SNP calling *vs.* sequencing coverage

Cost-effective approaches for variant calling and analysis in complex plants

JOSE DE VEGA

jose.devega@earlham.ac.uk

Earlham Institute

Decoding Living Systems

IFR Institute of Food Research

John Innes Centre

Norfolk and Norwich University Hospitals NHS Foundation Trust

The Sainsbury Laboratory TSL

Earlham Institute

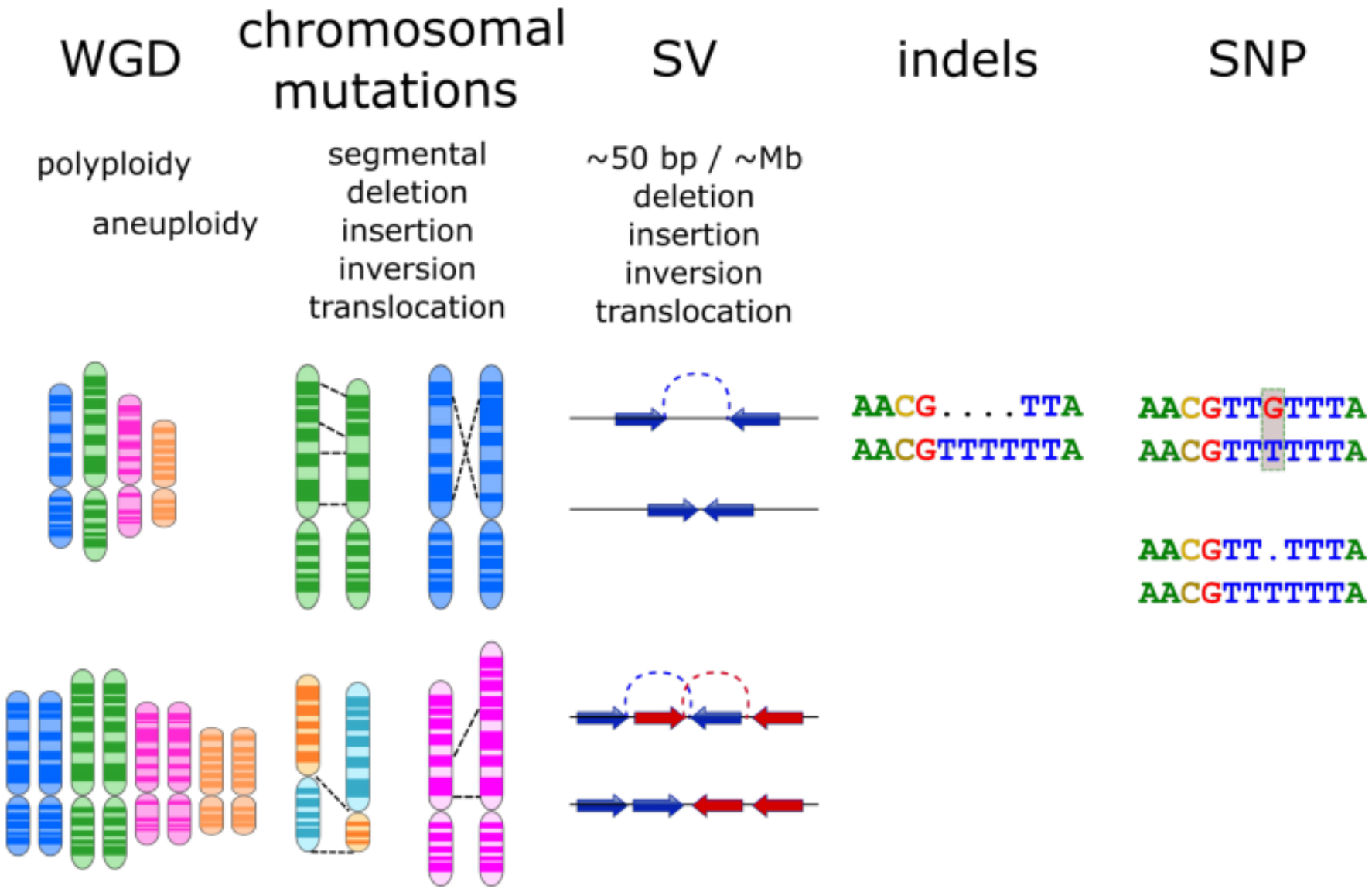
UEA University of East Anglia

http://www.earlham.ac.uk/science-strategy

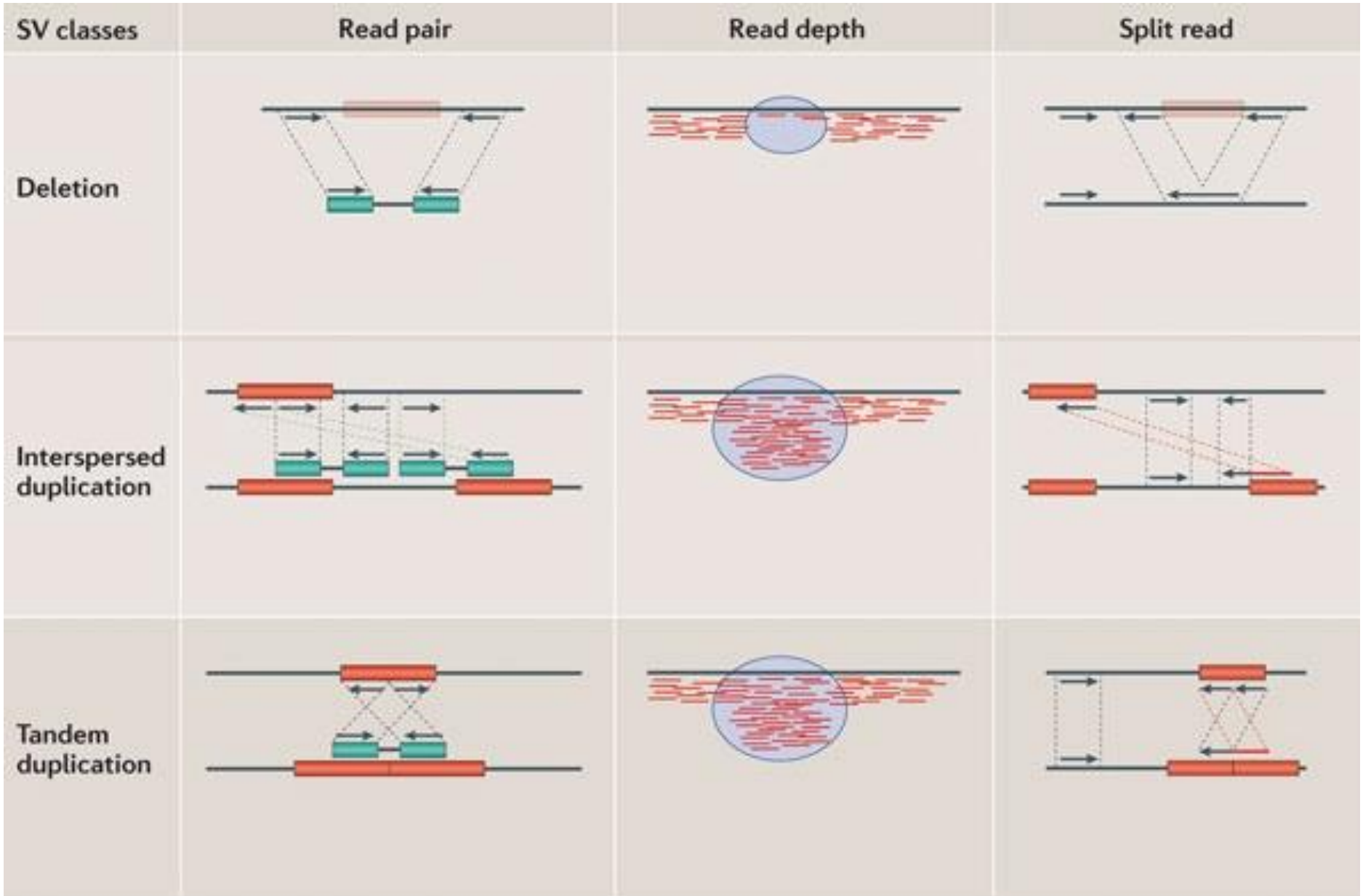
Bringing together multi-disciplinary expertise in living systems with computational science and biotechnology to answer ambitious biological questions and generate enabling resources.

Four UK National capabilities

Types of genomic variation



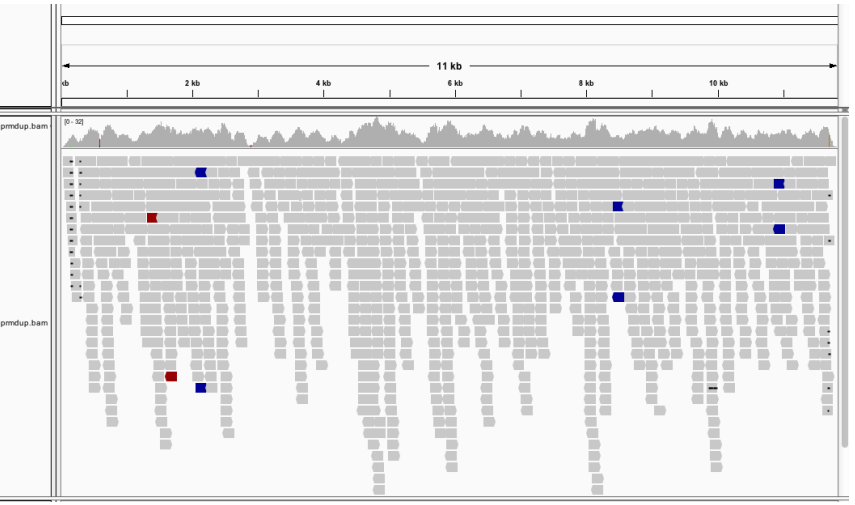
EMBL-EBI

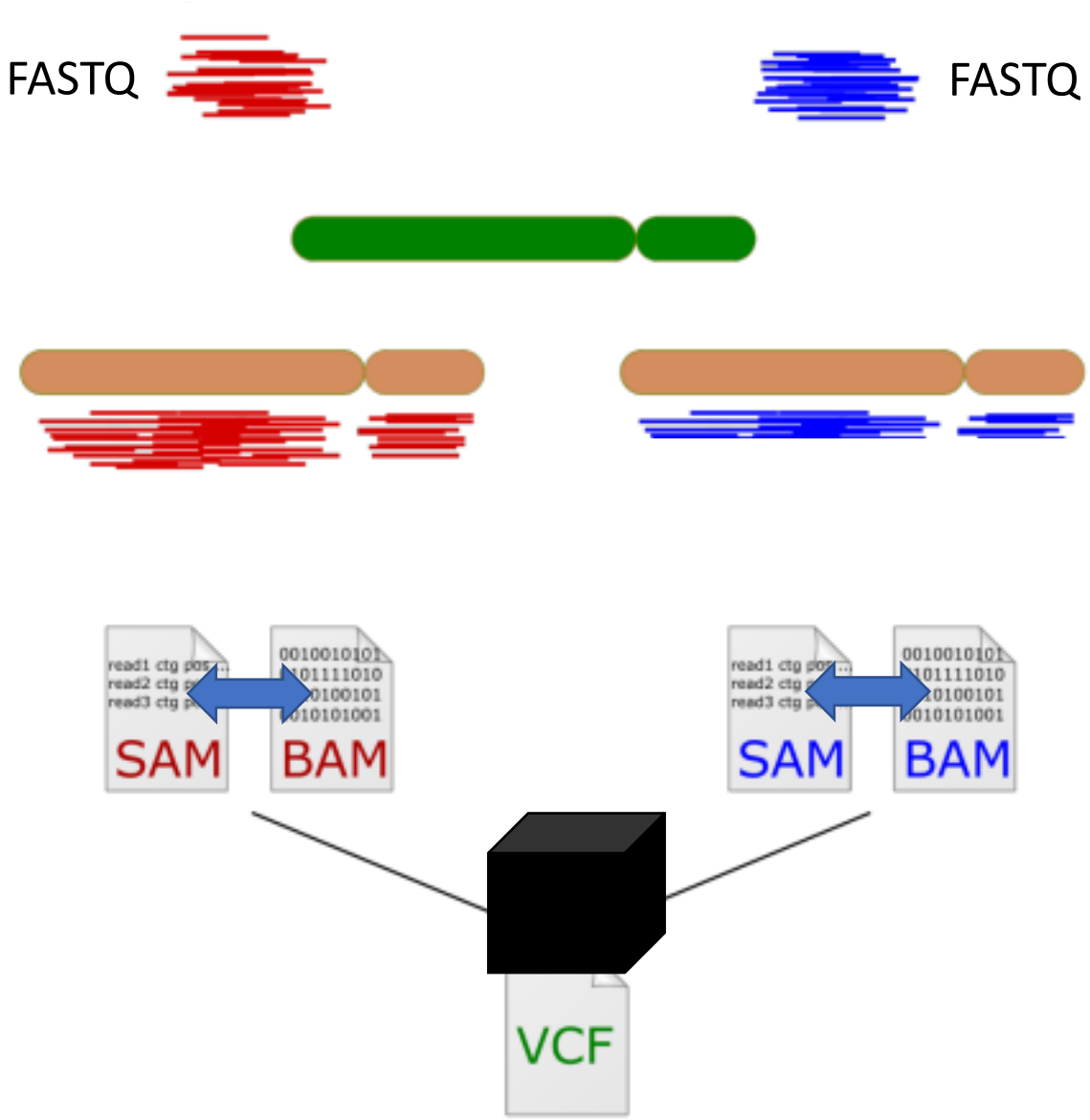


Alkan et al. (2011) Nature Rev Genet 12: 363-376

The basic workflow…

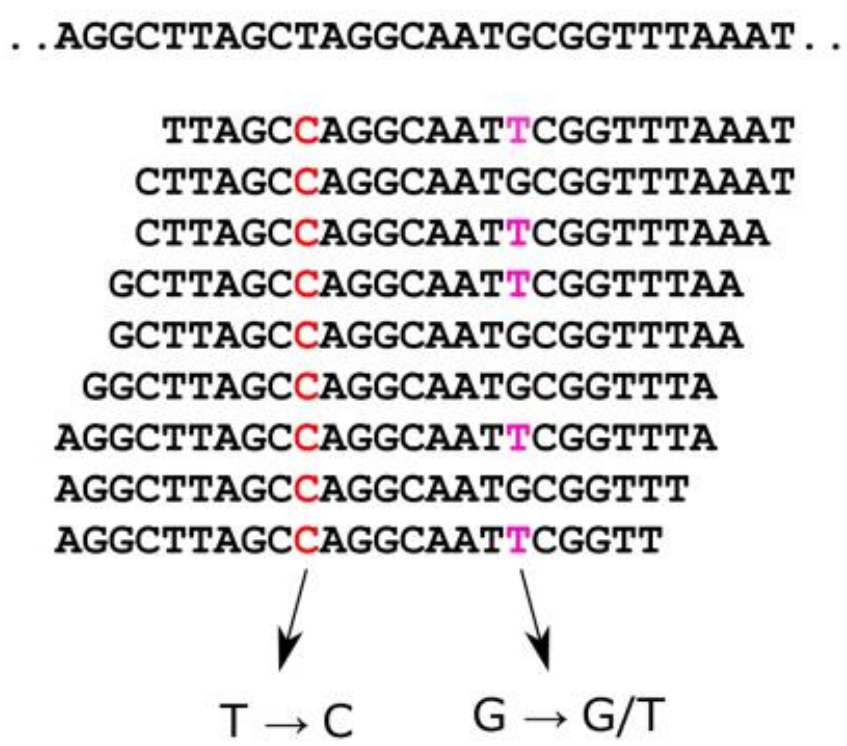
- Align reads to the reference
- Check for differences

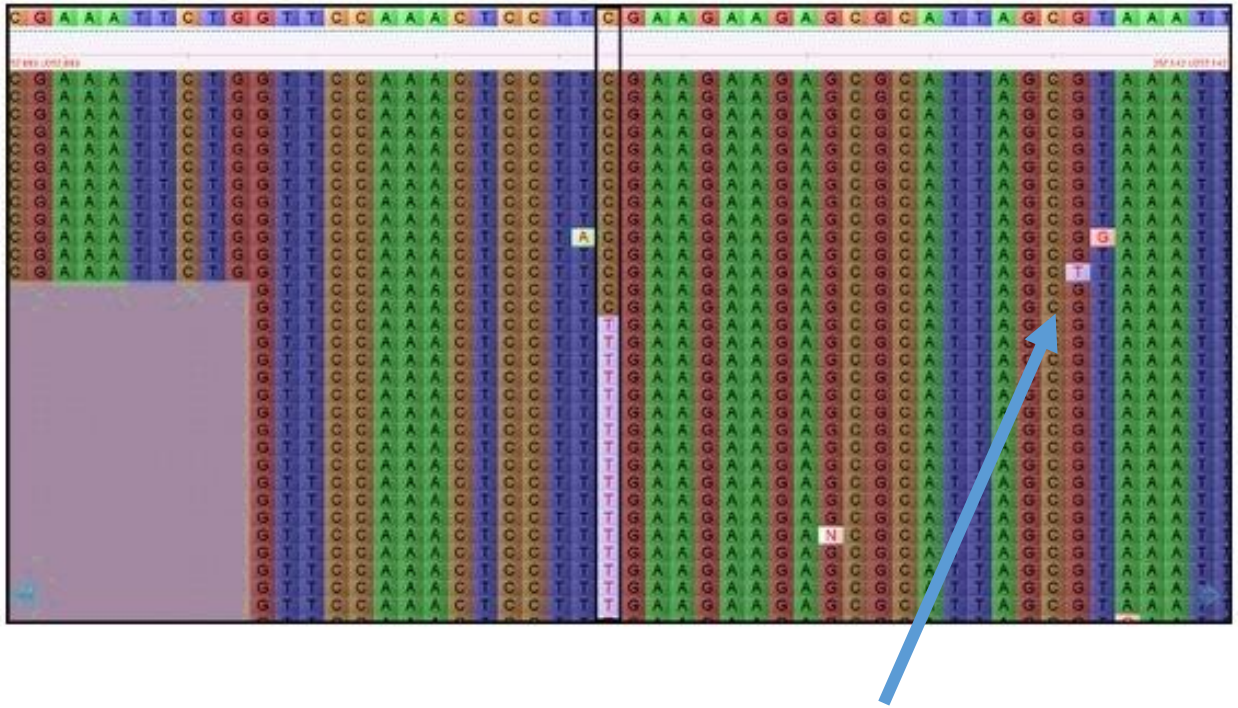




Sources of errors…

- Align reads to the reference
  - Reads are short (less every day)
  - Genomes are repetitive
- Check for differences
  - Sequencing introduces errors
  - Rare alleles





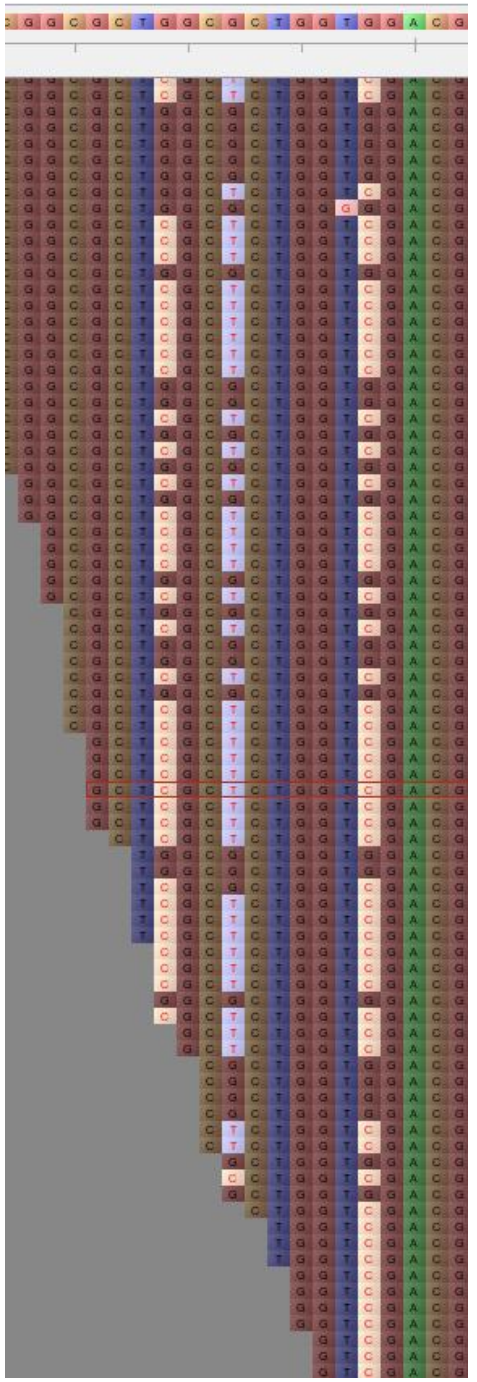
Sequencing error or Rare allele?

Sources of false positives (1 of 3): bad mapping

- Reads mapped to somewhere other than their true origin (e.g. recently duplicated genes)
- Accounts for approx. 40% of false positives

- Symptoms:
  - locations are heterozygous
  - several of these very close together
  - two or more clearly distinguishable classes of reads with variants in phase
  - difficult to distinguish from genuine haplotypes

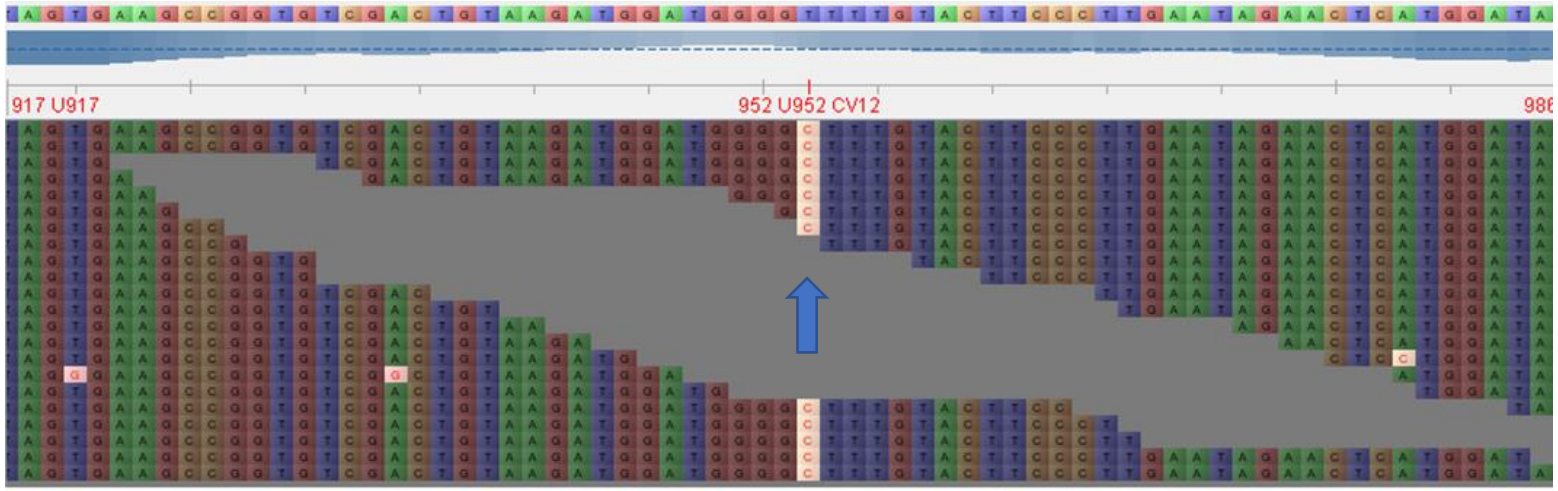
- Remediation: good mapping tool with appropriate means of suppressing mismapping (e.g. Bowtie/Bowtie2)



Sources of false positives (2 of 3): reference assembling errors

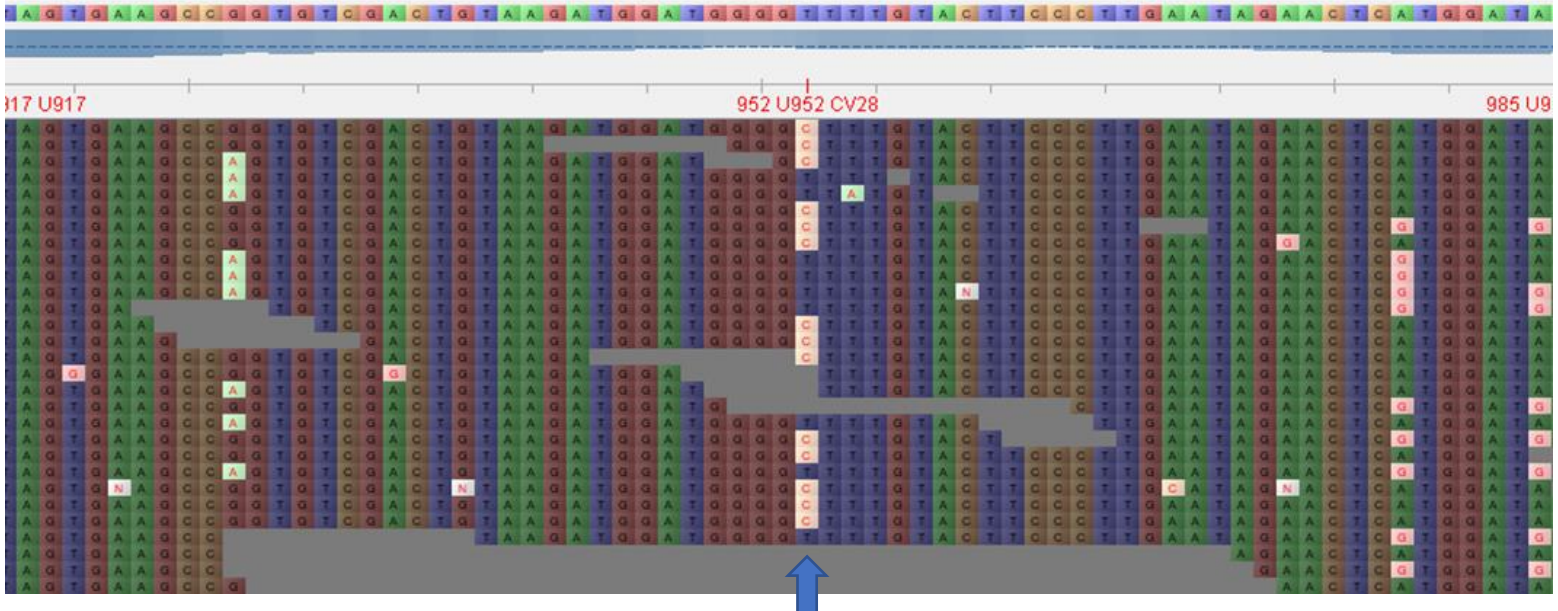
- Symptoms:
  - With the strict mapping of single samples, some SNPs are apparently homozygous
  - In the relaxed mapping, reads appear with many mismatches that have the genome allele

Strict mapping



- Implies *misassembly* of the reference
  - Assembler has likely produced a faulty sequence from two set of reads from pairs of paralogs

Relaxed mapping

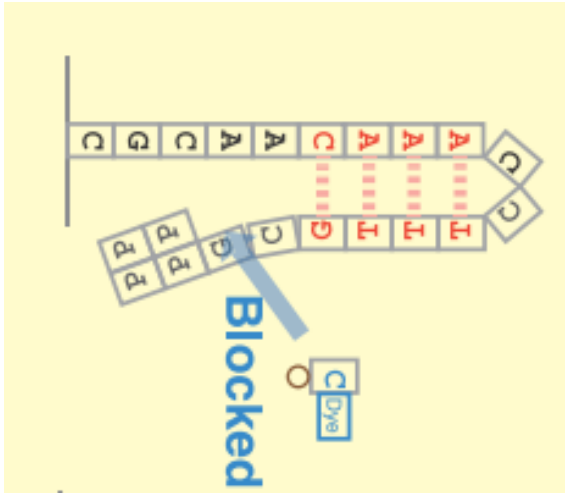


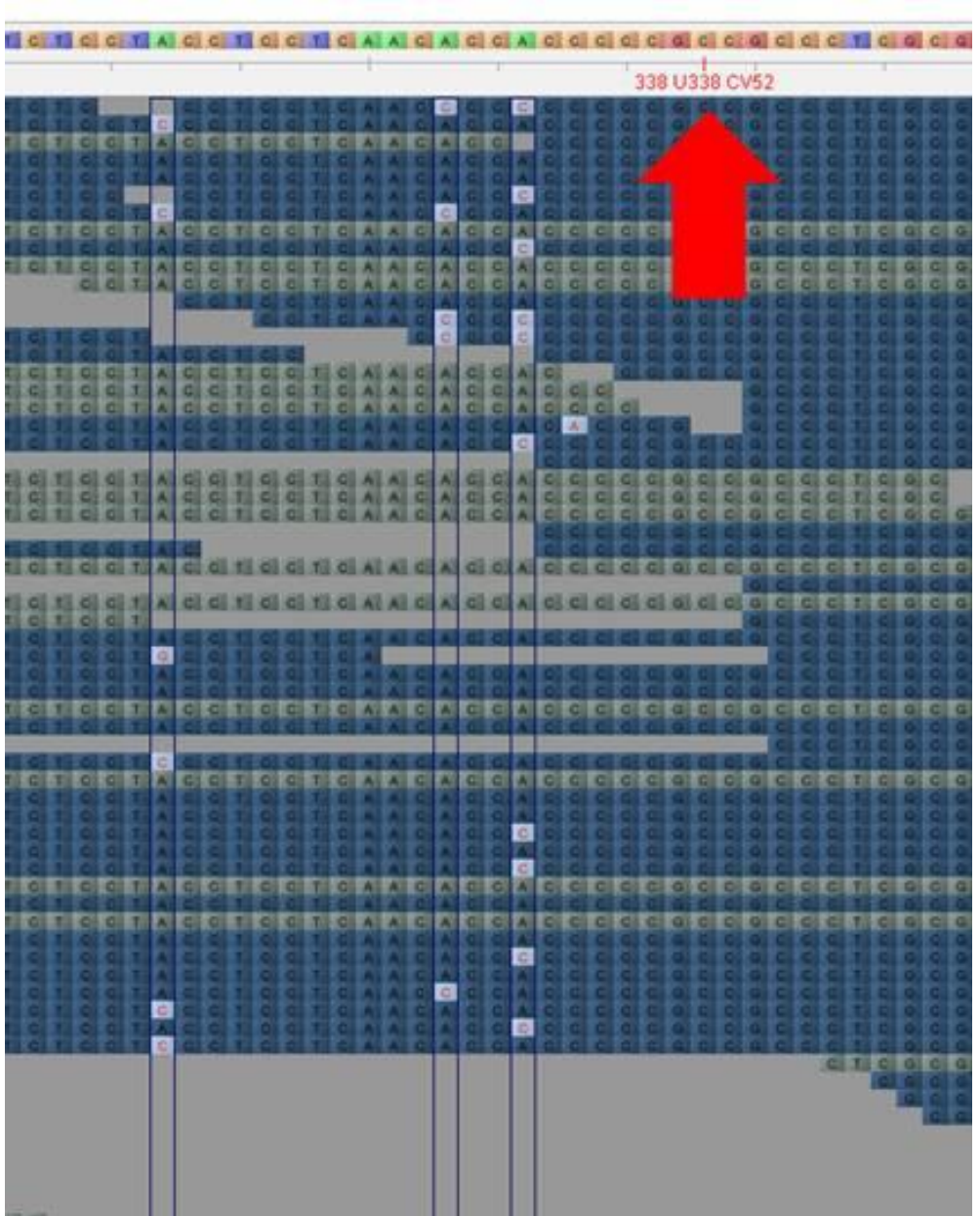
Agnieszka Golicz

Antonio Ribeiro

Sources of false positives (3/3): Illumina sequencing errors

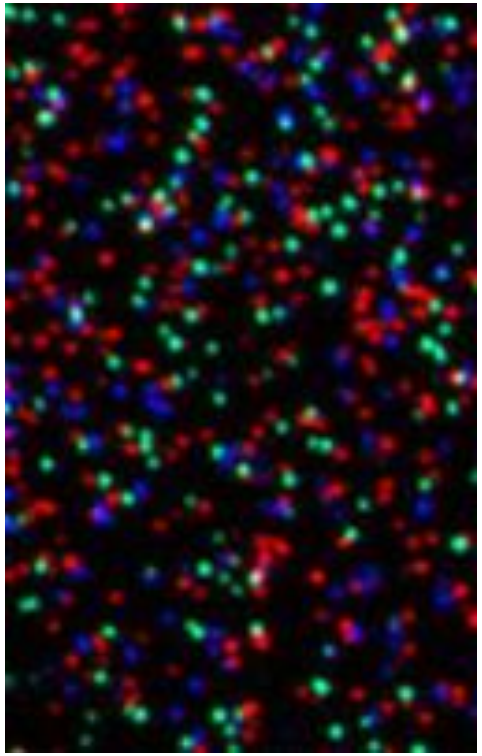
- Sequencing error in Illumina data in GGC motifs
  - GGC motifs inhibit DNA polymerase -- inverted repeats lead to folding of sequenced DNA strand
  - Both blocks base incorporation and leads to dephasing of signal in cluster



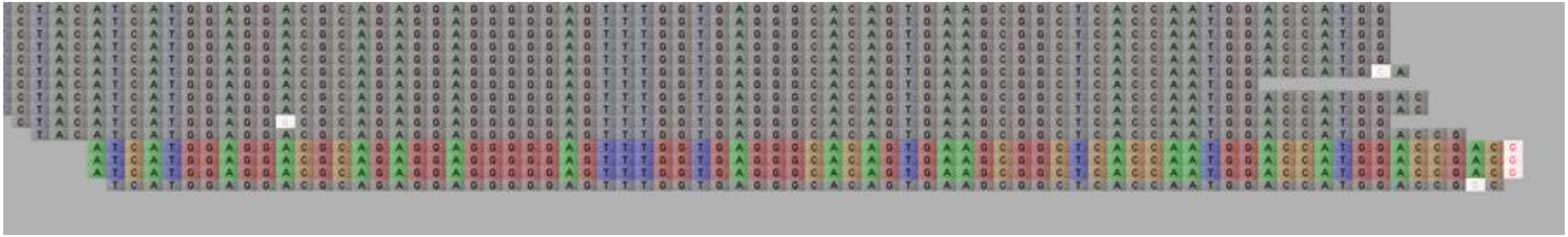


Sources of false positives (3/3): Illumina (minor) sequencing errors

- Sequencer image analysis error: Multiple identical reads are called from what should be a single cluster in flow cell

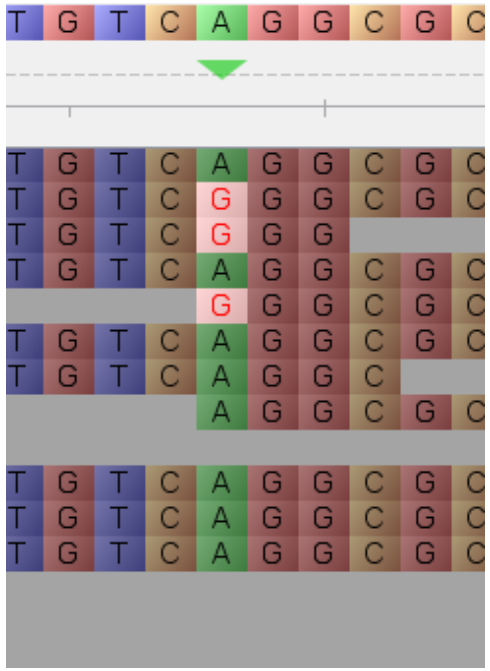


- Symptom: clusters of identical sequences with identical start/end positions and read error in same position across all reads
- Remediation: remove identical duplicates (also remove PCR duplicates in WGS, but not in GBS/RADSeq)

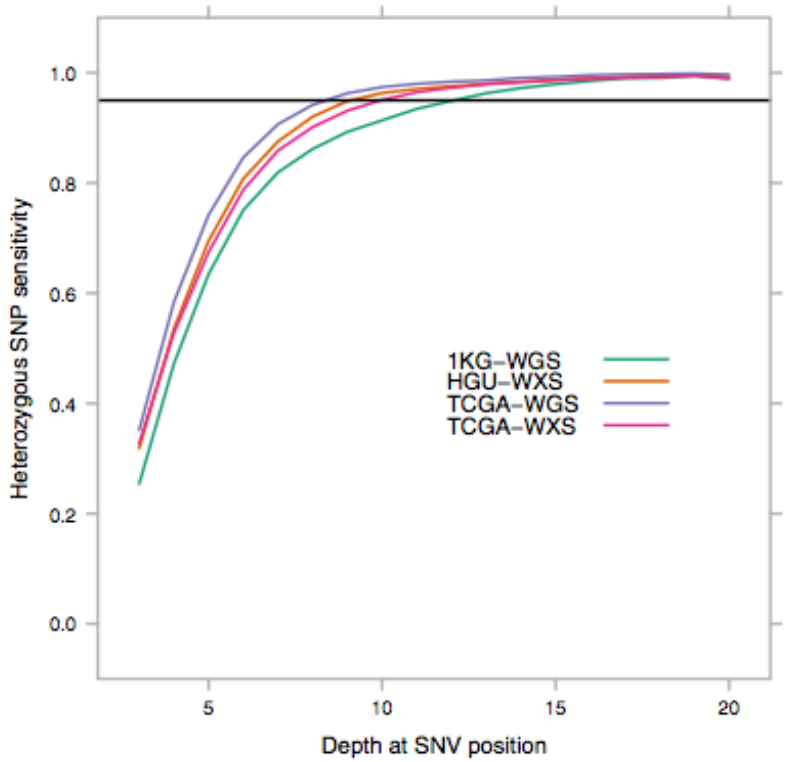


Coverage is key to reliable SNPs

- Coverage = times a given region has been sequenced
- Trace-off with cost



- Risk of not sampling all chromosomal regions
- To reliably call genotypes we need good coverage

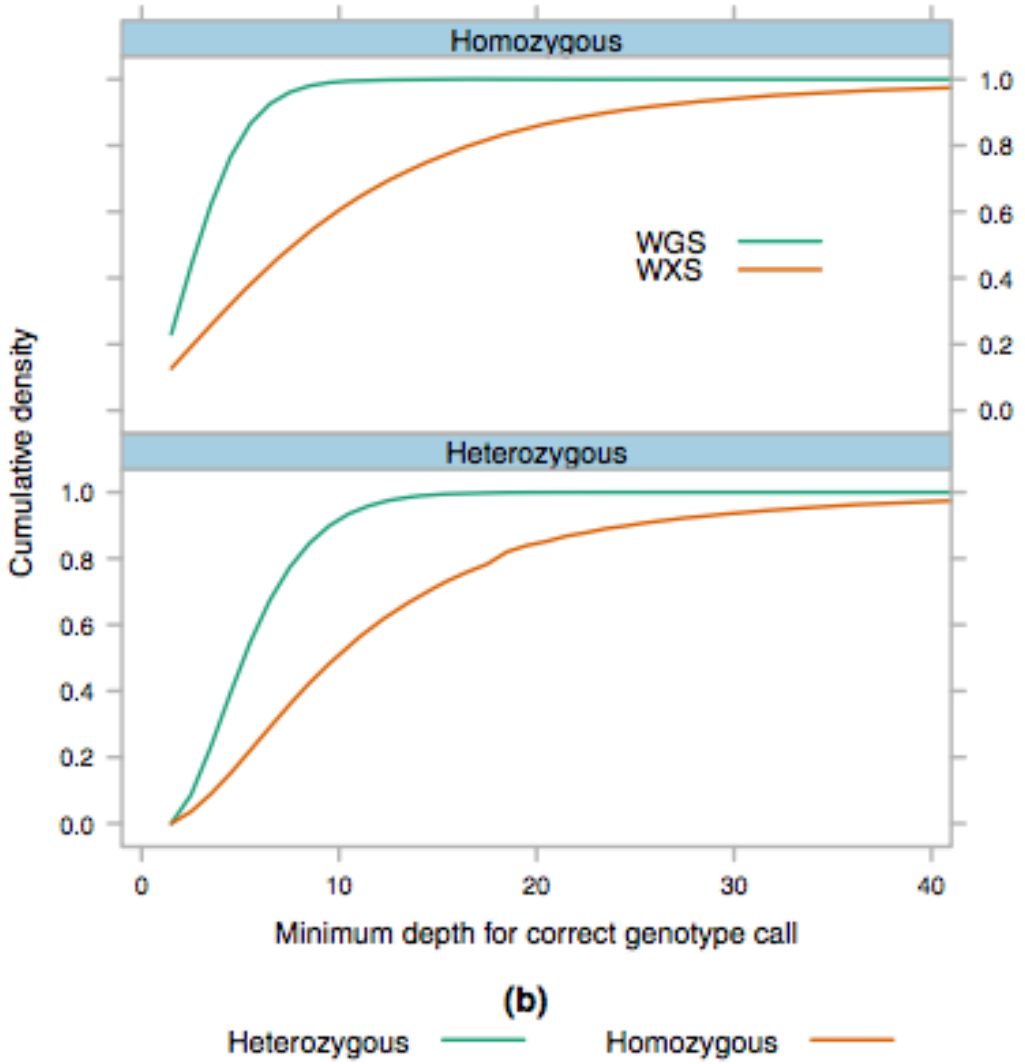


SNP detection sensitivity for exome and whole genome sequencing samples

Meynert et al. BMC Bioinformatics 2014, 15:247

Minimum coverage

Depends on technique



Meynert et al.. BMC Bioinformatics 2014, 15:247

Depends on reads length

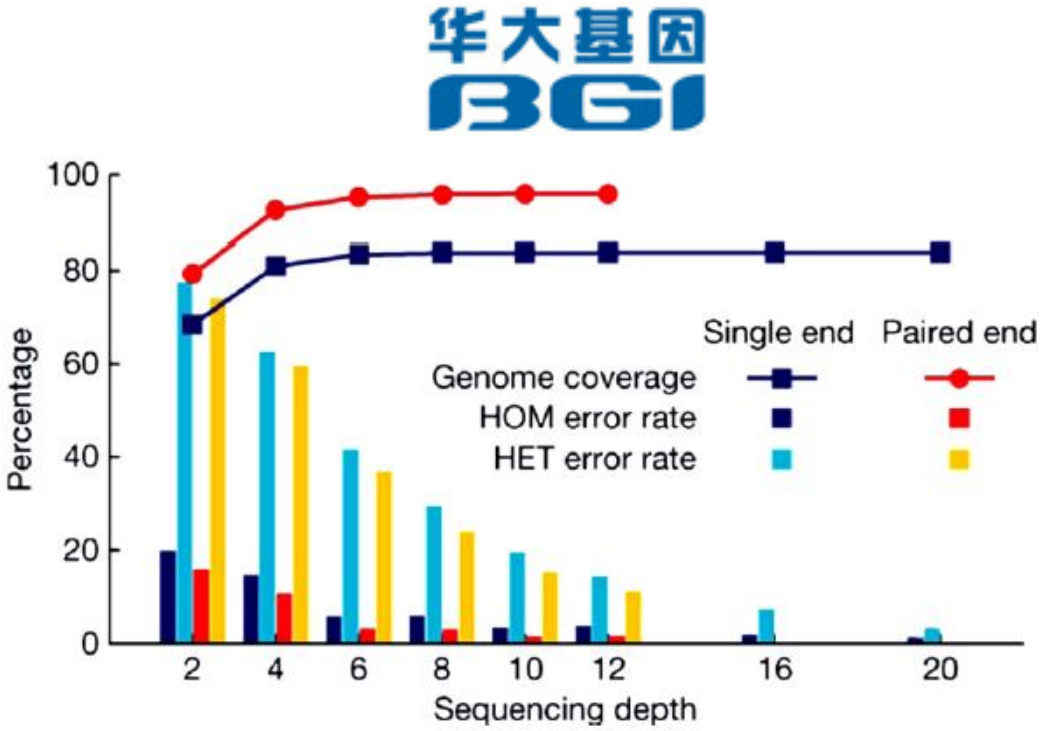
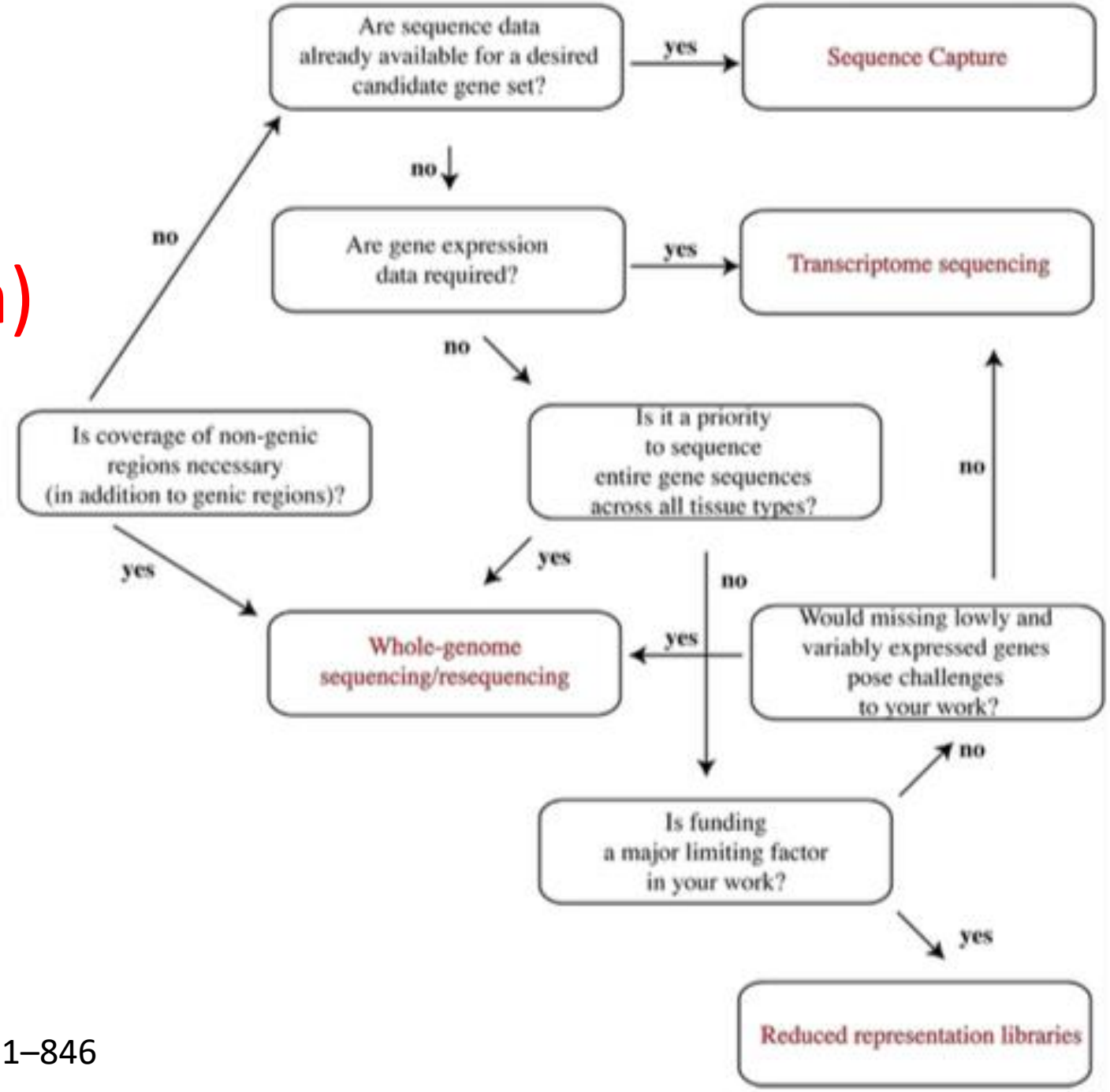


Figure 1-2-2 Genome coverage of the assembled consensus sequence and the accuracy of SNP detection as a function of sequencing depth

http://www.genomics.hk/PlantWhole.htm

Strategy for SNP Discovery Projects (Library preparation)



Clevenger et al (2015) Molecular Plant 8, 831–846

Genome reduced-representations

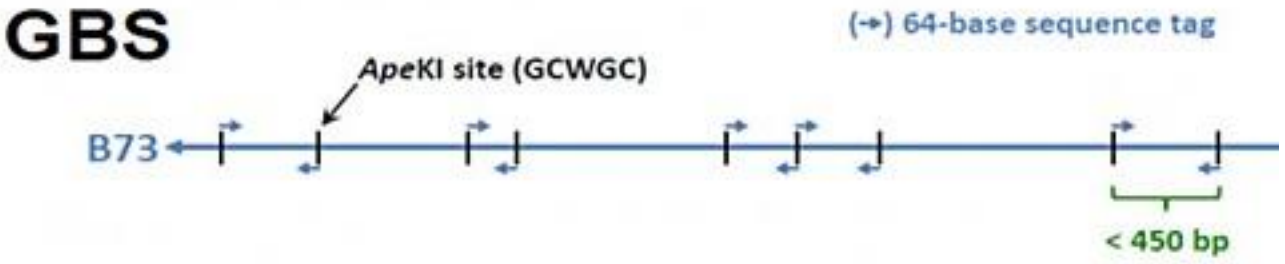
Increase coverage in specific loci

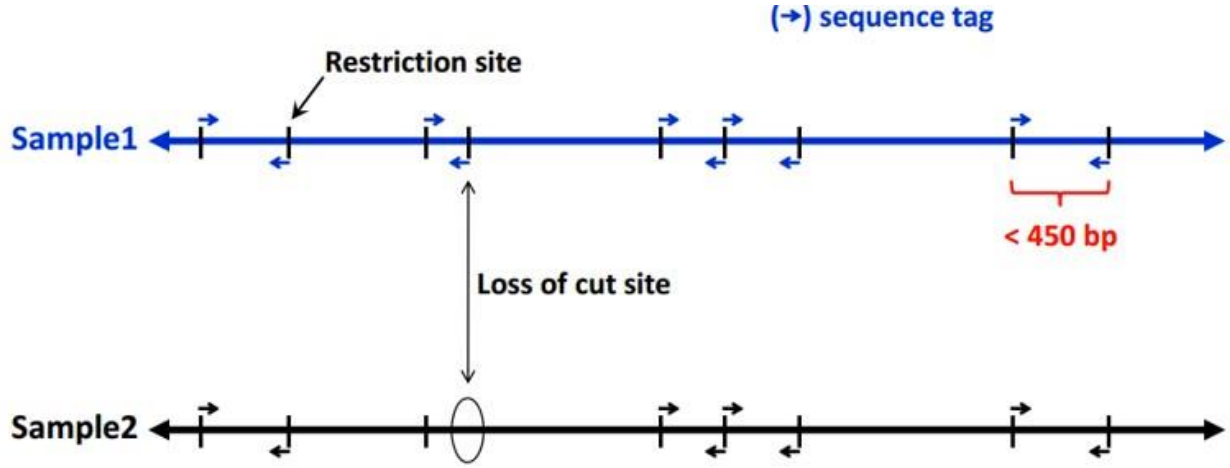
Enzymatic (Single or double)

- GBS, RAD-Seq
- Keygene Patent (March 2016) SBG

Homology

- Capture (Exon)
- rAmpSeq
- DArTseq



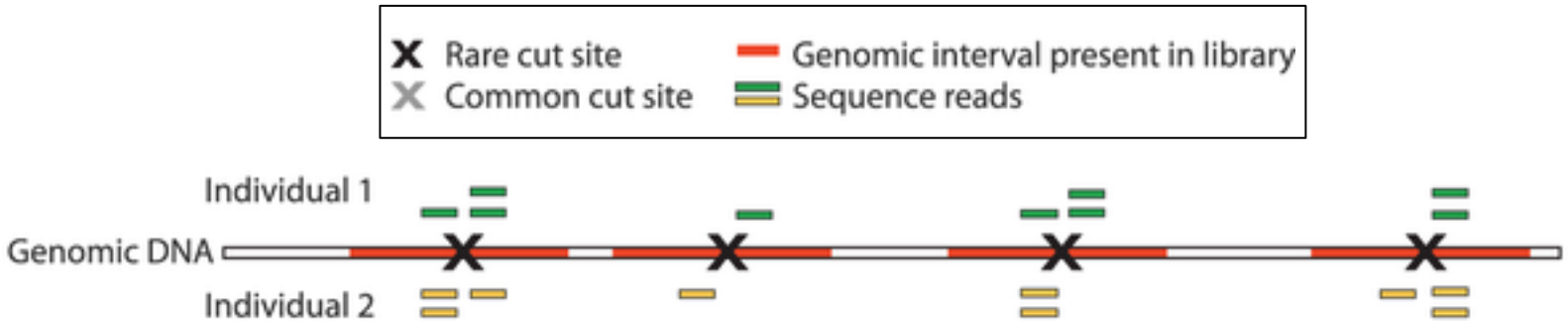


Genome reduced-representations

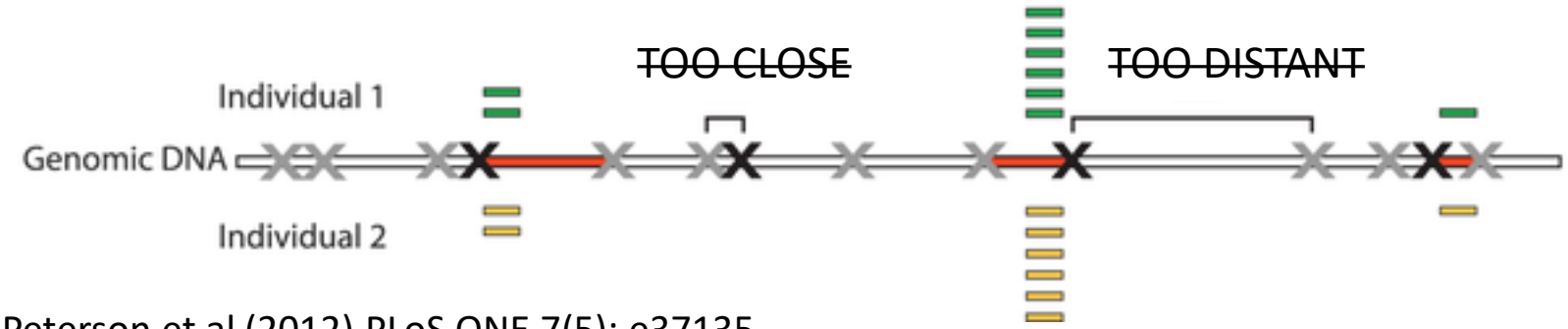
Increase coverage in specific loci

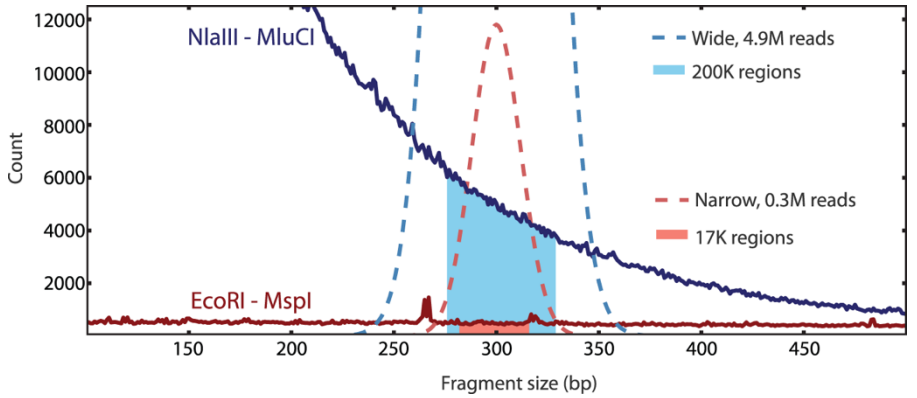
Enzymatic (Single or double)

- GBS, RAD-Seq
- Keygene Patent (March 2016) SBG



Double-digested GBS





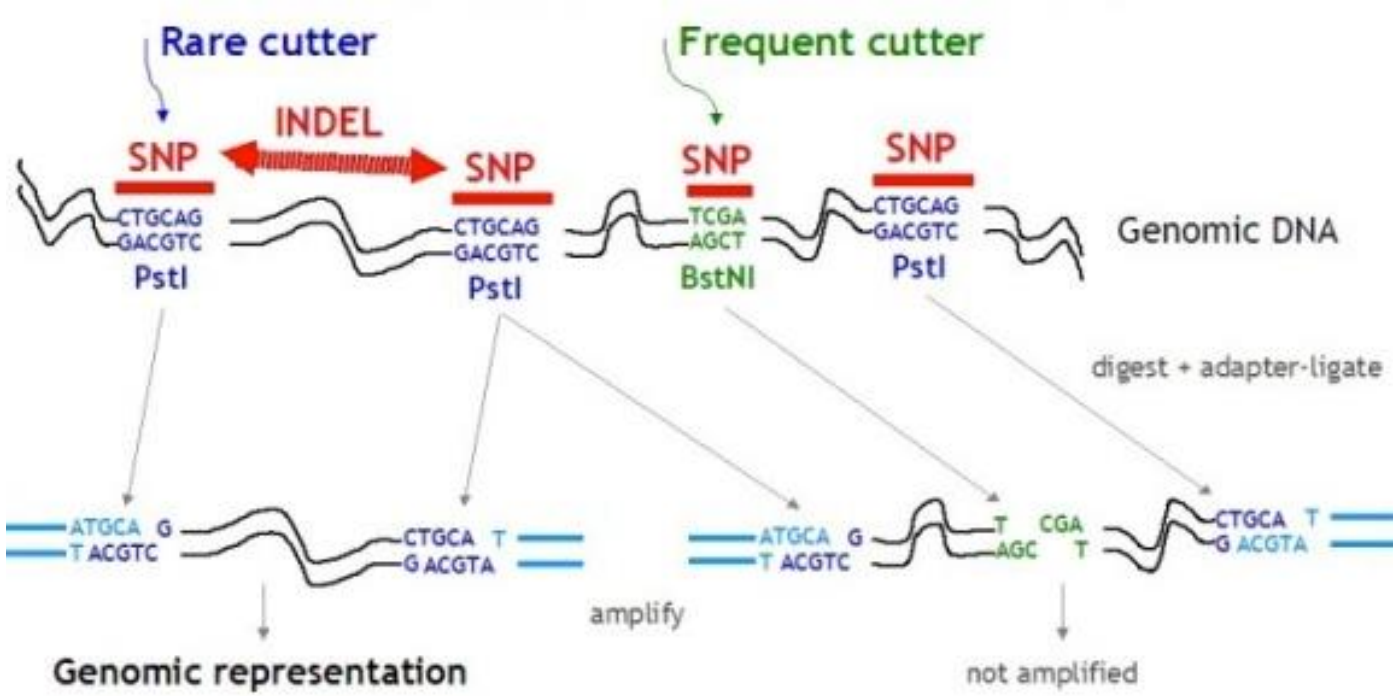
Peterson et al (2012) PLoS ONE 7(5): e37135.

Genome reduced-representations

Increase coverage in specific loci

Homology

- Capture (Exon)
- rAmpSeq
- DArTseq



http://www.diversityarrays.com/dart-application-microarray-process-complexity-reduction

New Results

rAmpSeq: Using repetitive sequences for robust genotyping

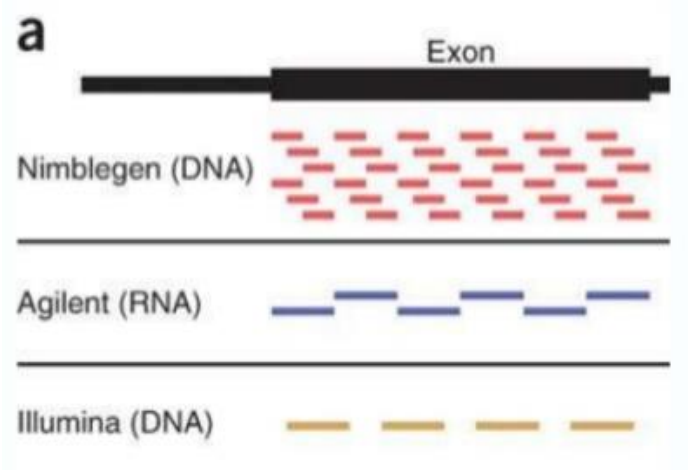
Edward S Buckler, Daniel C. Ilut, Xiaoyun Wang, Tobias Kretzschmar, Michael A. Gore, Sharon E. Mitchell

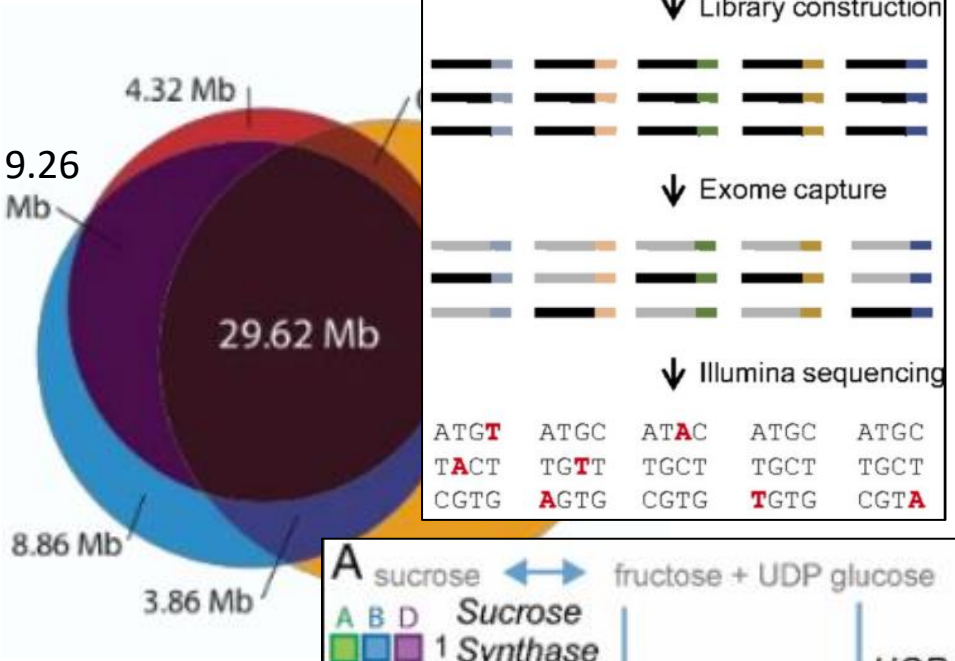
doi: https://doi.org/10.1101/096628

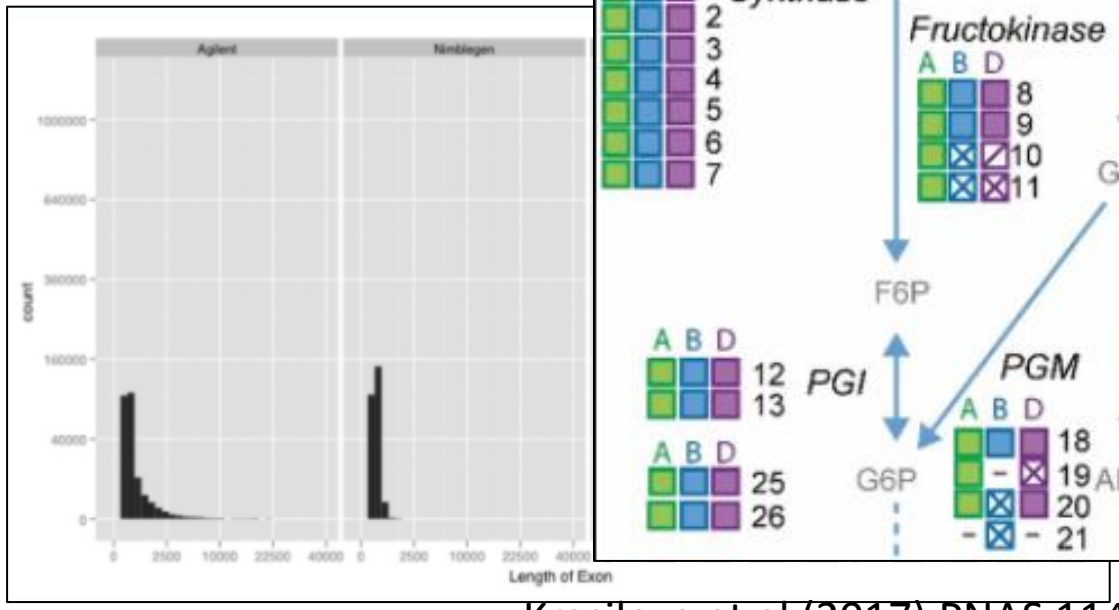
Exon capture

- Most functionally understood regions
- Capture reaction is a bottleneck

Bait lib design:

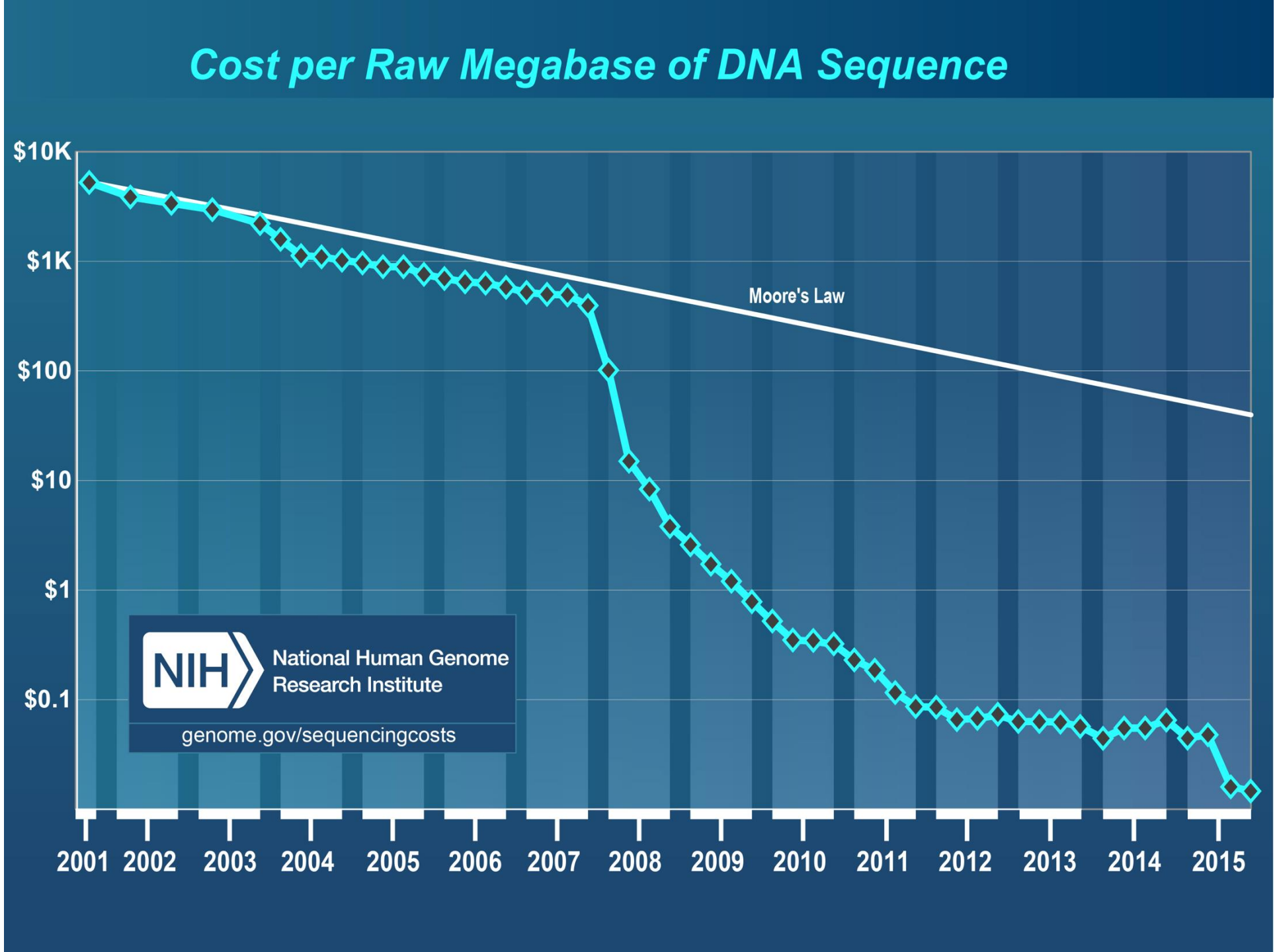






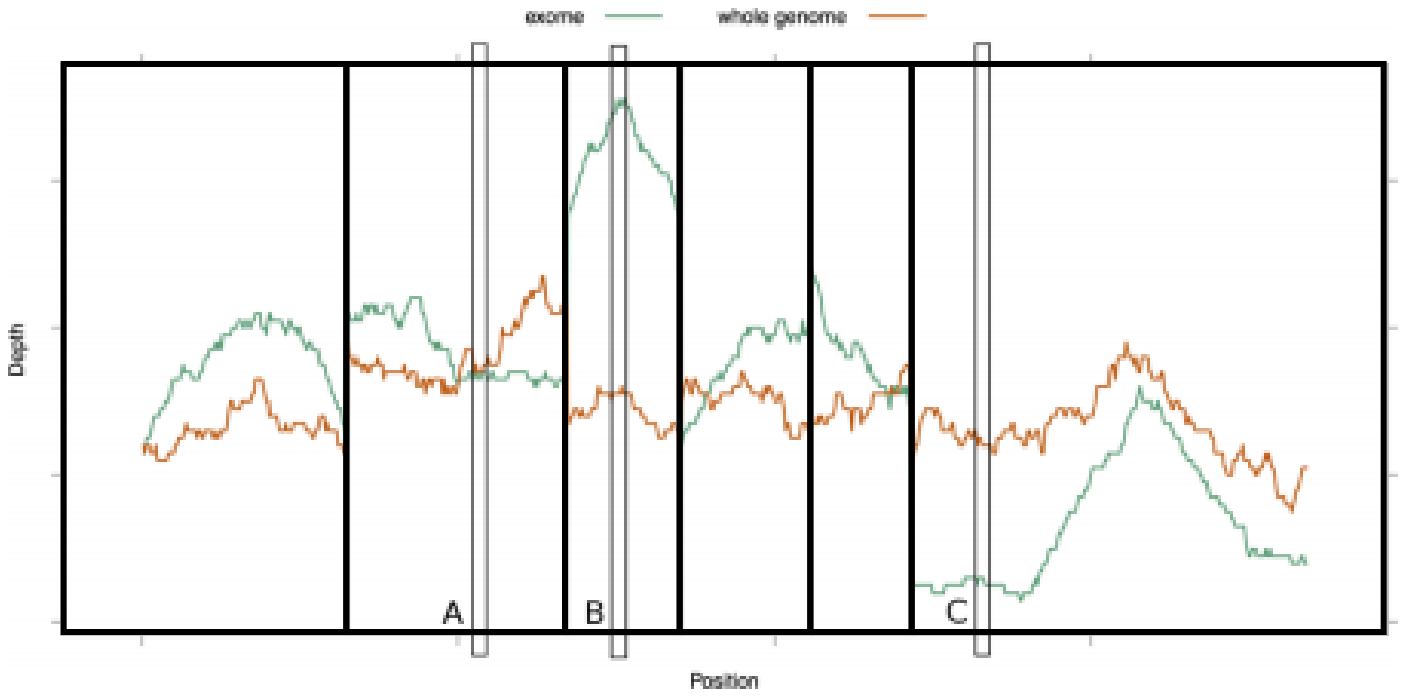
Clark et al 2011 Nat Biotech 29 (10): 907

Krasileva et al (2017) PNAS 114 (6): 913

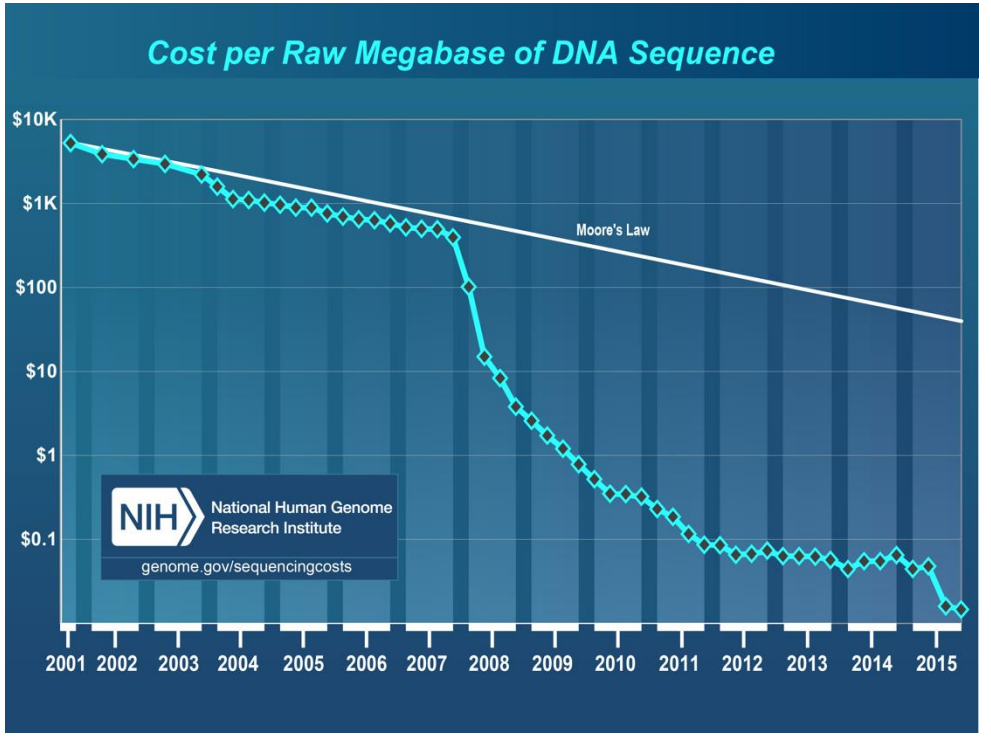

Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Why whole-genome?

- Uniform of read coverage and more balanced allele ratio calls
- Sequencing costs (But still an issue for population/cohorts analysis)
- Rare/structural variants discovery
- “This is it!” / “The time has come” …

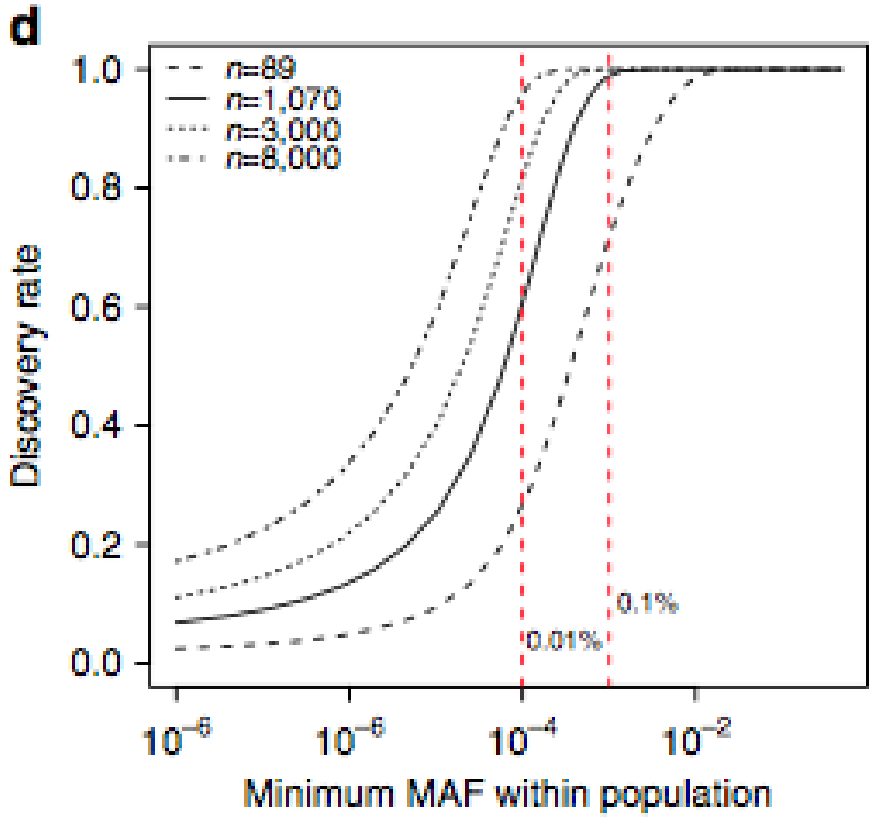


Meynert et al.. BMC Bioinformatics 2014, 15:247

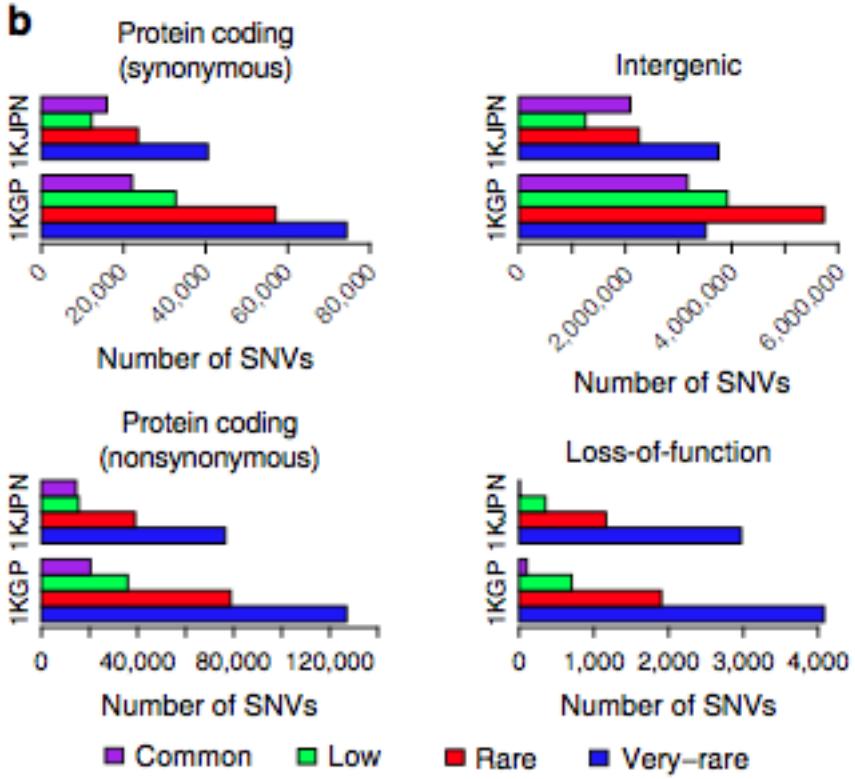


Rare alleles

Very rare alleles are difficult to discover…



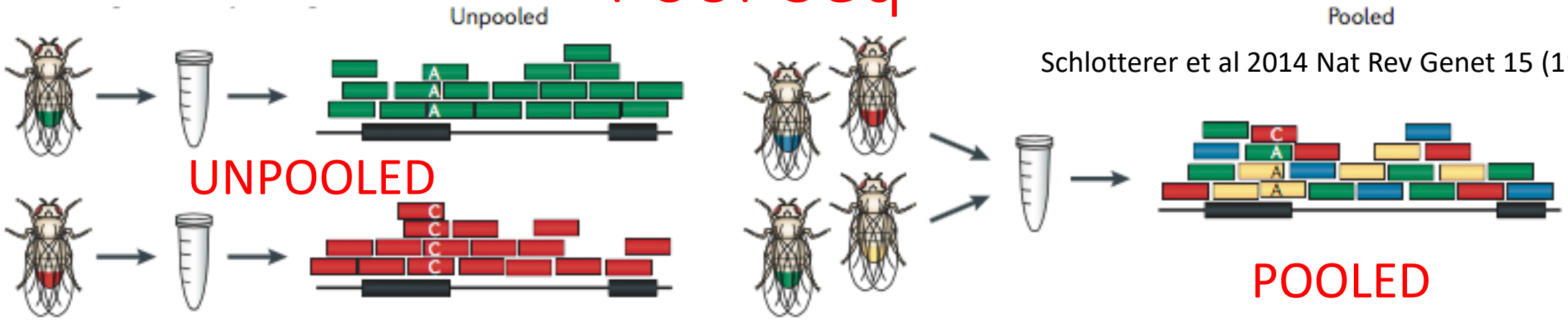
…and have functional genetic impact



Nagasaki et al 2015 Nat Comm 6:8018

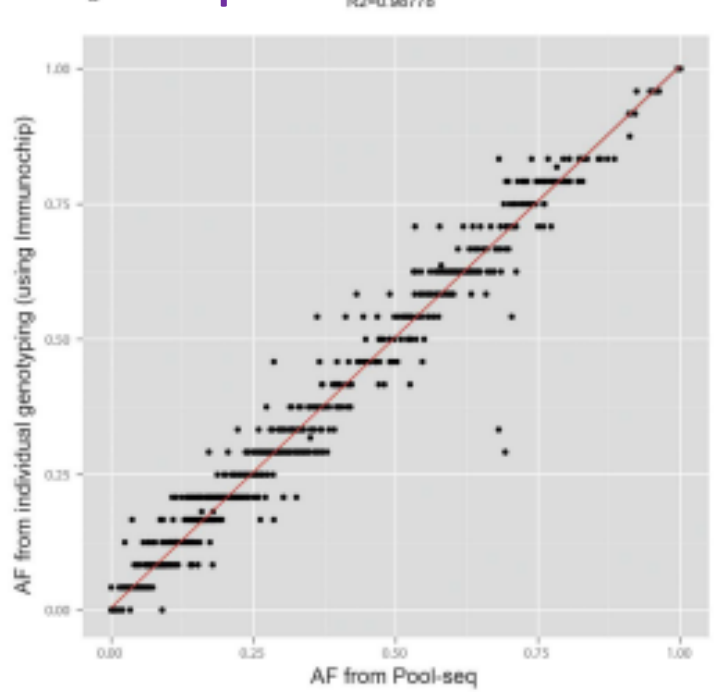
- High-coverage
- Find rarer variants if new samples are sequenced (cohort dependency)
- Difficult to impute (population specific)
- Defined fine characteristics of SVs, CNVs and HLA types in a population

Pool-Seq



Schlotterer et al 2014 Nat Rev Genet 15 (11): 749

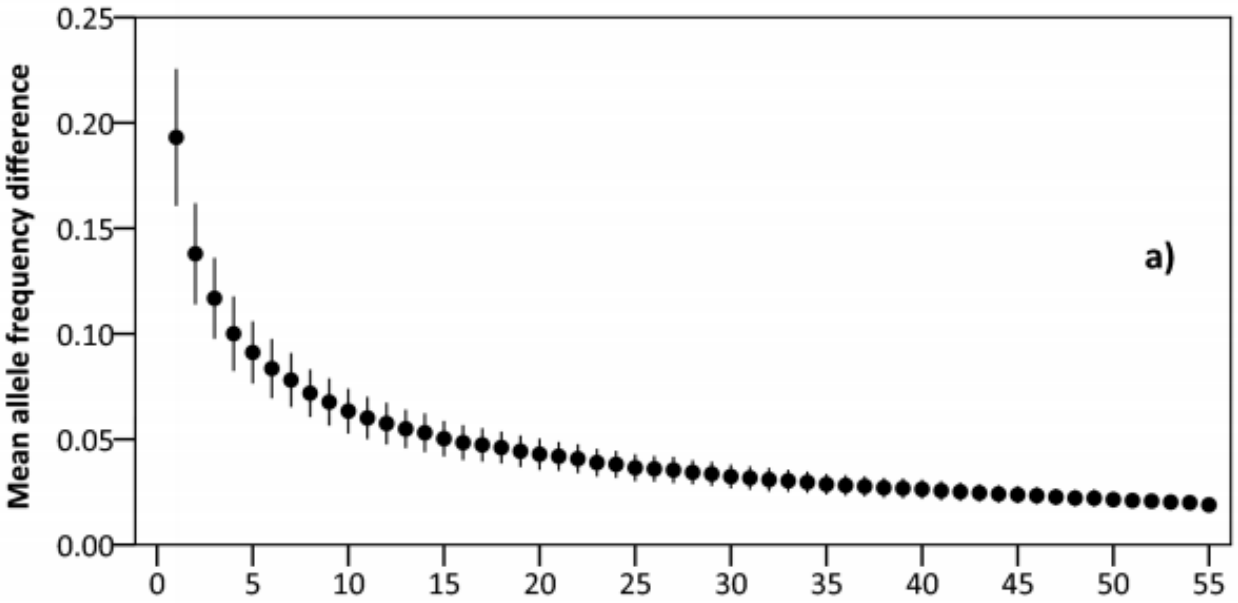
Pool-seq AFs are reliable



Anand et al 2015 Sci Rep 6: 33735

Effect of sequencing coverage

on the accuracy of the Pool-Seq in Arabidopsis haller

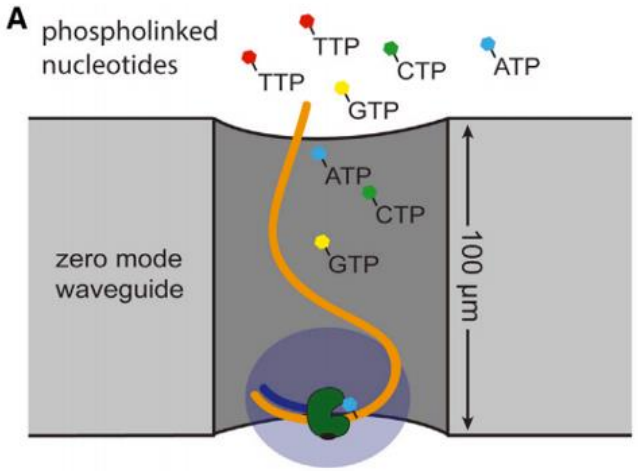


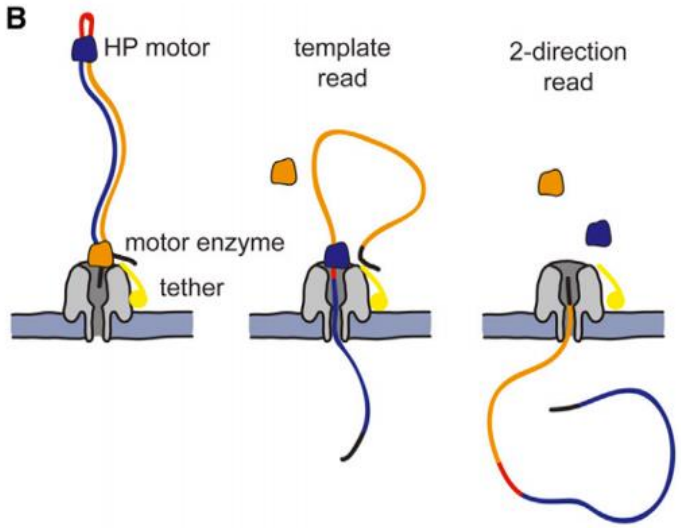
Rellstab et al 2013 PLoS ONE 8(11): e80422

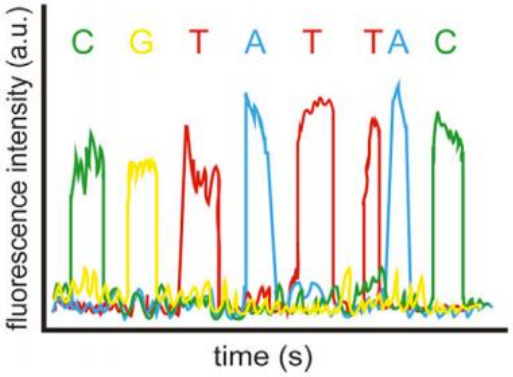
Cost-effective WGS?

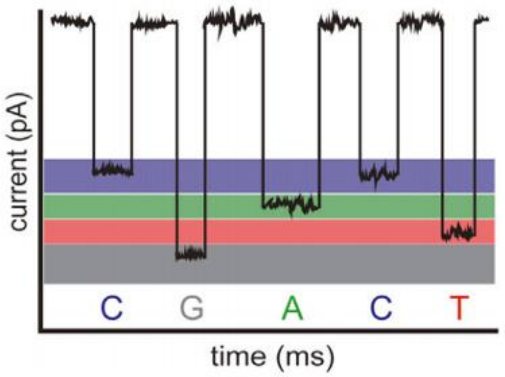
- Automatized sample barcoding before pooling (LITE)
  - Normal distribution of coverage-per-sample (How low can we go?)
- Exploit population features (kinship, WGD)
  - “Population power” (as in Pool-Seq)
  - Different bioinformatic pipeline for rare alleles (Low MAF)

Nanopore sequencing









Acknowledges…

- *"Crop Genomics and Diversity":* Paul Bailey, Jon Wright, Kirstie, Ji Zhou, Ned Peel, Gonzalo Garcia, Ricardo Ramirez, Bernardo Clavijo, *Office 202*.
- Federica Di Palma, Anthony Hall, Sarah Ayling
- P&P, project managers, EI's Faculty
- IBERS
  - Leif Skot, Charlotte Jones, Kerrie Farar, Sarah Purdy, Ian Armstead, Narcis Fernandez, Iain Donnison…
- CIAT and U. Yale
  - Joe Tohme, Margaret Worthington, Stephen Dellaporta…

And you all for your time…

Earlham Institute

Decoding Living Systems

www.earlham.ac.uk

cenicaña

IBERS

Yale

CIAT

BBSRC

Newton Fund

RESEARCH COUNCILS UK

BRITISH COUNCIL

BRIDGE COLOMBIA

www.bridgecolombia.org