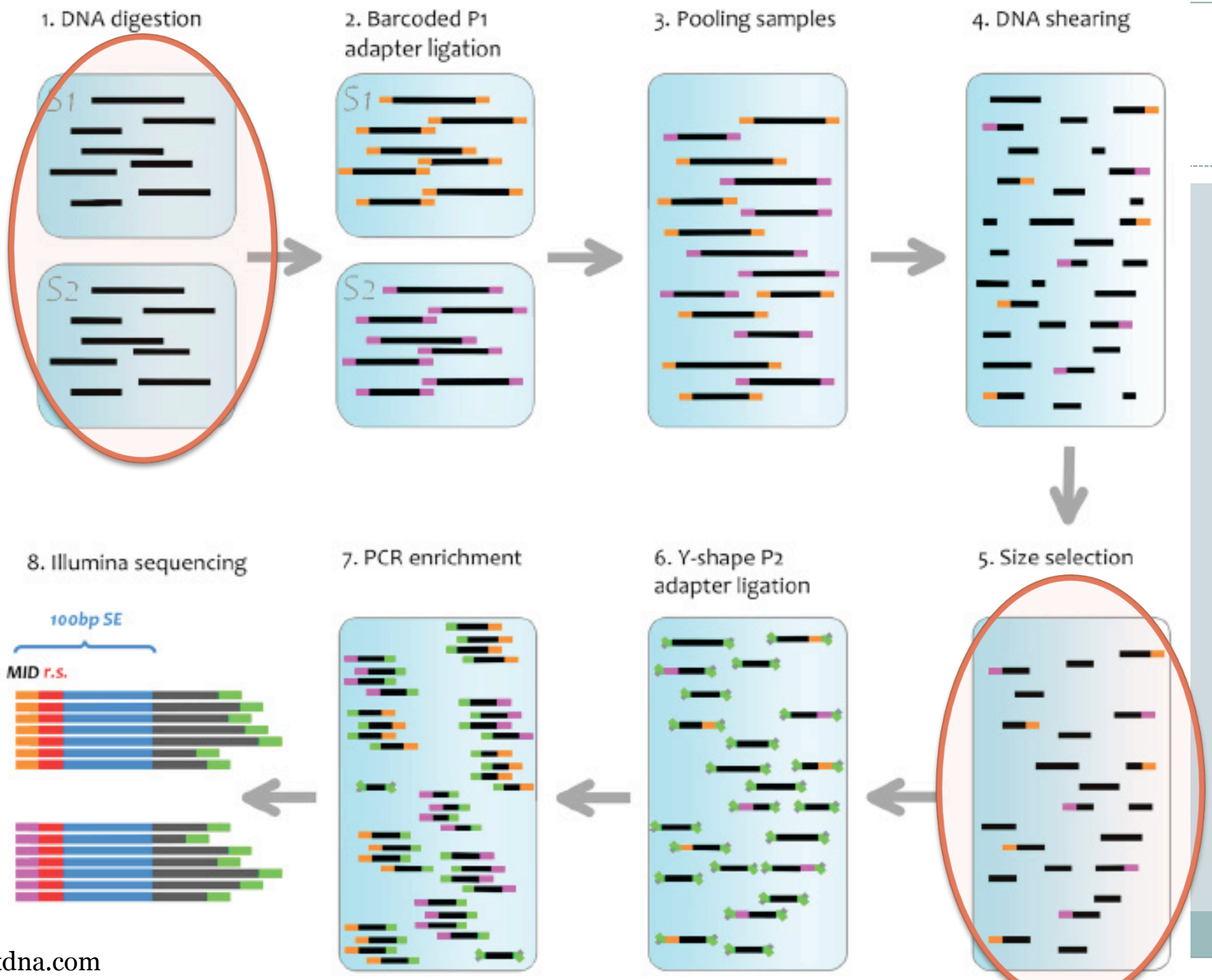
- Need high quality DNA
- Large quantities of DNA
- Taxa specific optimisation

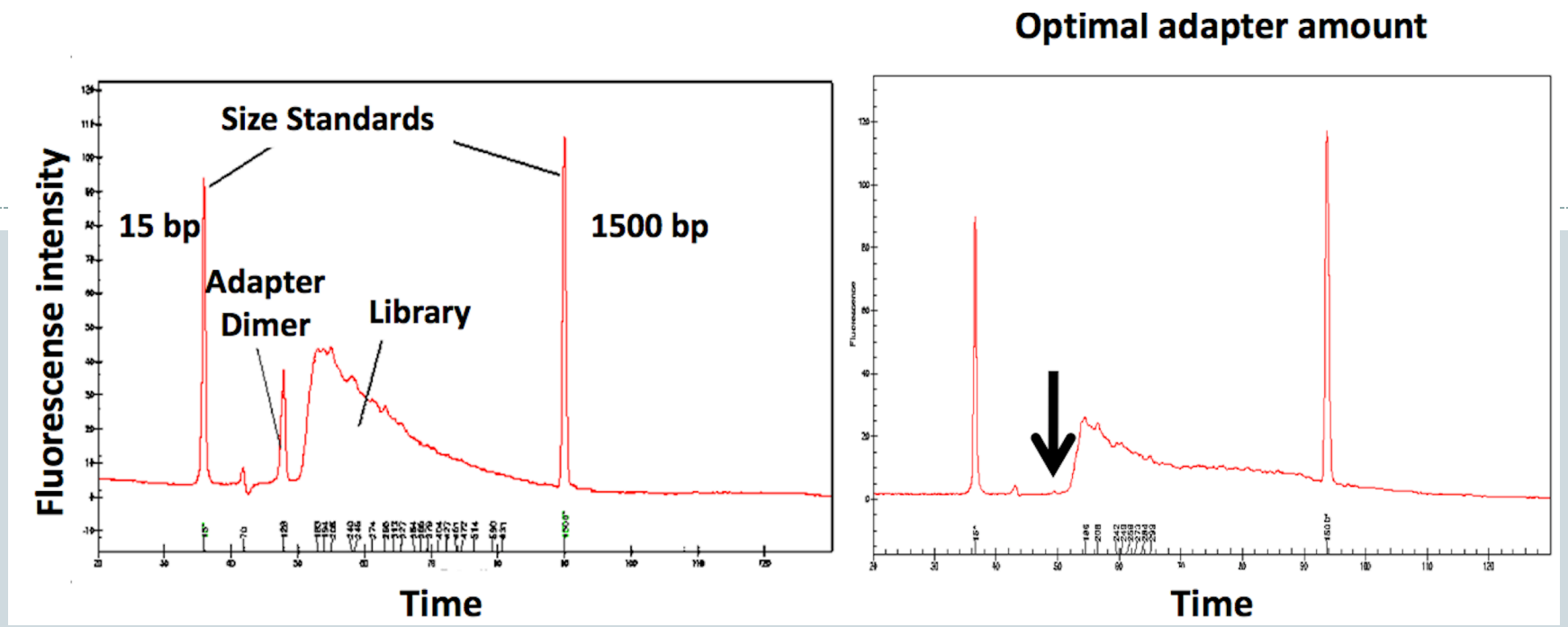
**Optimisation:**

- Restriction digest variation between samples.
- PCR bias
- Variation in size selection

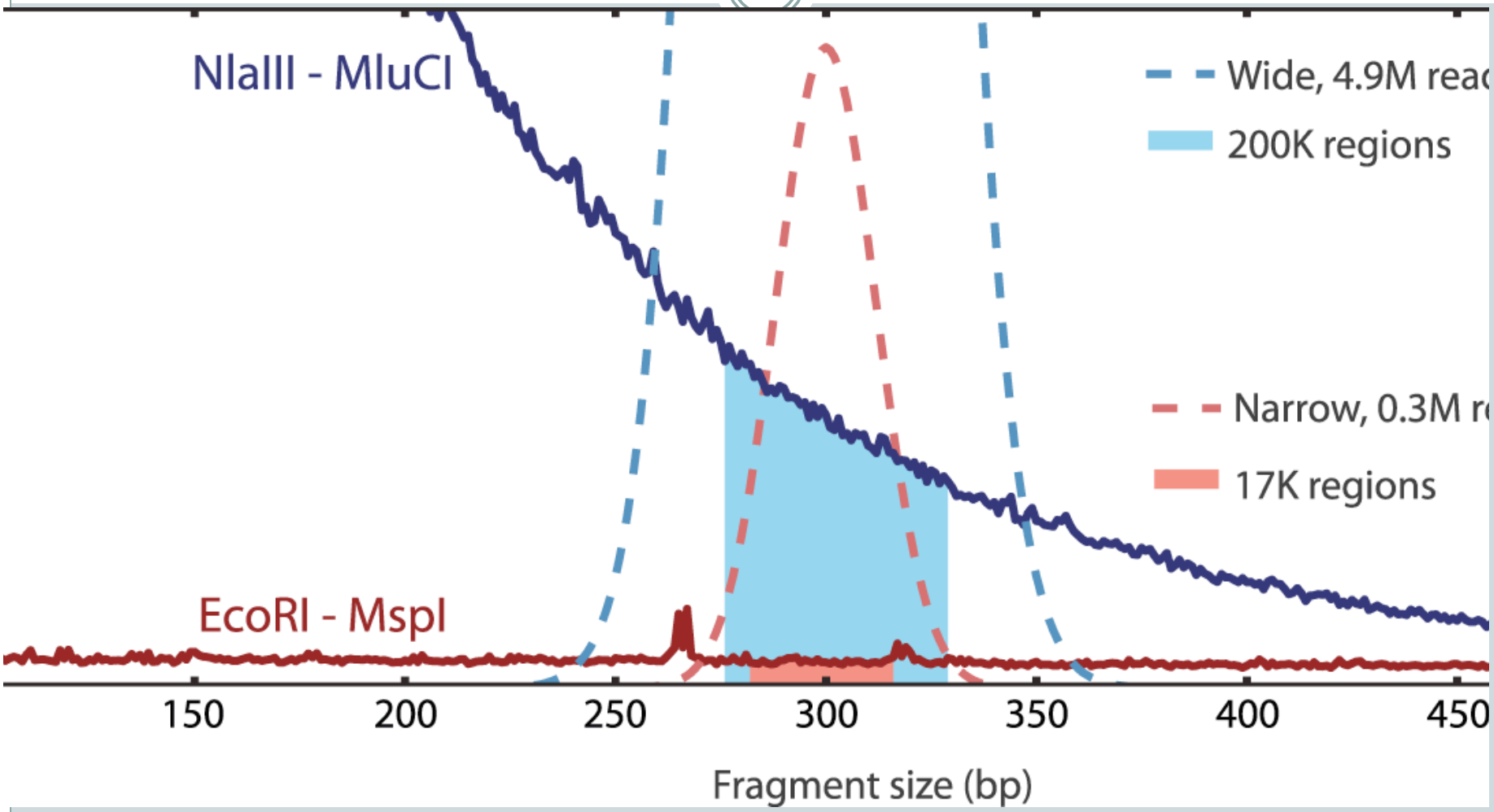

1. DNA digestion
S1
S2
2. Barcoded P1 adapter ligation
S1
S2
3. Pooling samples
4. DNA shearing
5. Size selection
6. Y-shape P2 adapter ligation
7. PCR enrichment
8. Illumina sequencing
100bp SE
MID r.s.


1. DNA digestion
S1
S2
2. Barcoded P1 adapter ligation
S1
S2
3. Pooling samples
4. DNA shearing
5. Size selection
6. Y-shape P2 adapter ligation
7. PCR enrichment
8. Illumina sequencing
100bp SE
MID r.s.

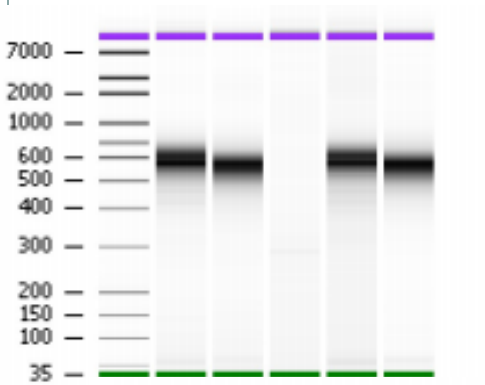

1. DNA digestion
S1
S2
2. Barcoded P1 adapter ligation
S1
S2
3. Pooling samples
4. DNA shearing
5. Size selection
6. Y-shape P2 adapter ligation
7. PCR enrichment
8. Illumina sequencing
100bp SE
MID r.s.

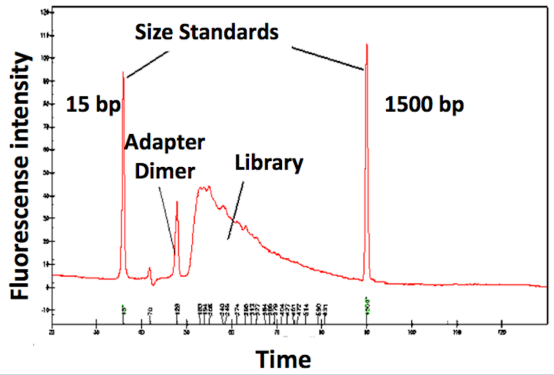

Size Standards
15 bp
1500 bp
Adapter
Dimer
Library
Fluorescense intensity
Time
Optimal adapter amount
Time

# ddRAD – importance of size selection

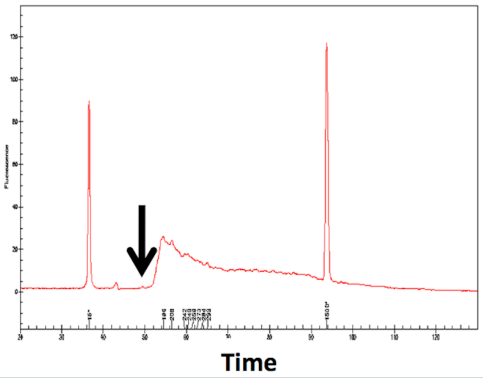



Peterson *et al.* 2012

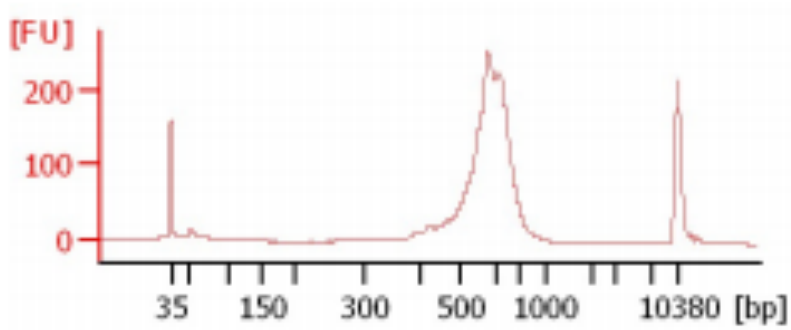


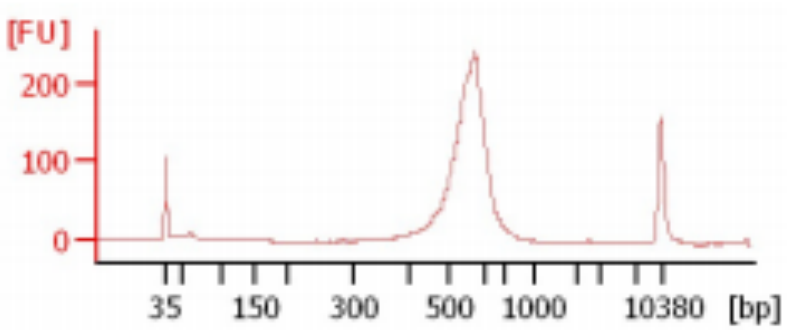

Optimal adapter amount



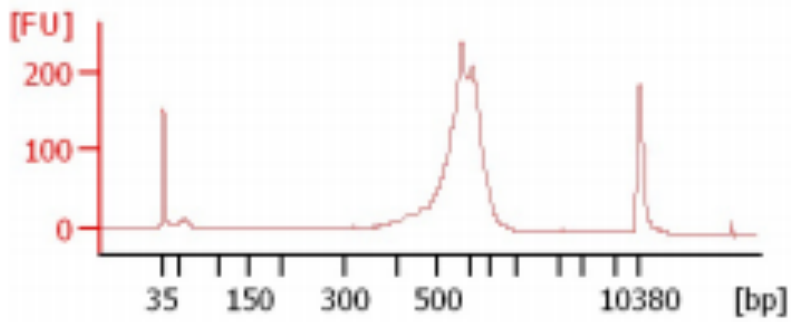

Dsyl 3



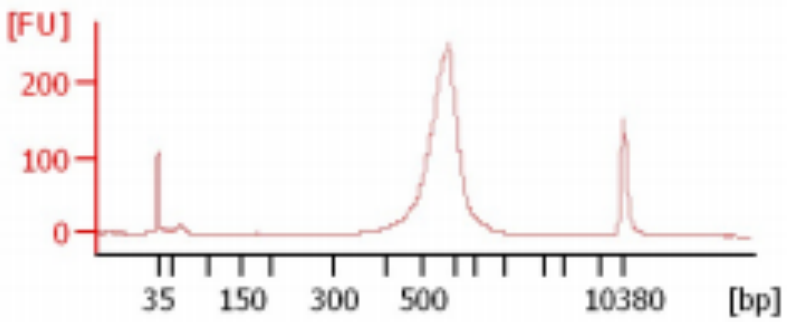

Nep 1

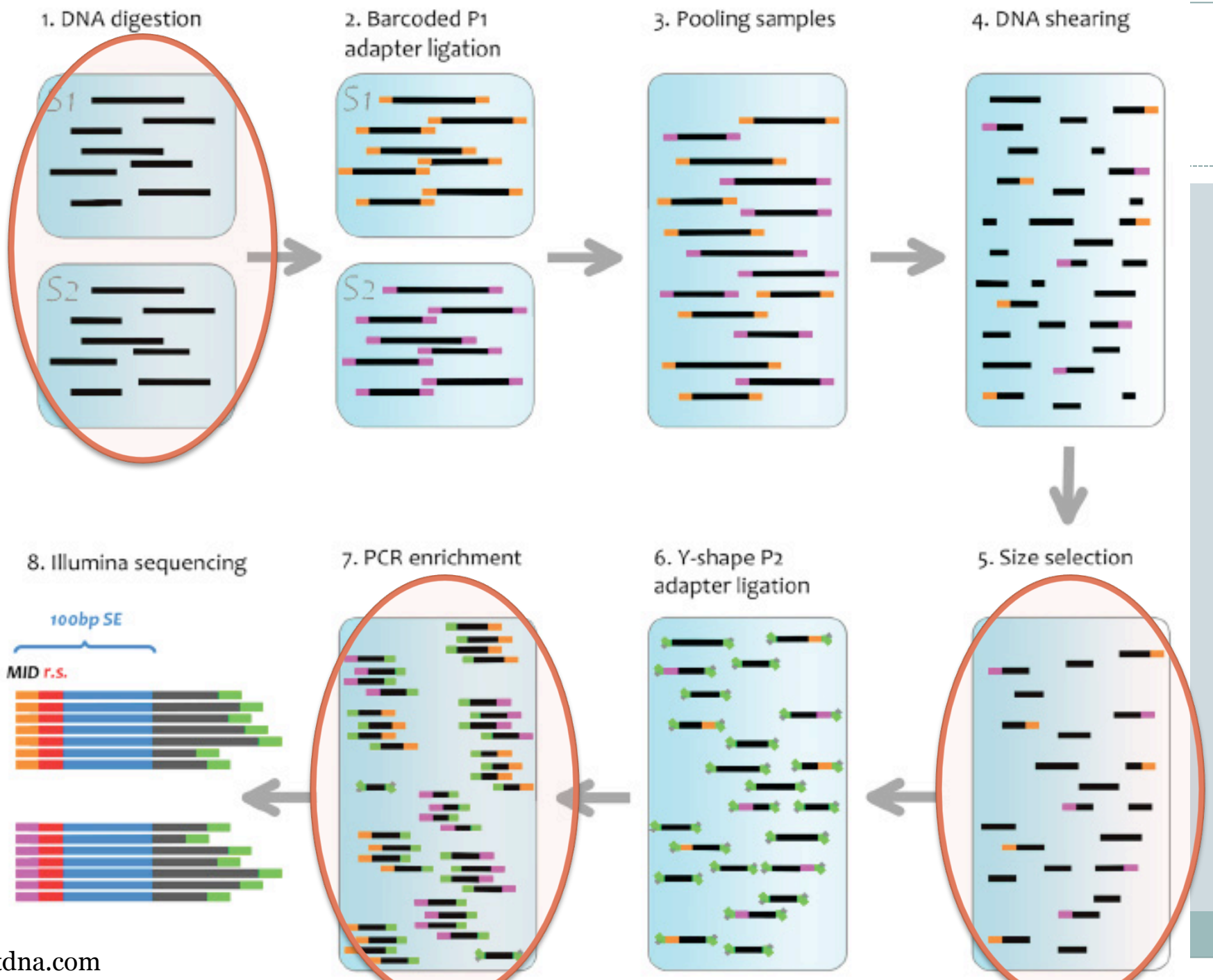



Dsyl 3




Nep 1



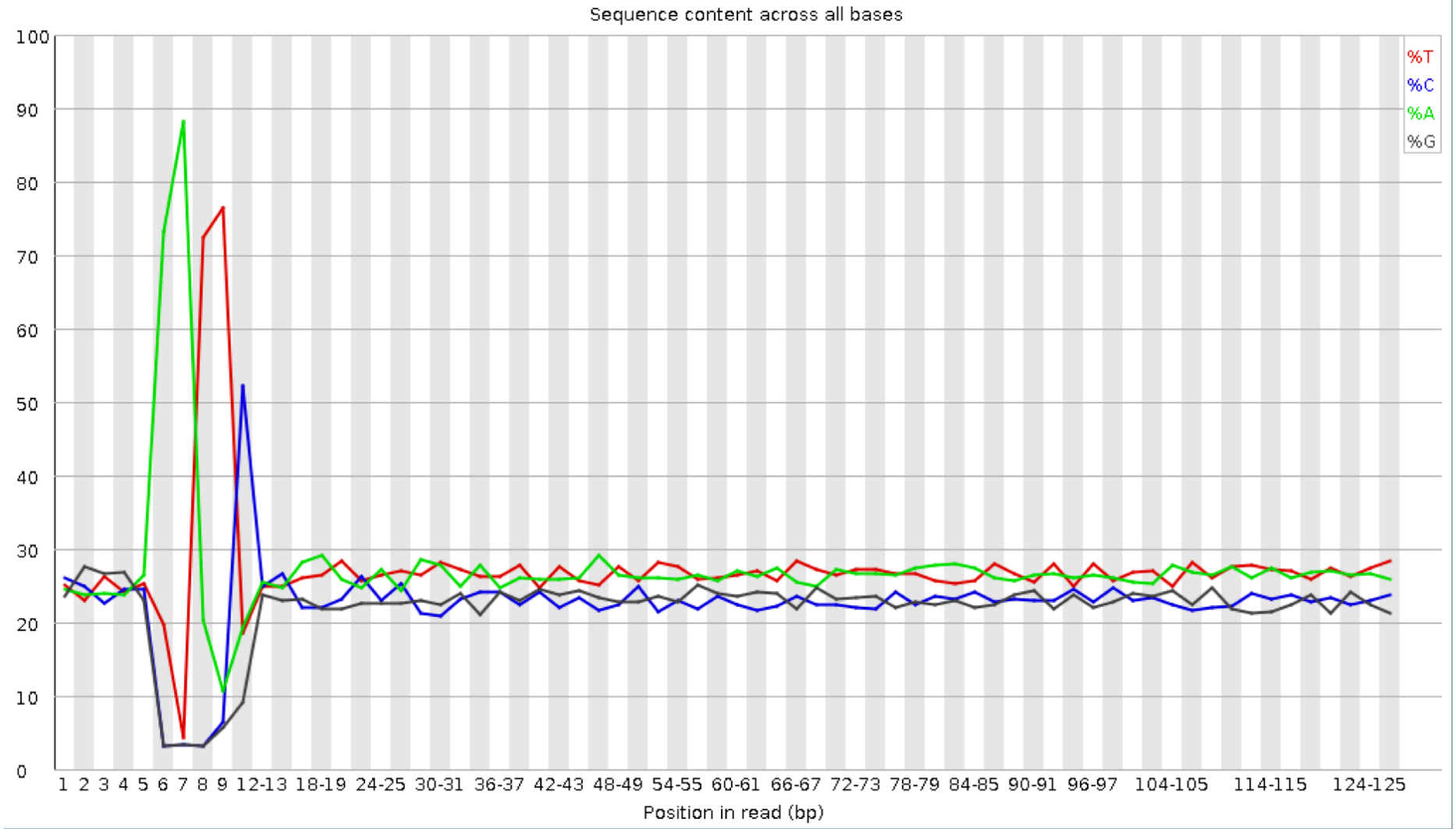

1. DNA digestion
S1
S2
2. Barcoded P1 adapter ligation
S1
S2
3. Pooling samples
4. DNA shearing
5. Size selection
6. Y-shape P2 adapter ligation
7. PCR enrichment
8. Illumina sequencing
100bp SE
MID r.s.

# Lab & Sequencing facilities

FGCZ

- HiSeq  (250Mil reads; 100-150bp)
- MiSeq (100Mil reads; 150-300bp)

# Lab & Sequencing facilities



# Lab & Sequencing facilities

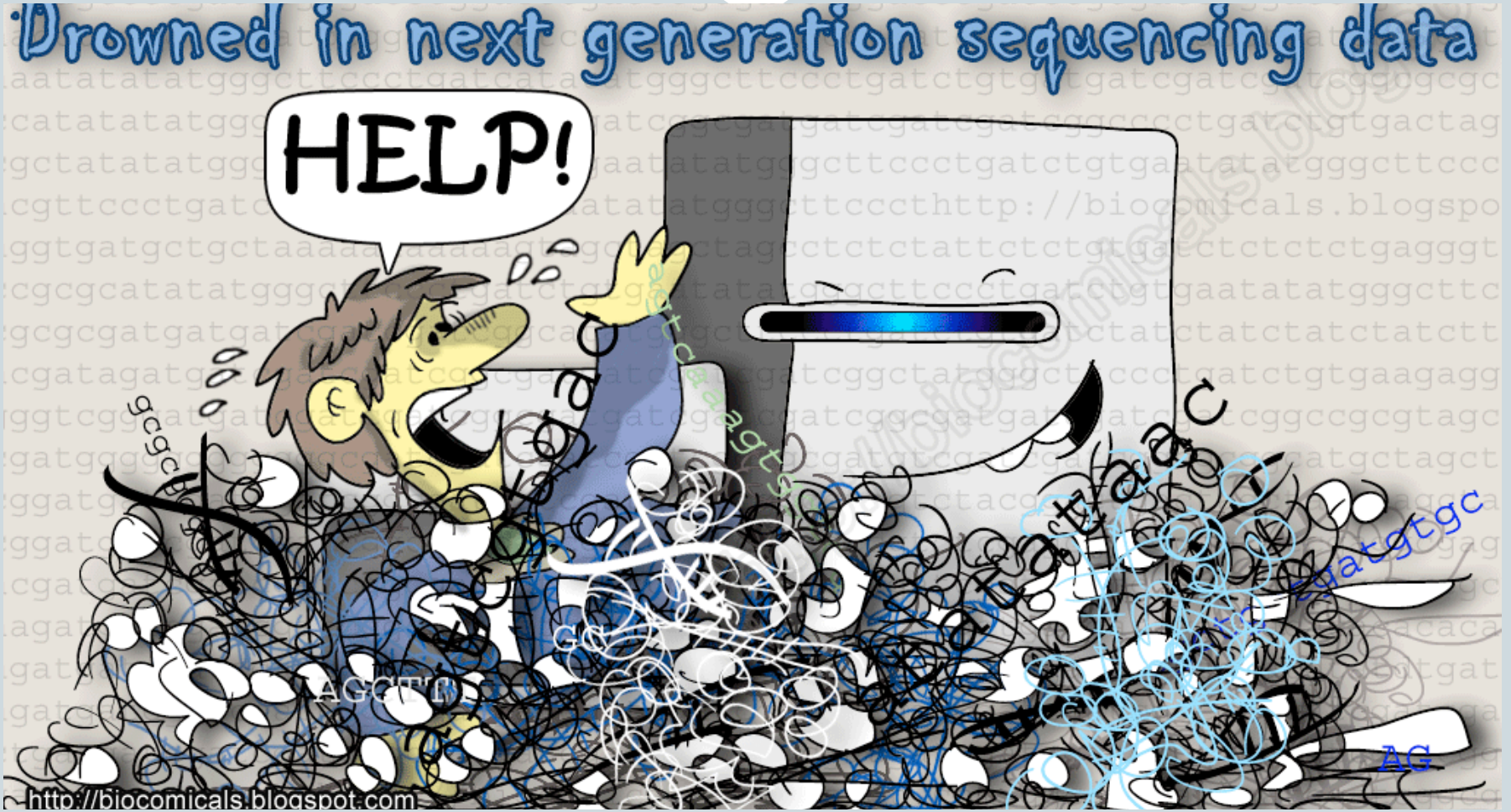
FGCZ

- HiSeq (250Mil reads; 100-150bp)
- MiSeq (100Mil reads; 150-300bp)

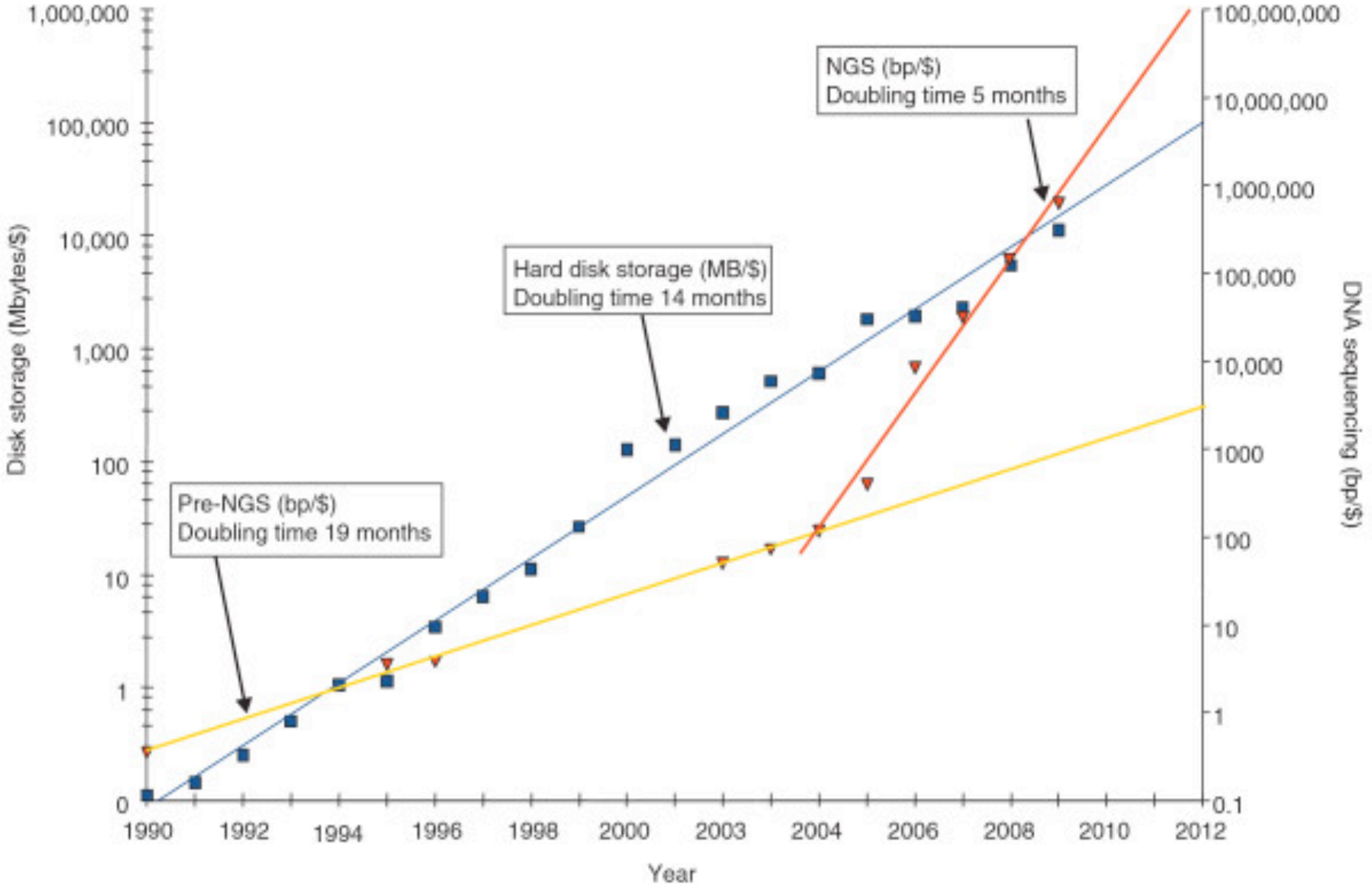
Final nr of reads & coverage:

- -20% for PhiX

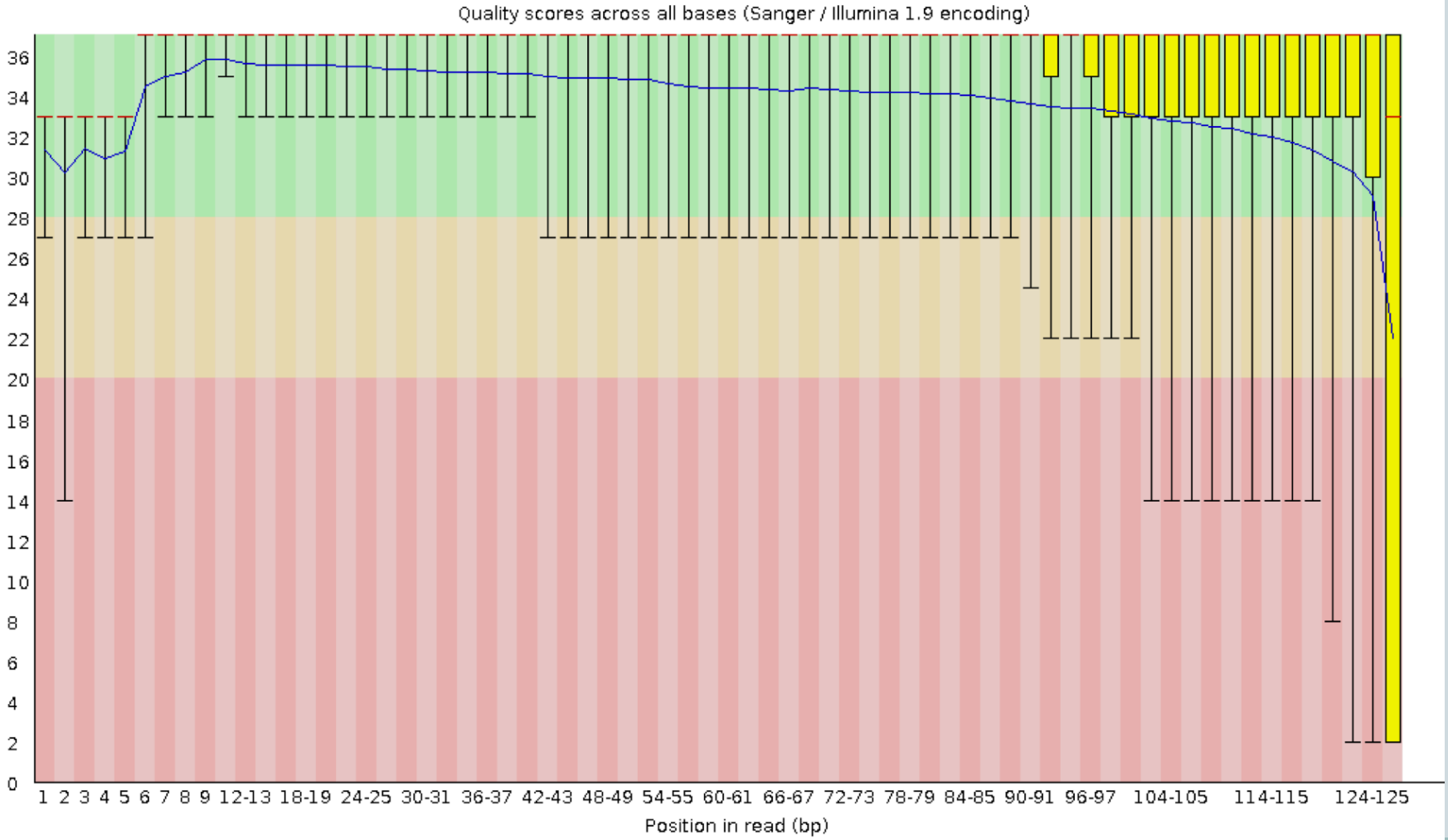
# … 2-6 weeks later



# Large data


NGS (bp/$)
Doubling time 5 months
Hard disk storage (MB/$)
Doubling time 14 months
Pre-NGS (bp/$)
Doubling time 19 months
Disk storage (Mbytes/$)
DNA sequencing (bp/$)
Year
1,000,000
100,000
10,000
1,000
100
10
1
0
100,000,000
10,000,000
1,000,000
100,000
10,000
1000
100
10
1
0.1
1990
1992
1994
1996
1998
2000
2003
2004
2006
2008
2010
2012

# Data!


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97 104-105 114-115 124-125
Position in read (bp)

# RAD bioinformatics pipeline

1. Demultiplex sequences
   - Based on barcode + cutsite + mutation(s)
2. Clean data
   - Poor quality reads
   - Shorter reads
   - Remove adapter dimer
3. Find loci
   1. Within sample clustering
   2. Between sample clustering
   3. Call SNPs
4. Filter SNPs

# RAD bioinformatics pipeline

1. Demultiplex sequences
   - Based on barcode + cutsite + mutation(s)
2. Clean data
   - Poor quality reads
   - Shorter reads
   - Remove adapter dimer
3. Find loci
   1. Within sample clustering
   2. Between sample clustering
   3. Call SNPs
4. Filter SNPs

**Optimise**

**Optimise**

# RAD bioinformatics pipeline

- Stacks (Catchen et al. 2013)
  - First available pipeline
  - Designed for RAD but handles other methods
- dDocent (Puritz et al. 2014)
  - Simple, customisable bash backbone for bioinformatics
  - Designed for ddRAD & ezRAD
  - FreeBayes or GATK
- pyRAD (Eaton 2014)
  - Analysis pipeline written in Python
  - Many different RAD types
  - Clustering using Usearch or Vsearch
- aftrRAD (Sovic et al. 2015)
  - Newest
  - Blend between stacks and pyrad

# 1. Demultiplex data

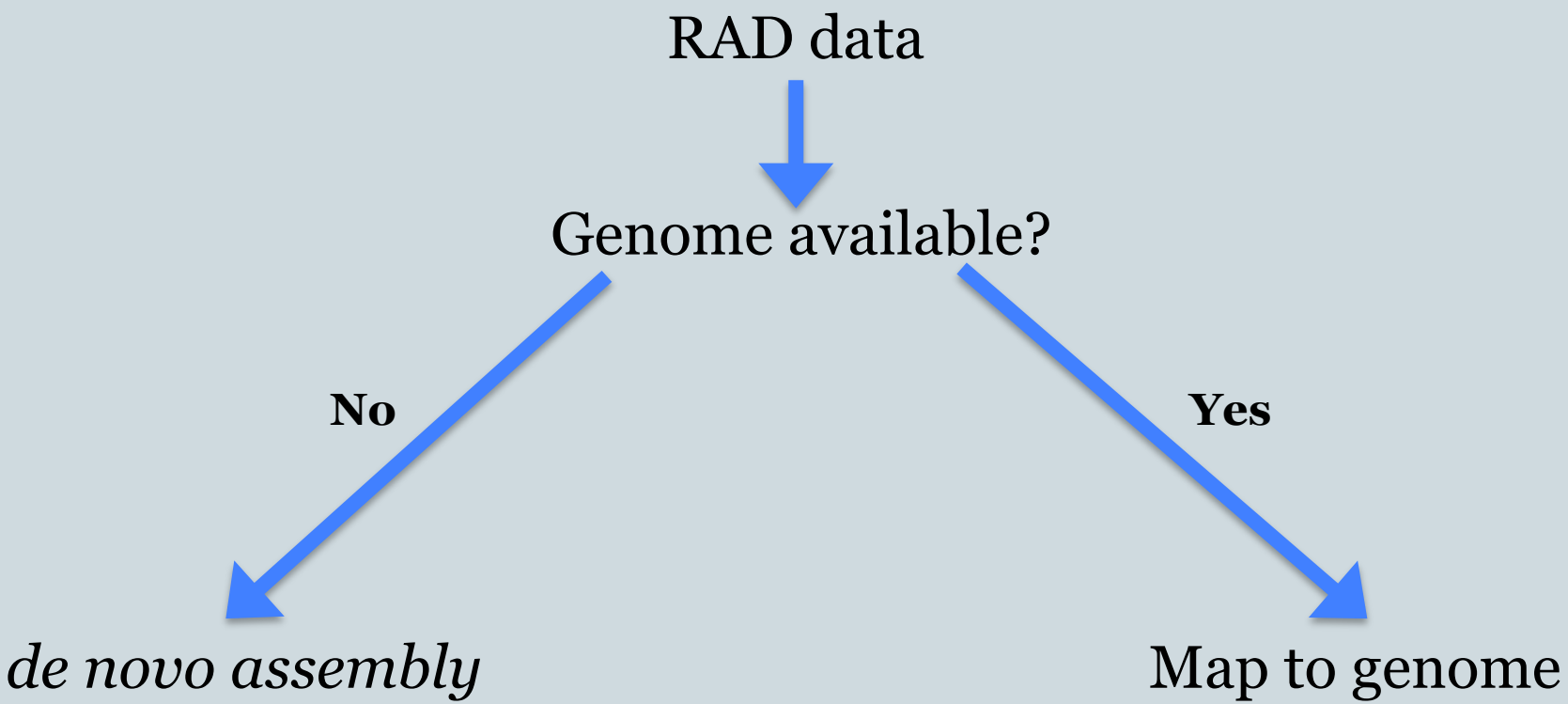
- Barcode OR Barcode + cutsite
- Check your data

CTATTG TGC

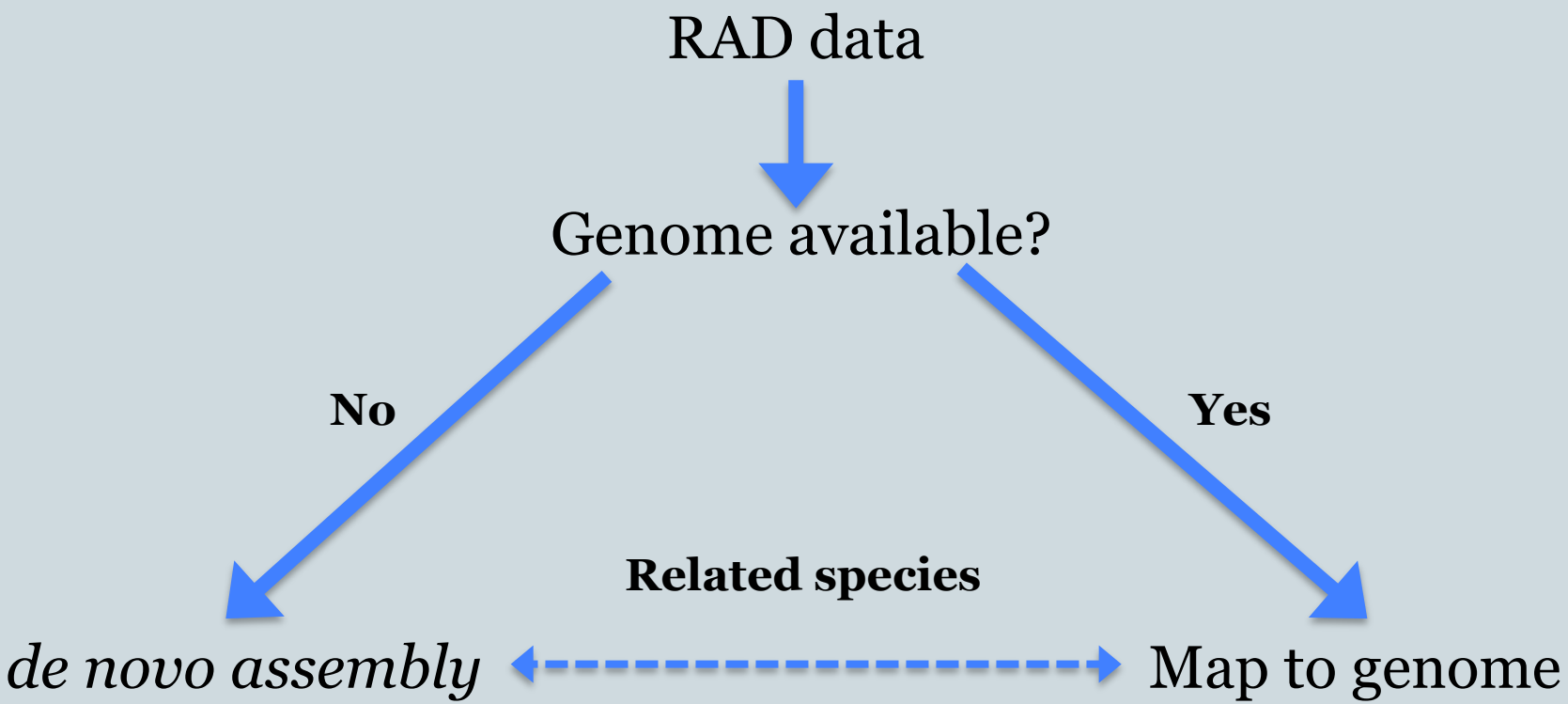
TCTATTGTGCAGGAACCCAGTGTGGACCGGCAGGCAGACCCG
CACGACTGCTTACTGCTAATGGCAGCTCCTGCCCACAGCTCAG
CCCCCGACACAGGCCTTACCAGCTTGTCTCTGAAAGCCCCG

CCTATTGTGCAGGTGGATTATTTACCACTGAGCCACGTGGGA
AGTCCTCTTACACAACAGGCACTCAATAAACATCATCCCTTCA
GCAGACATTTATCAAGCTTTTACTGGGAACCCAACCTGCAT

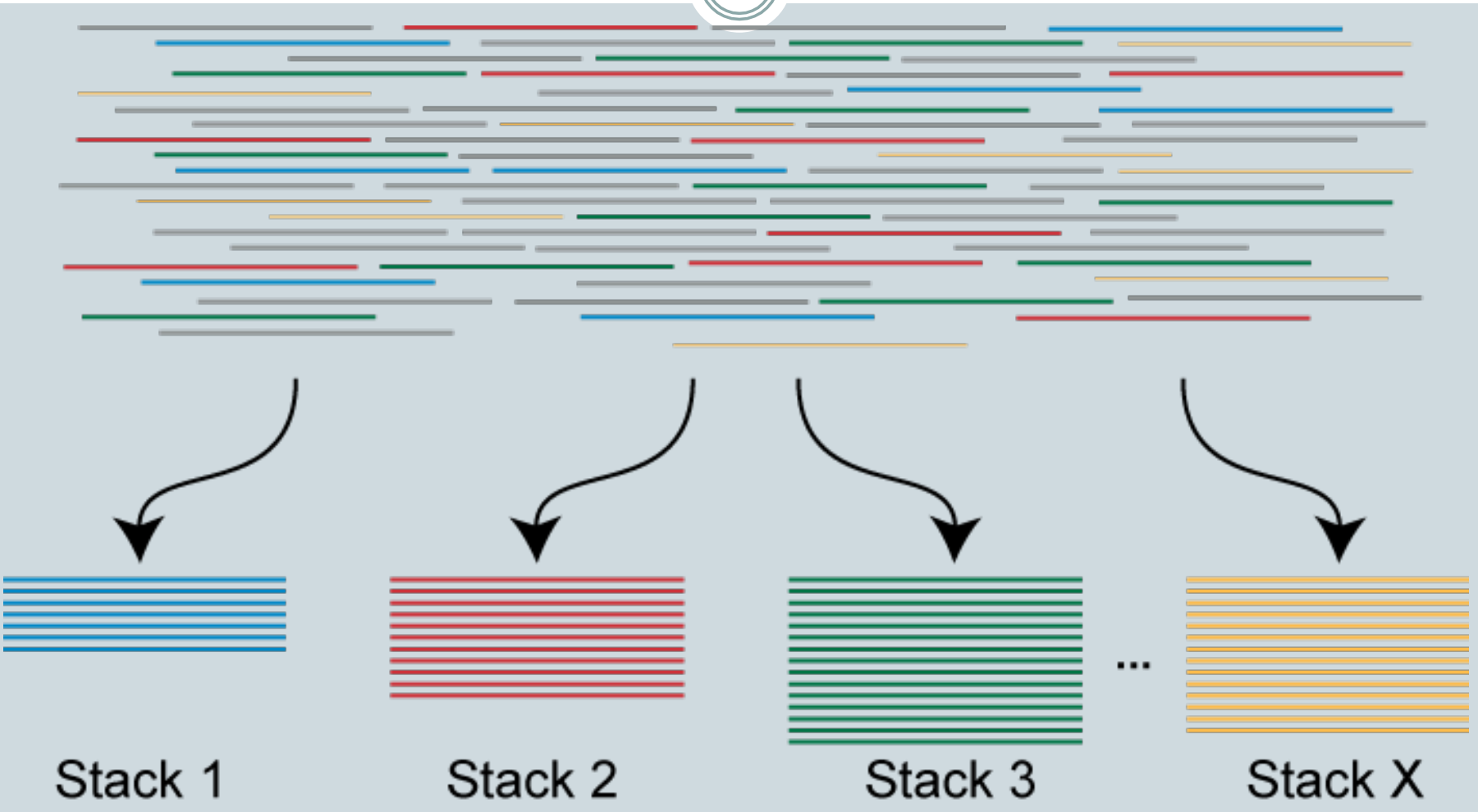
# 2. Find loci


RAD data
Genome available?
No
Yes
*de novo assembly*
Map to genome

# 2. Find loci

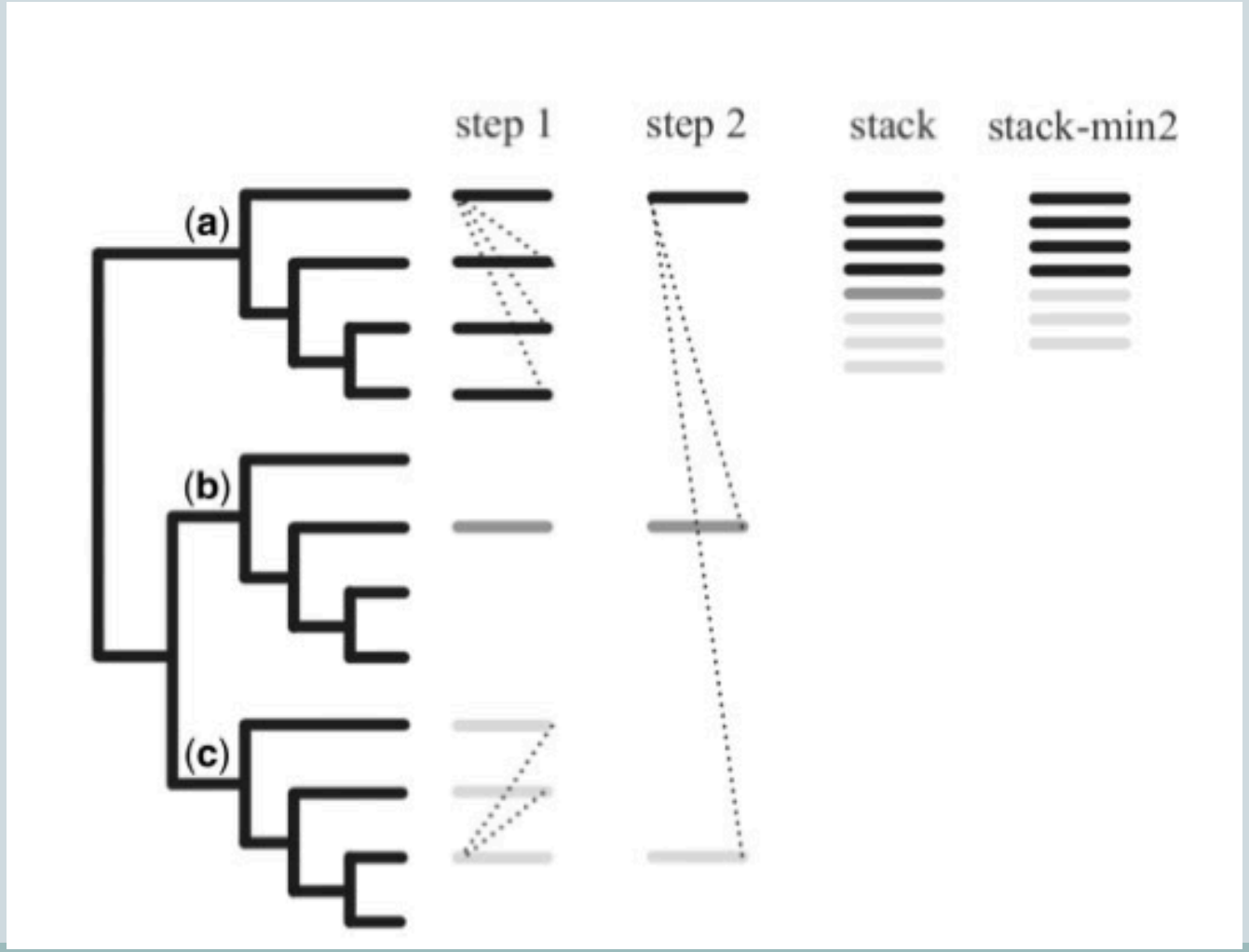

RAD data
Genome available?
No
Yes
Related species
*de novo assembly*
Map to genome

# 2. Find loci: *De Novo* Assembly



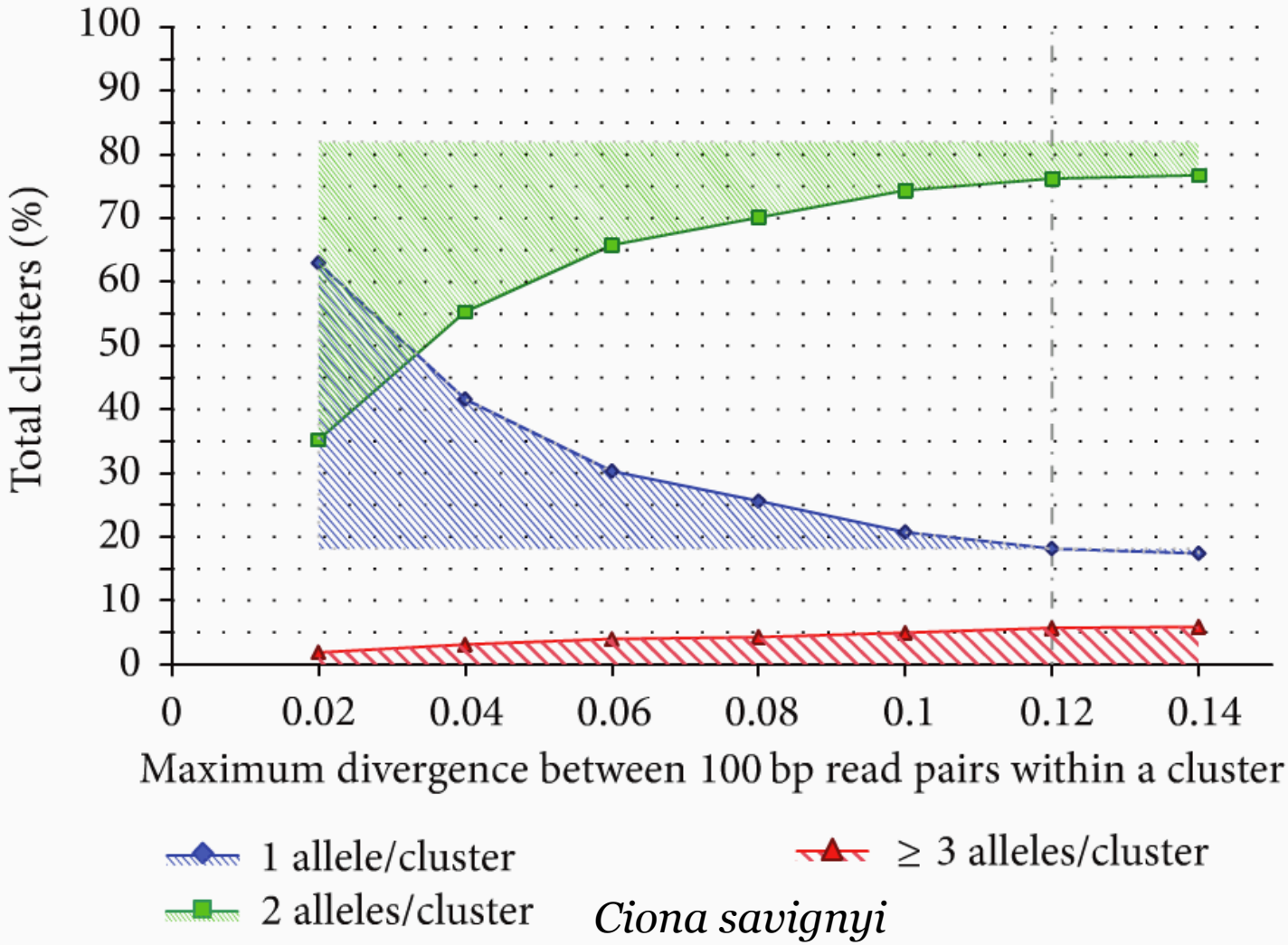

Catchen *et al.* 2011

# 2. Find loci: *De Novo* Assembly




Eaton *et al.* 2014

# 2. Find loci: *De Novo* Assembly




*Ciona savignyi*

Ilut *et al.* 2014

Mastretta_Yanes *et al.* 2015

# 2. Find loci: Map to genome

- Educated guess of read origin

```
CCTGCAGGGATTCC
|||||| || | | ||
CCTGCAGG–ATACC
```

- Global alignment- uses all characters in a read
- Local alignment- maximizes alignment scores through clipping

# 2. Find loci: Call SNPs

- Not all SNP callers are created equal:
  - INDEL sensitivity
  - Base quality awareness
  - Speed
  - Probability of false positives and negatives
  - Biases against homozygotes/heterozygote calls

# 3. SNP filtering

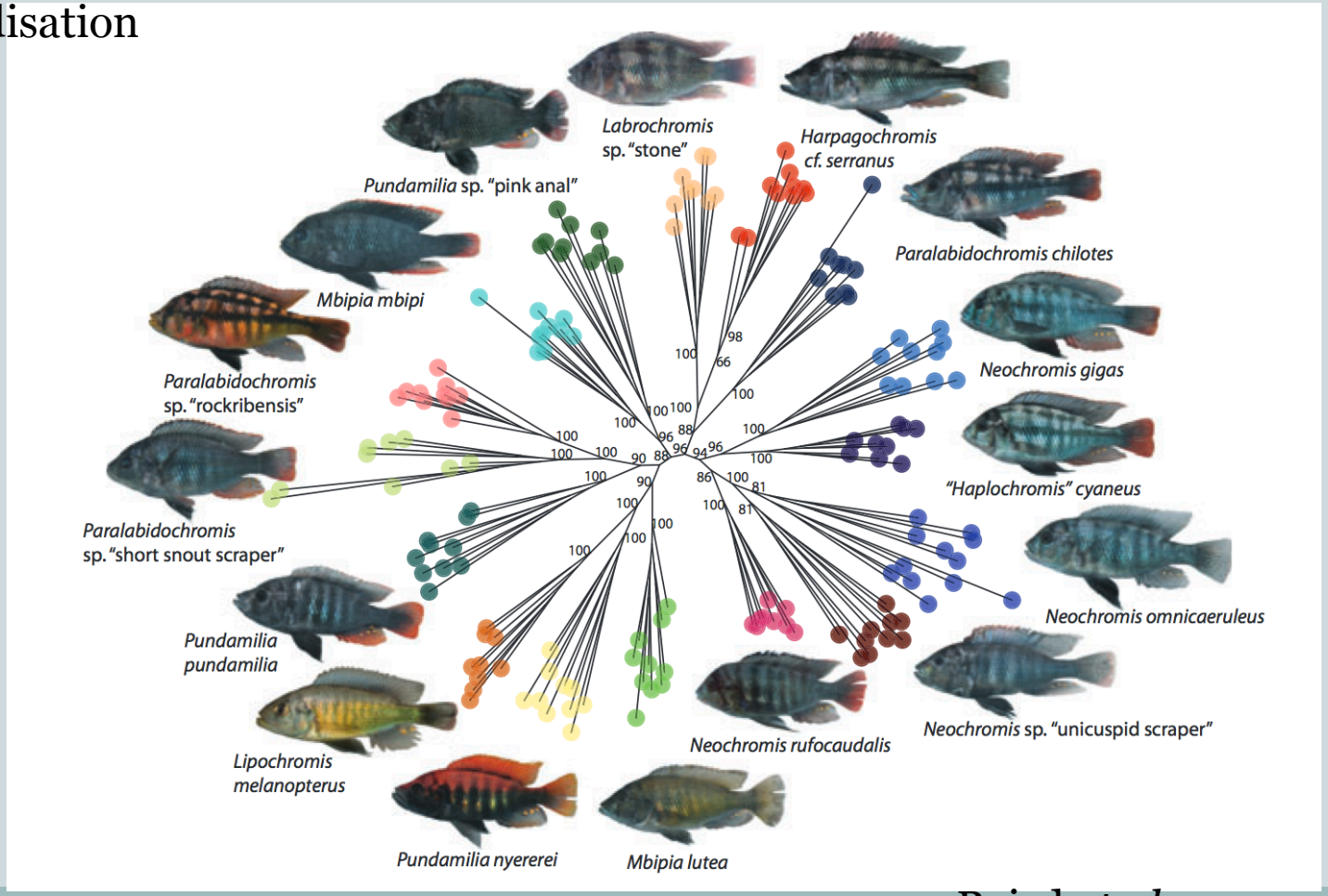
- SNP callers are over sensitive
- Final SNP output: false alleles and allelic dropout
- Filtering these out is essential…
  but the filters have to make sense!
- Filters have to keep in mind
  - Errors from sequencing/ library prep
  - Biological expectations
  - Downstream application of the data

# RAD: Tissue to data

- Thousands of loci
- Non model organism
- Relatively cheap

# Phylogenetics: Species boundaries of Lake Victoria Cichlids

- Radiation <15 000 ya
- Ongoing hybridisation




Baird *et al.* 2012

# RAD: Tissue to data

- Thousands of loci
- Non model organism
- Relatively cheap

- Caveat: Keep sources of error in mind
  - Experimental design
  - Wet Lab
  - Illumina Sequencing
  - SNP calling & filtering

# Thanks for listening!