The Genome Analysis Centre

Building Excellence in Genomics and Computational Bioscience

TGAC
The Genome Analysis Centre™

BBSRC Greater Norwich Development Partnership

Resequencing approaches

Sarah Ayling
Crop Genomics and Diversity
sarah.ayling@tgac.ac.uk





Why re-sequence plants?

TGAC
The Genome Analysis Centre™

To identify genetic diversity:

- Search for alleles
- Identify indels
- Develop molecular markers
- Phylogenetics
- etc…

Sequencing technologies

TGAC
The Genome Analysis Centre™

Illumina HiSeq

Illumina MiSeq

Ion Proton

Oxford Nanopore
MinION Early Access

PacBio RS

Illumina

TGAC
The Genome Analysis Centre™

Illumina HiSeq2500

Hiseq 2500: (High vs Rapid)

- Read length:
  - 2x 125 / 250bp
- Number of reads per flowcell
  - 1.5 billion / 300 million
- Run time
  - 6 days / 40h

Illumina MiSeq

Miseq

- Read length:
  - 2x 300bp
- Number of reads per flowcell
  - 25 million
- Run time
  - 5-55h

Pacific Biosciences RS & Life technologies

TGAC
The Genome Analysis Centre™

PacBio RS
- Read length:
  - Average 8.5kb
- Number of reads per SMRT cell
  - 50,000
- Run time
  - 3h

Ion Proton
- Read length:
  - Average 200bp
- Number of reads
  - 60-80 million
- Run time
  - 2-4h

Oxford Nanopore

TGAC
The Genome Analysis Centre™

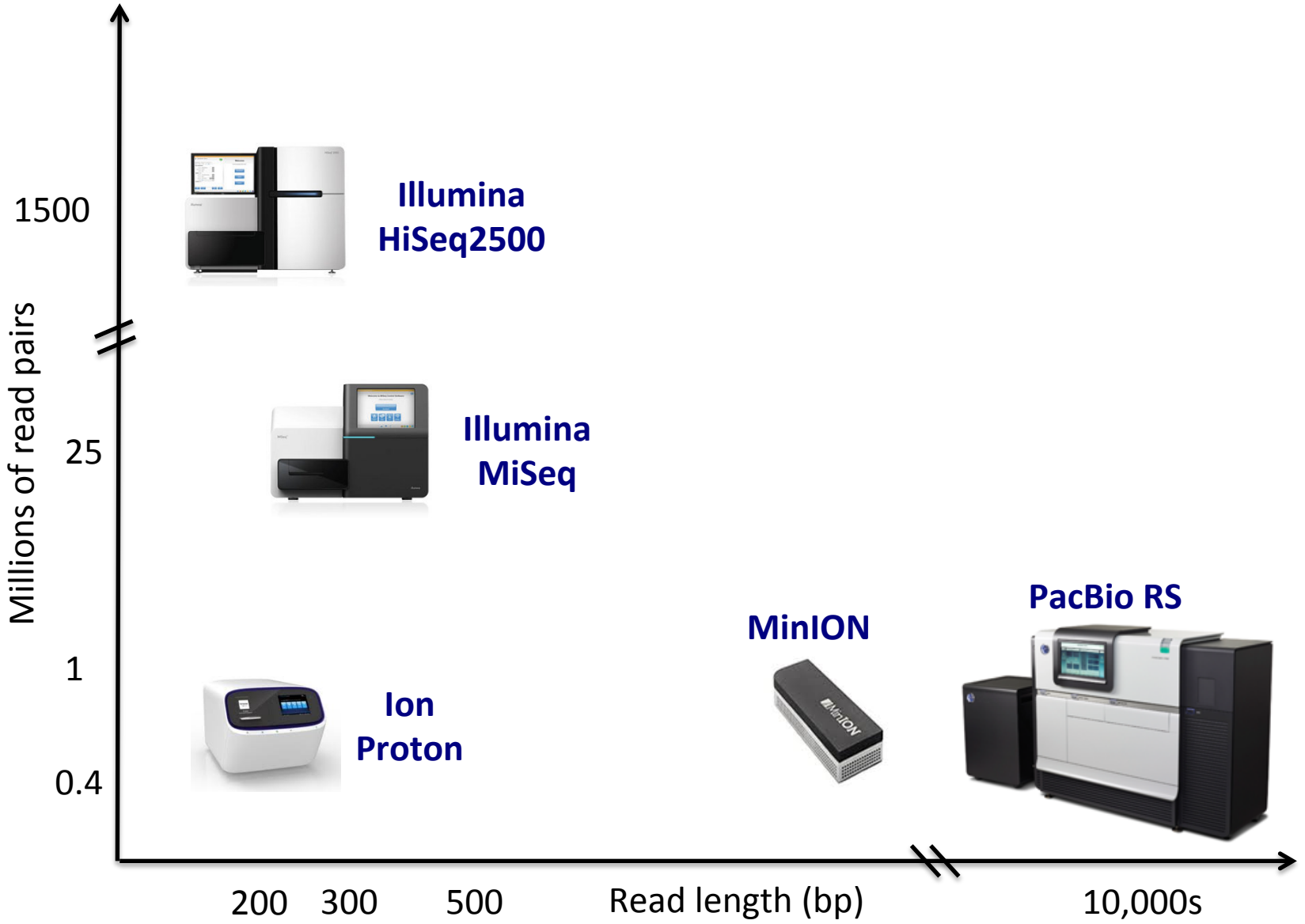
MinION
- Read length:
  - Average 2kb, up to ~100 kb
- Run time
  - Streaming – hours
- Error rate:
  - 37-27% - chemistry improving…
- MinION Access Program

PromethION
- Contains 48 flow cells
- PromethION Early Access Program coming soon…

Laver *et al.,* (2015) Biomolecular Detection and Quantification

Next Generation Sequencing

TGAC
The Genome Analysis Centre™



Trends in sequencing technologies

TGAC
The Genome Analysis Centre™

- **Longer reads**

- **Shorter run times**

- **Cheaper**

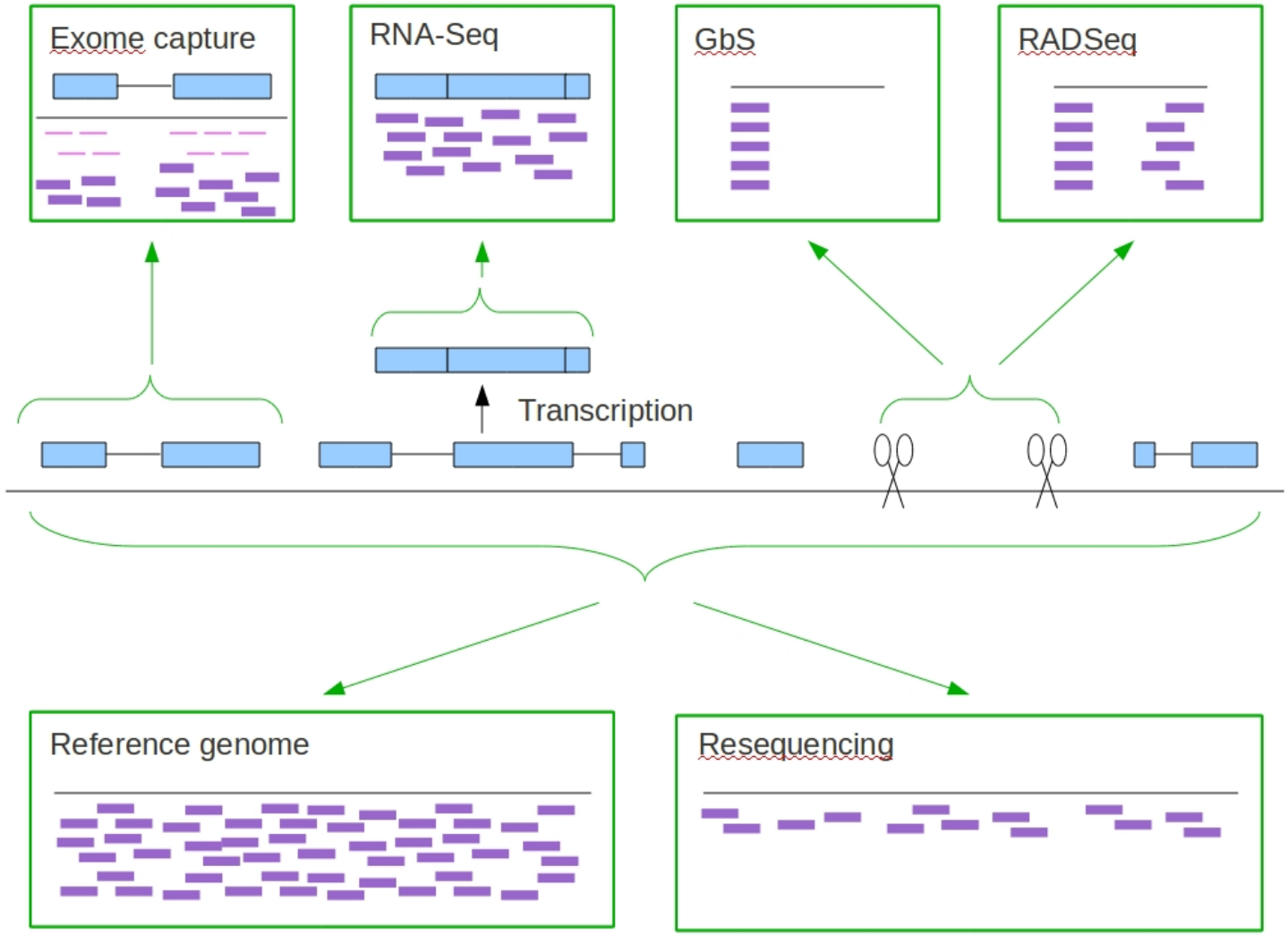
- **Single-molecule sequencing**

What can be (re)sequenced

TGAC
The Genome Analysis Centre™

Size Reduction

Whole genome



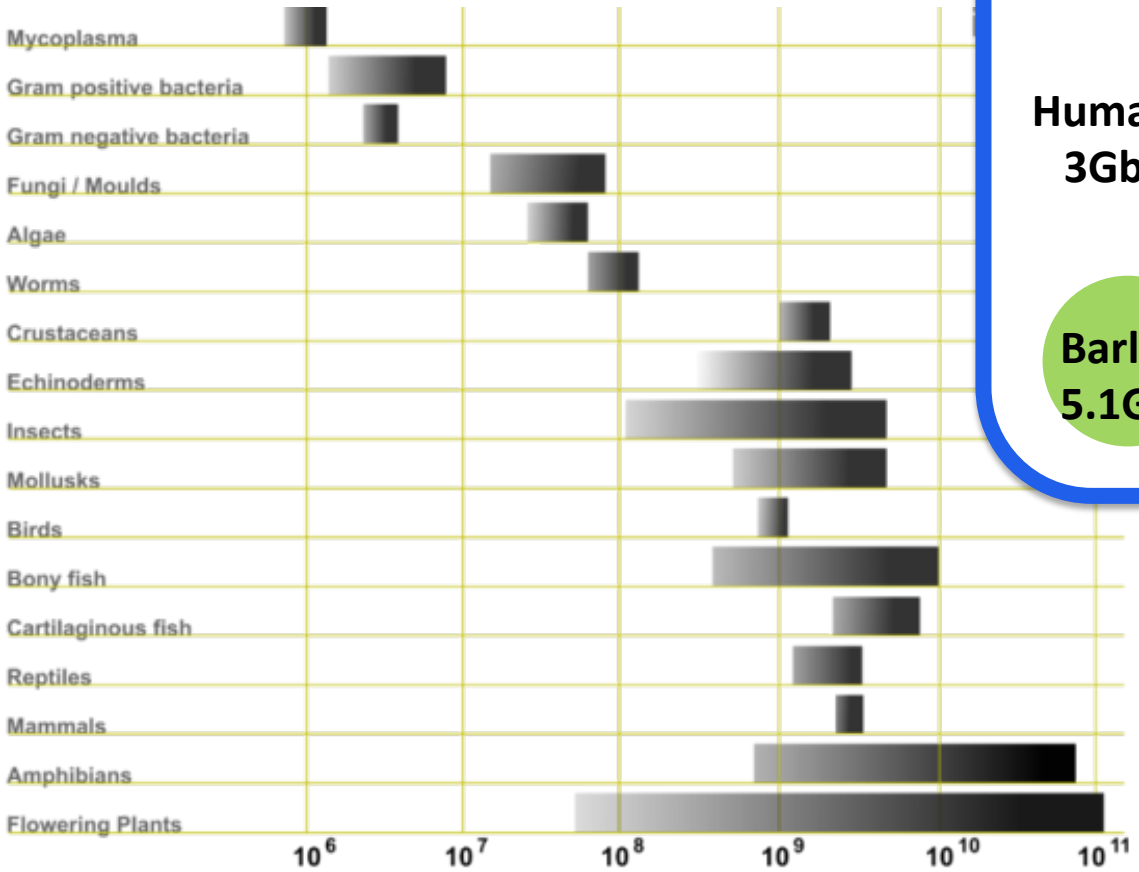
Which strategy?

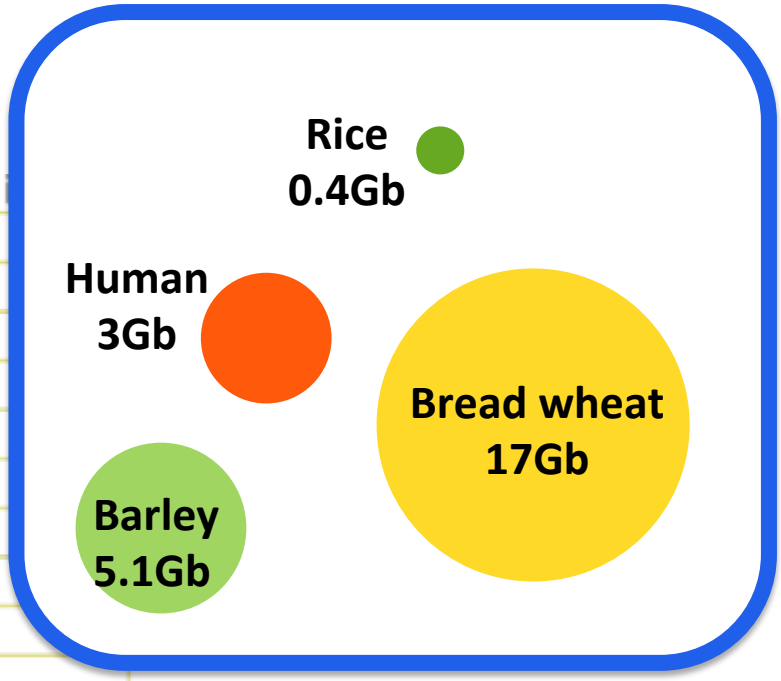
TGAC
The Genome Analysis Centre™

- Depends on:
  - Your objective
  - Your species
  - Your analysis capacity
  - Your budget!

Genome sizes

TGAC
The Genome Analysis Centre™

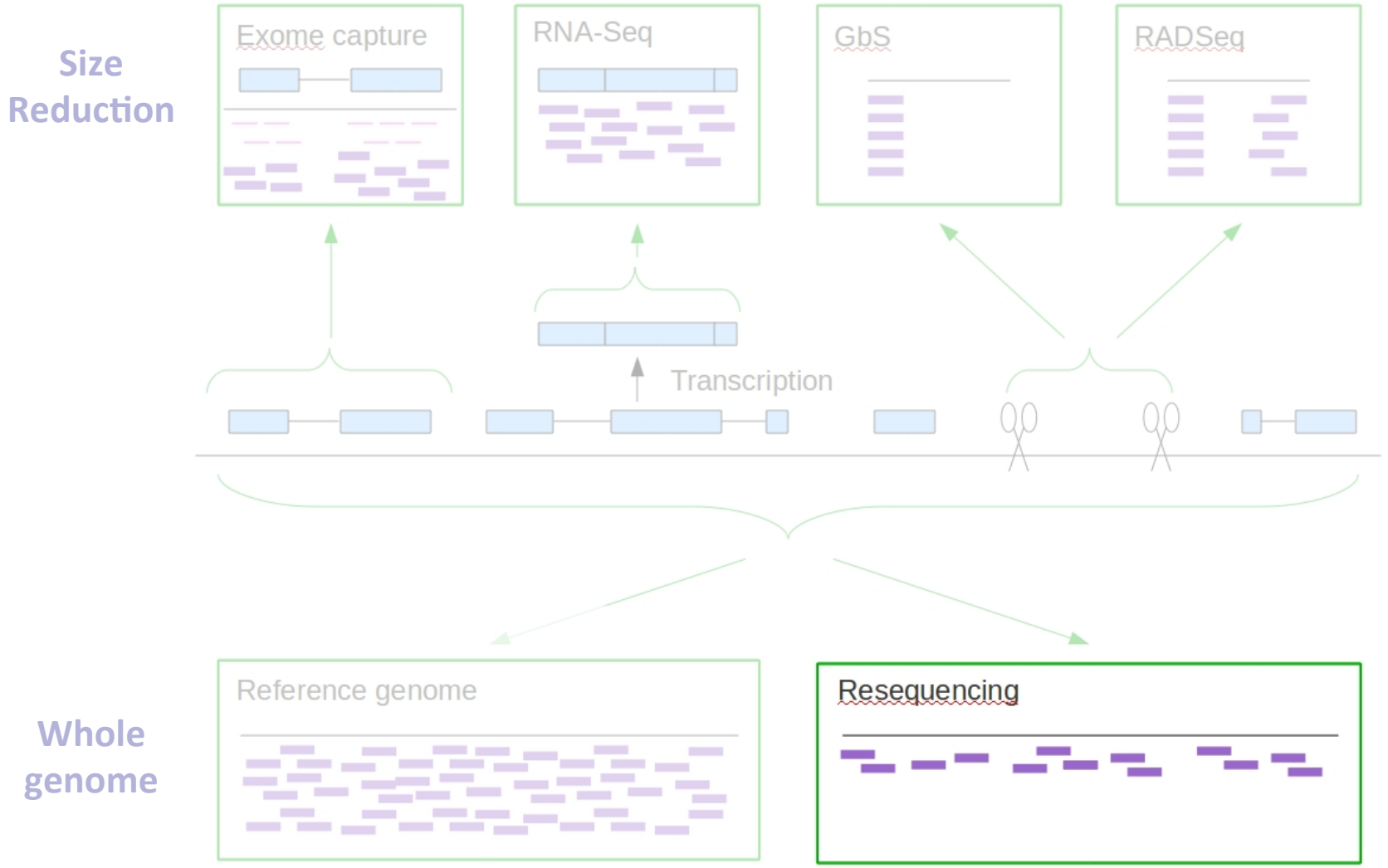




Pic: Abizar at Wikipedia

Sequencing strategies

TGAC
The Genome Analysis Centre™



Resequencing pipeline

TGAC
The Genome Analysis Centre™

Generation of Sequence

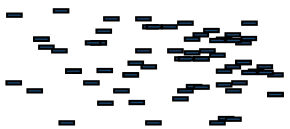
Mapping Reads

Variant calling (SNPs, SVs)

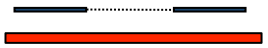
Association

Annotation

NGS Sequencing



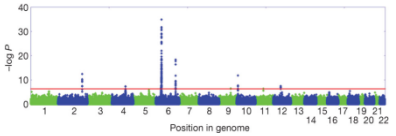
Bowtie
BWA etc



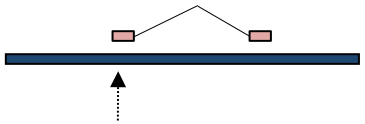
GATK
SAM tools
FreeBayes

```
ACGTCTCGCT C GCTCGCTCTCGCTCG
  GTCTCGCT A GCTCGCT
    CTCGCT A GCTCGCT
```

GWAS



Causal Variant



SNP A to C

Whole genome resequencing

TGAC
The Genome Analysis Centre™

Advantages:

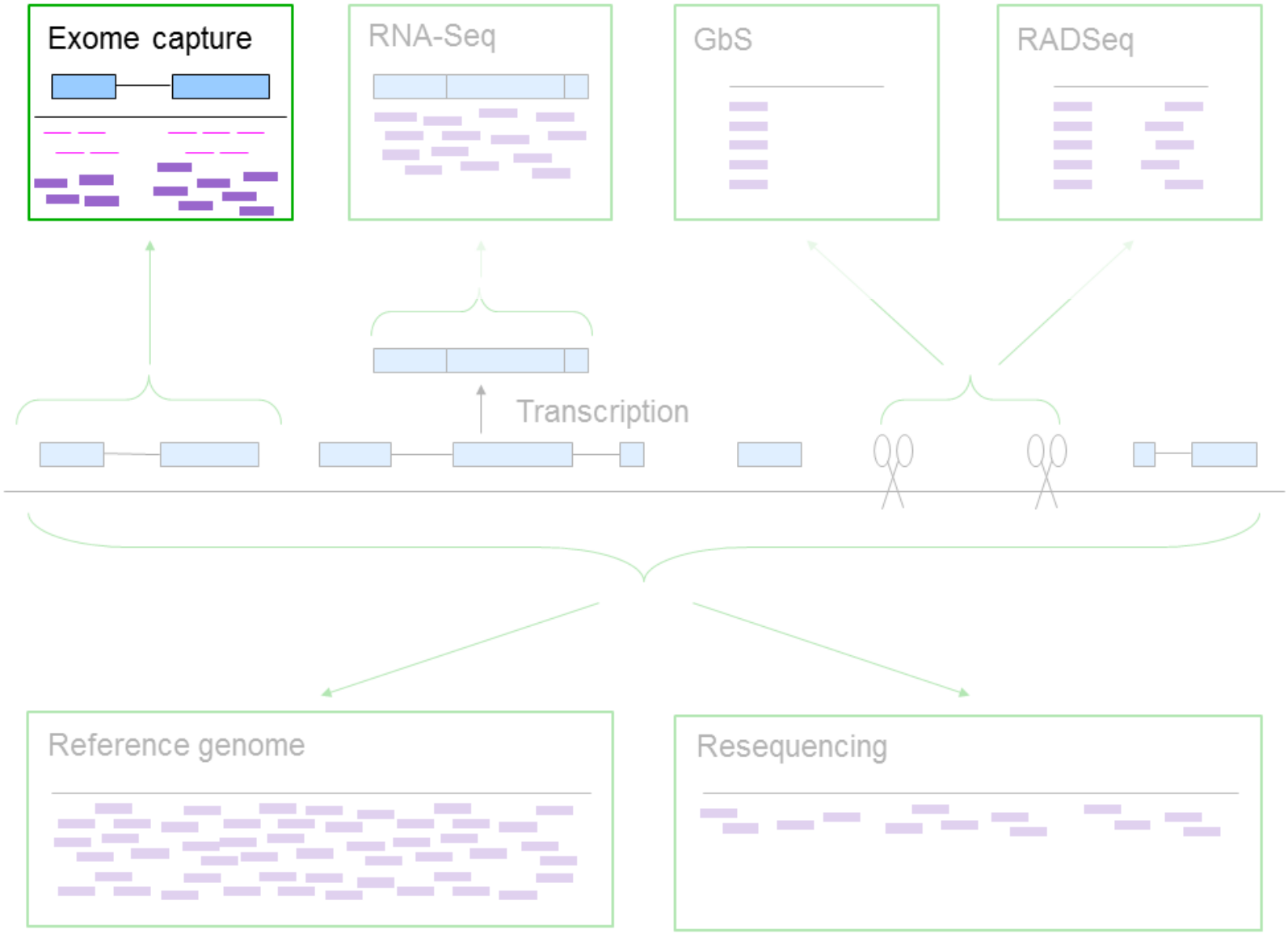
- Easy sample prep
- Whole genome resequenced
  - including sample-specific regions (depth…)

Disadvantages

- Can be costly for large genomes
- Missing data (depending on coverage)
- Large amounts of data to handle
  - Align to genome? Novel regions?
- Overkill?

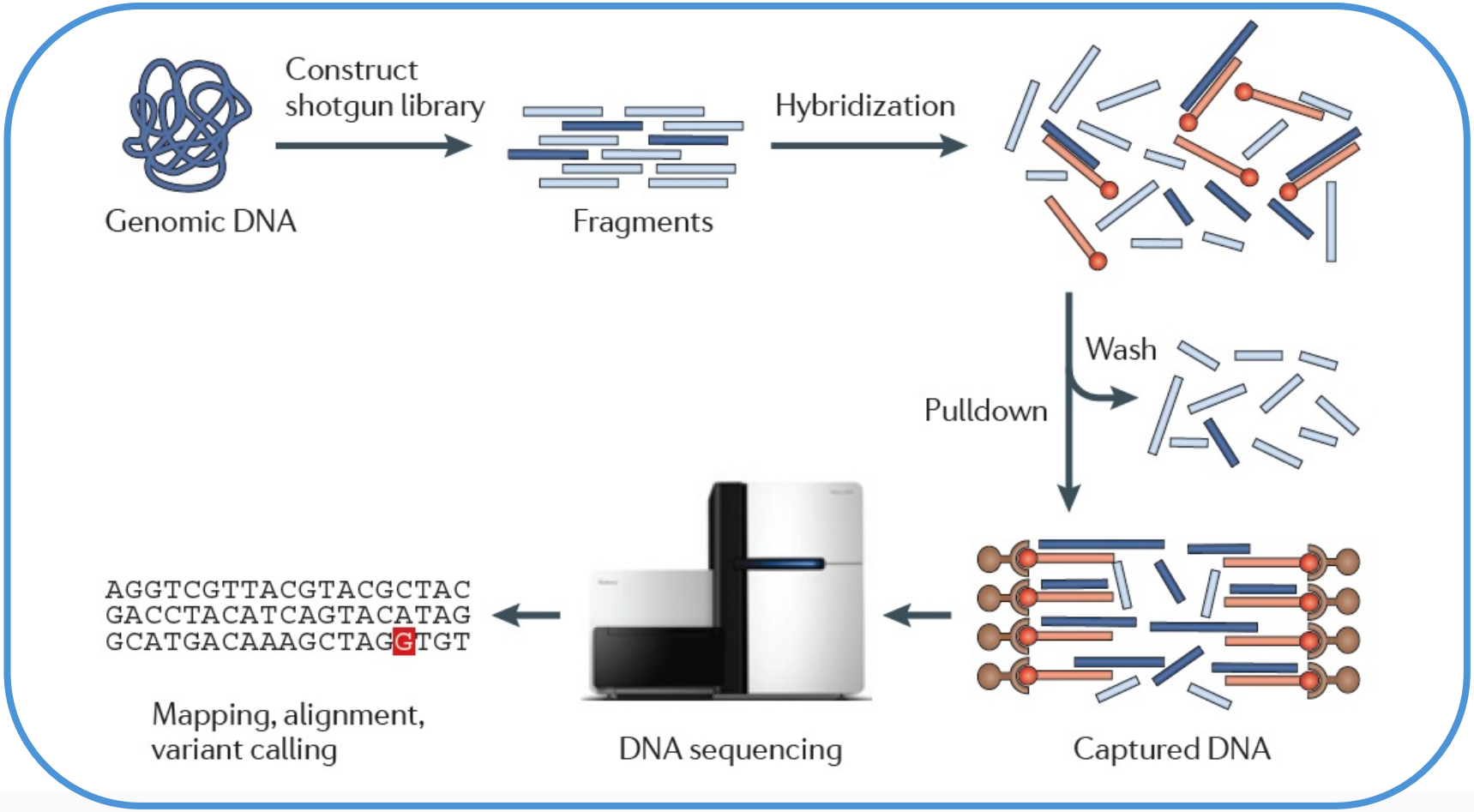
Sequencing strategies

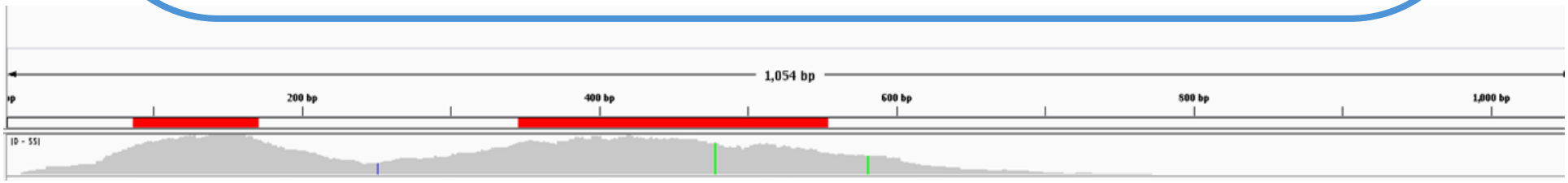
TGAC
The Genome Analysis Centre™



Sequence/exome capture

TGAC
The Genome Analysis Centre™





Bamshad et al., 2011 Nature Genetics

Exome capture

TGAC
The Genome Analysis Centre™

Advantages:

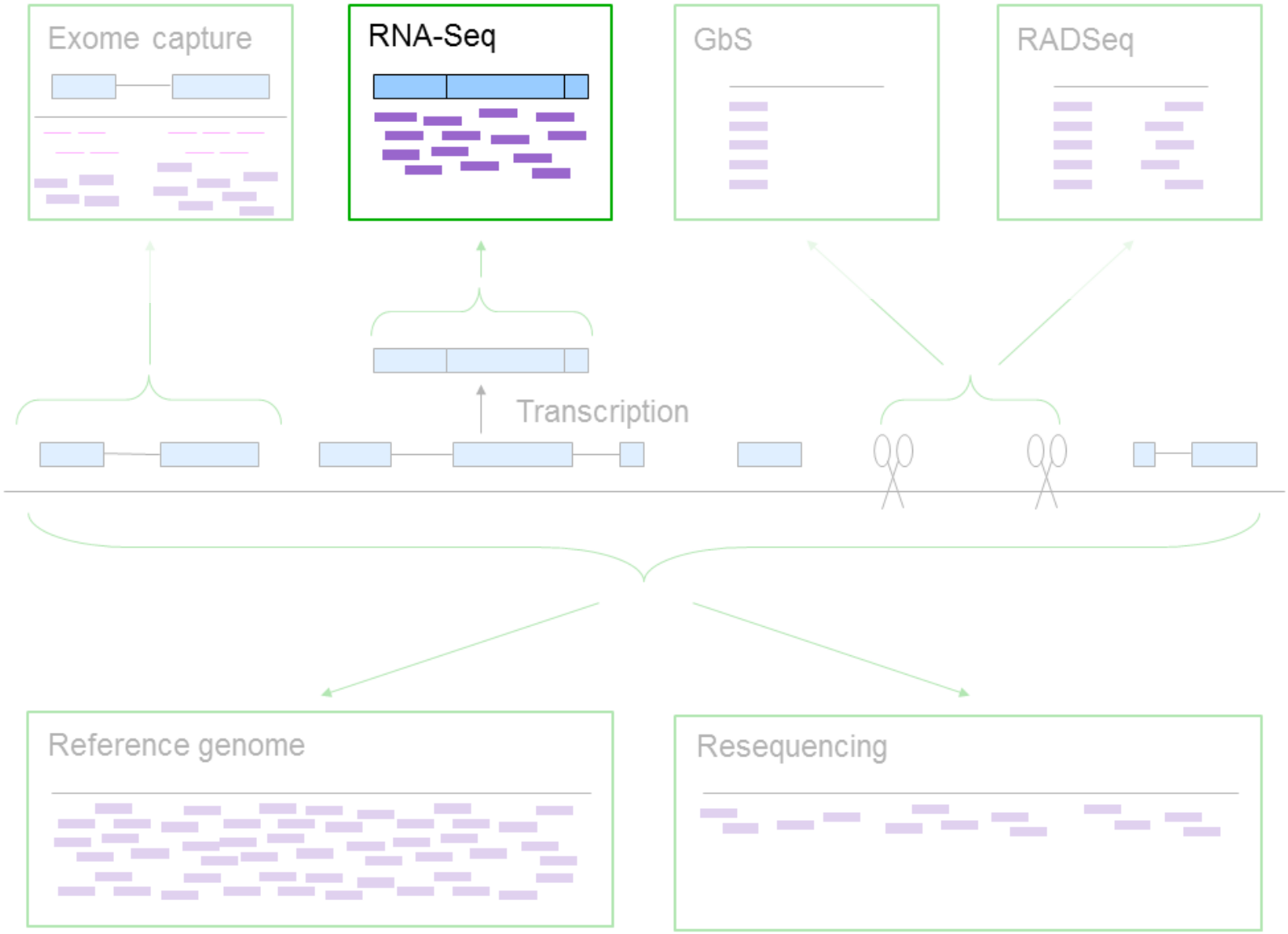
- Only targets regions of interest
- Reduced sequencing costs - multiplexing
- Less data to handle

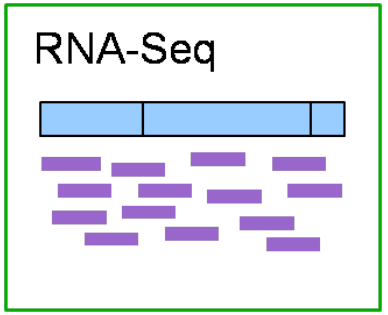
Disadvantages

- Need to know regions of interest beforehand
- Additional price of capture (may vary)
- More complicated library prep
- Will miss novel regions

Sequencing strategies

TGAC
The Genome Analysis Centre™





- Sequence mRNA
  - PolyA pulldown or ribo-depletion
- Reads proportional to expression
- Strand-specific protocols
- Splice variants
- Splice-aware aligners/assemblers required
  - Reference guided
  - *De novo*

Associative Transcriptomics

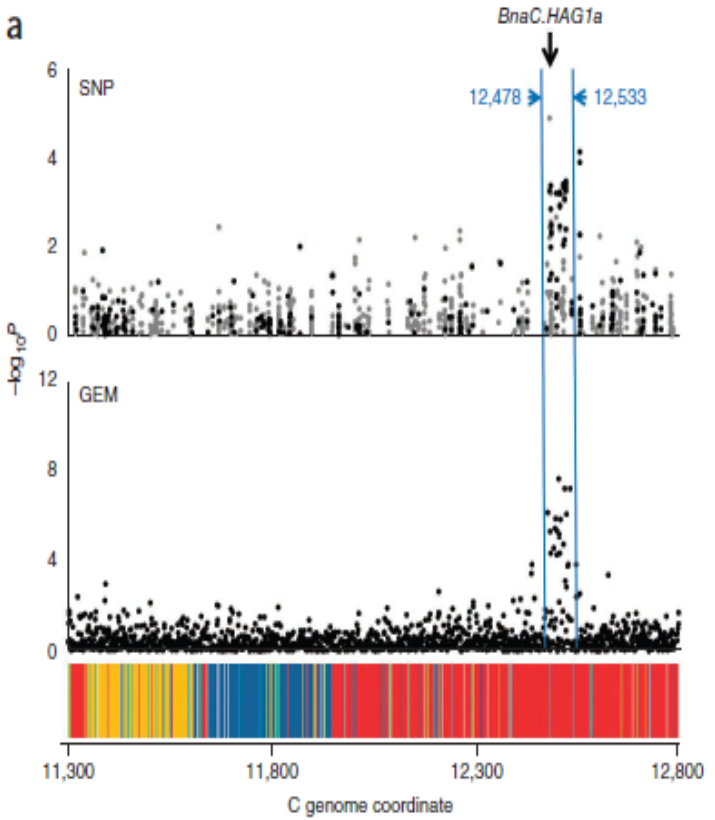
TGAC
The Genome Analysis Centre™

Map mRNAseq reads to 61,613 anchored unigene sequences

Call SNPs and transcript quantification values

Perform GWAS

Transcription factor *HAG1* gene family as a candidate in the quantitative control of glucosinolate content of rapeseed



Harper *et al.,* (2012) Nature Biotech.

Advantages:

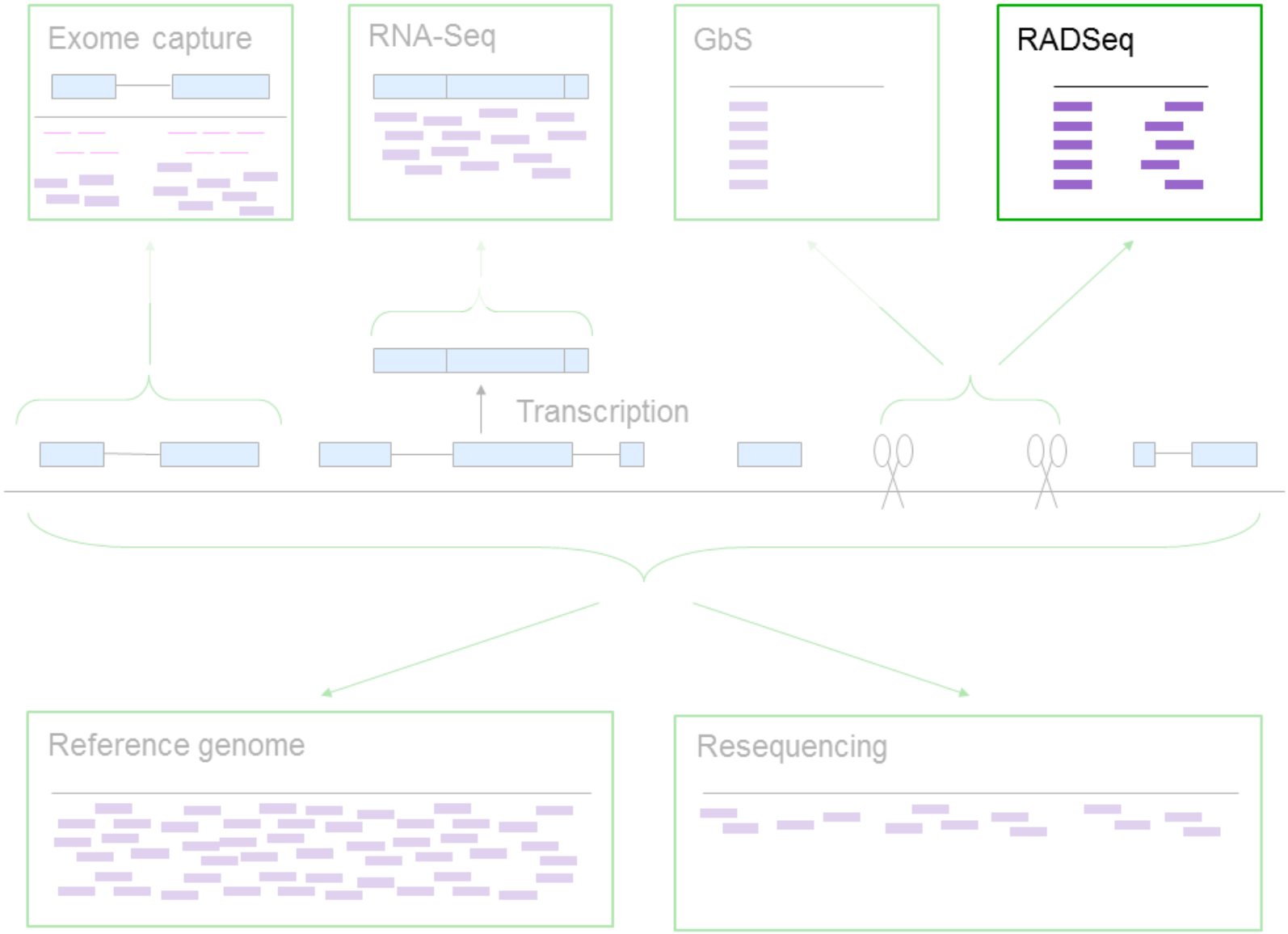
- Only targets expressed loci
- Widely used
- Reduced sequencing costs compared to genomic
- Less data to handle
- Can use variation and expression levels

Disadvantages

- Lowly / unexpressed transcripts will be missed
- RNA more difficult to ship

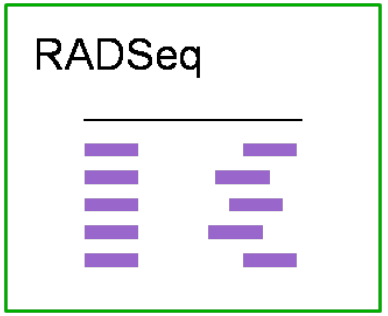
Sequencing strategies

TGAC
The Genome Analysis Centre™

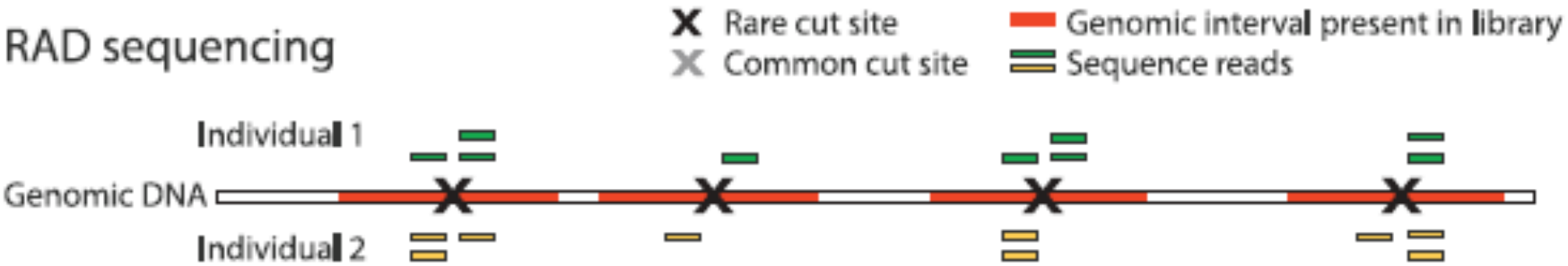


RADSeq - classic

TGAC
The Genome Analysis Centre™



- Enzyme restriction, mechanical shearing + broad size selection
- Identify loci based on read stacks
- Can assemble sheared end

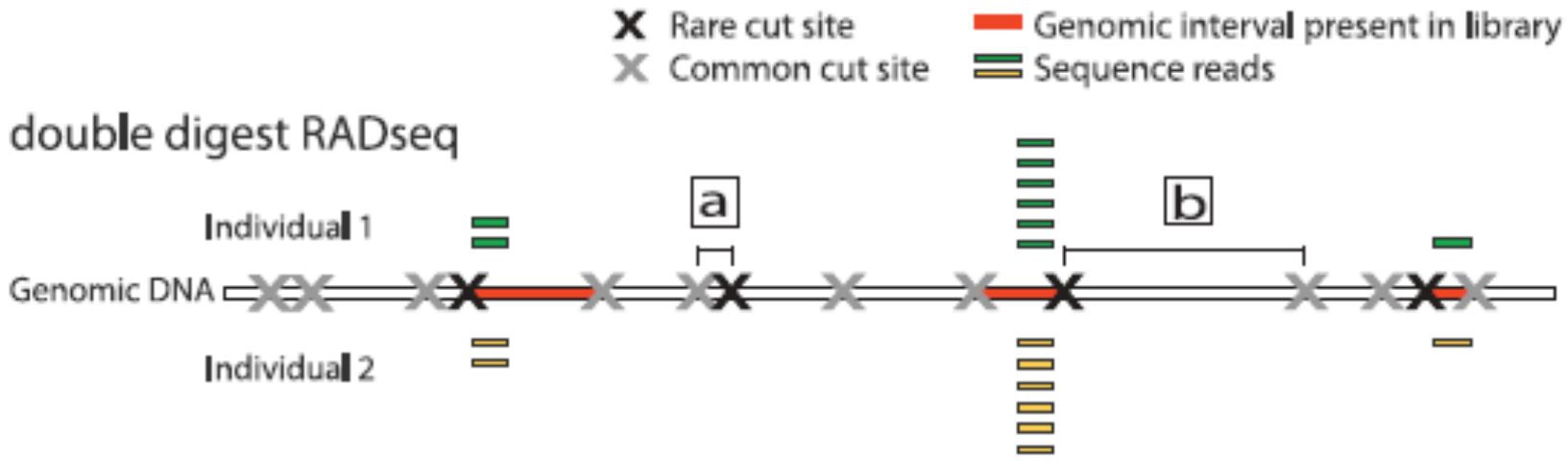


Baird *et al.* (2008) PLoS ONE

RADSeq – double digest

TGAC
The Genome Analysis Centre™

- Double enzyme restriction, precise size selection, easier protocol
- Identify loci based on read stacks



Peterson *et al.* (2012) PLoS ONE

Advantages:

- Reduced sequencing costs - multiplexing
- Less data to handle
- Reproducible

Disadvantages

- More complicated protocols
- Licence restrictions?
- Sites ‘randomly’ distributed throughout genome (maybe advantage?)

Sequencing strategies

TGAC
The Genome Analysis Centre™

Genotyping-by-Sequencing

TGAC
The Genome Analysis Centre™

- Enzyme restriction with frequent cutter
  - e.g *PstI* or *ApeKI* (methylation-sensitive)
  - Or 2 enzymes (one rare and one common cutter e.g. *PstI* + *MspI*)
- Implicit size selection by Illumina sequencing
- Identify loci based on stacks

Elshire *et al.* (2011) PLoS ONE

Poland *et al.* (2012) PLoS ONE

Genotyping-by-Sequencing

TGAC
The Genome Analysis Centre™

Advantages:

- Reduced sequencing costs - multiplexing
- Less data to handle – especially if single-end
- Reproducible

Disadvantages

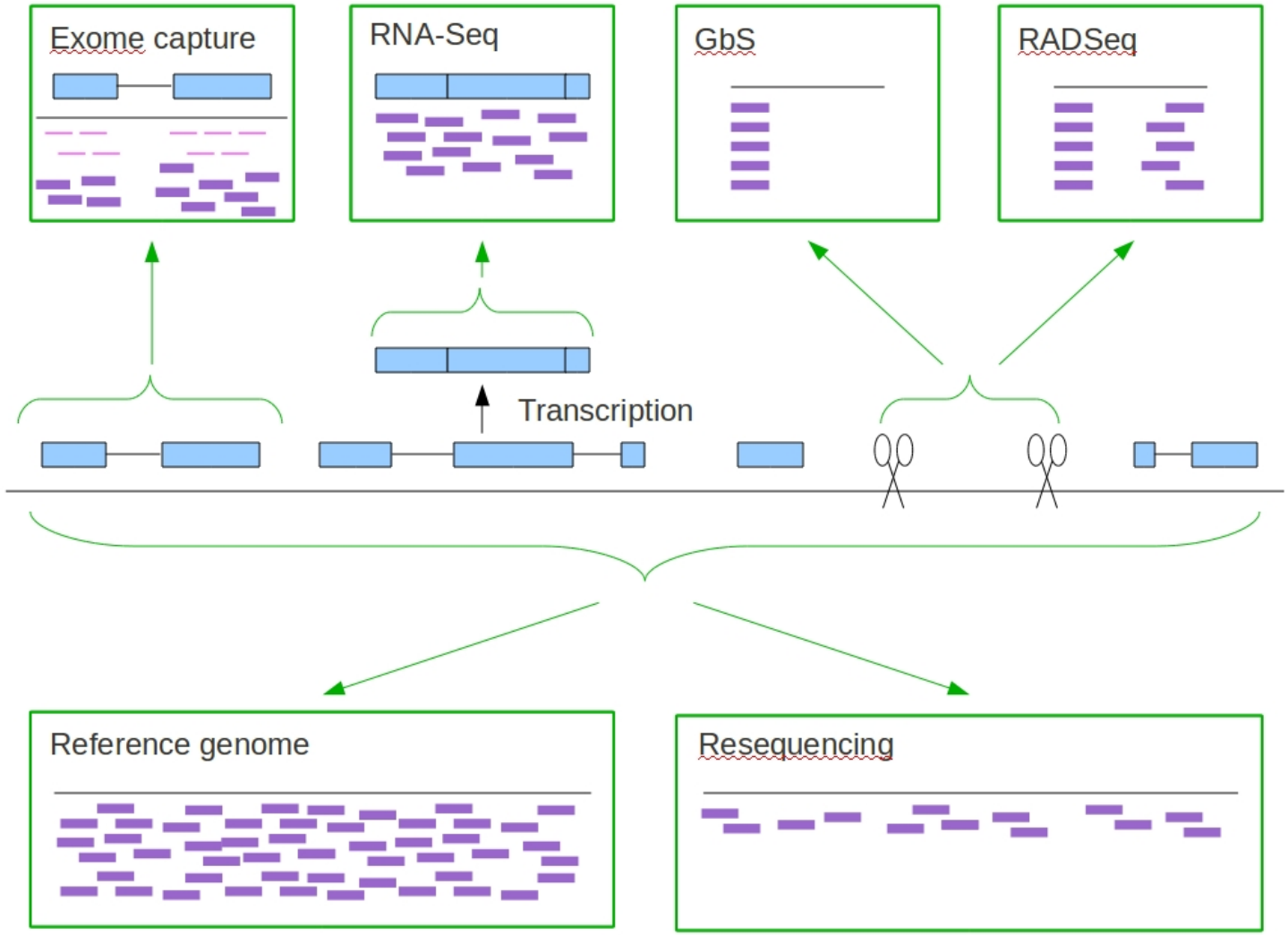
- Licence restrictions?
- Sites ‘randomly’ distributed throughout genome (maybe advantage?)

What can be (re)sequenced

TGAC
The Genome Analysis Centre™

Size Reduction

Whole genome

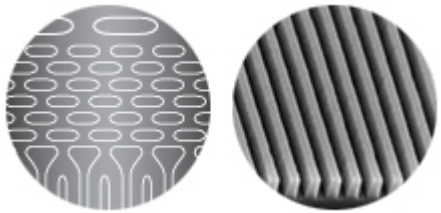


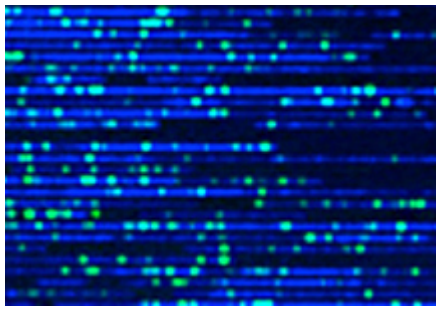
Optical mapping

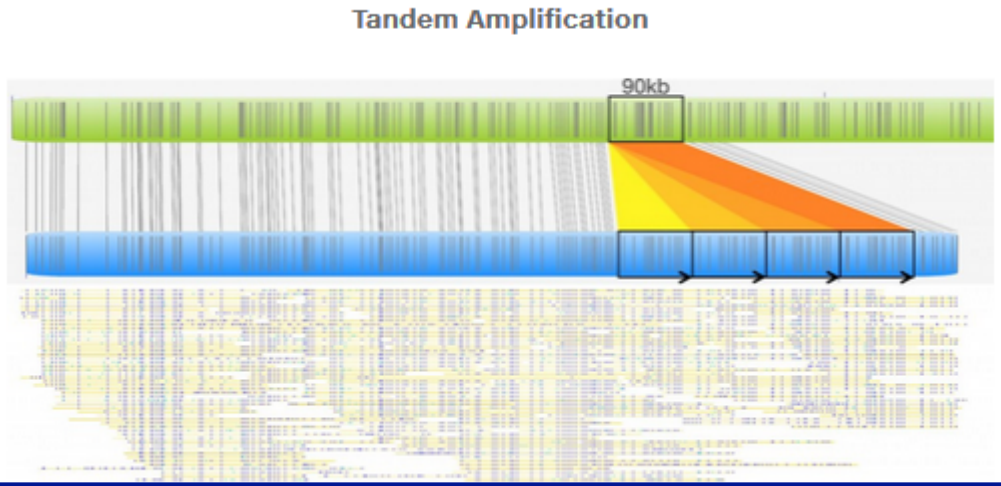
TGAC
The Genome Analysis Centre™

BIONANO GENOMICS

- Irys System for rearrangements
  - High molecular weight DNA (>150Kb)







Summary

TGAC
The Genome Analysis Centre™

- What do you want to know?
- How difficult is your genome?
- Do you need to sequence?
- What should you sequence and how?
  - Targets
  - Technologies
- Involve your bioinformaticians and statisticians from the start ☺


THANKS!