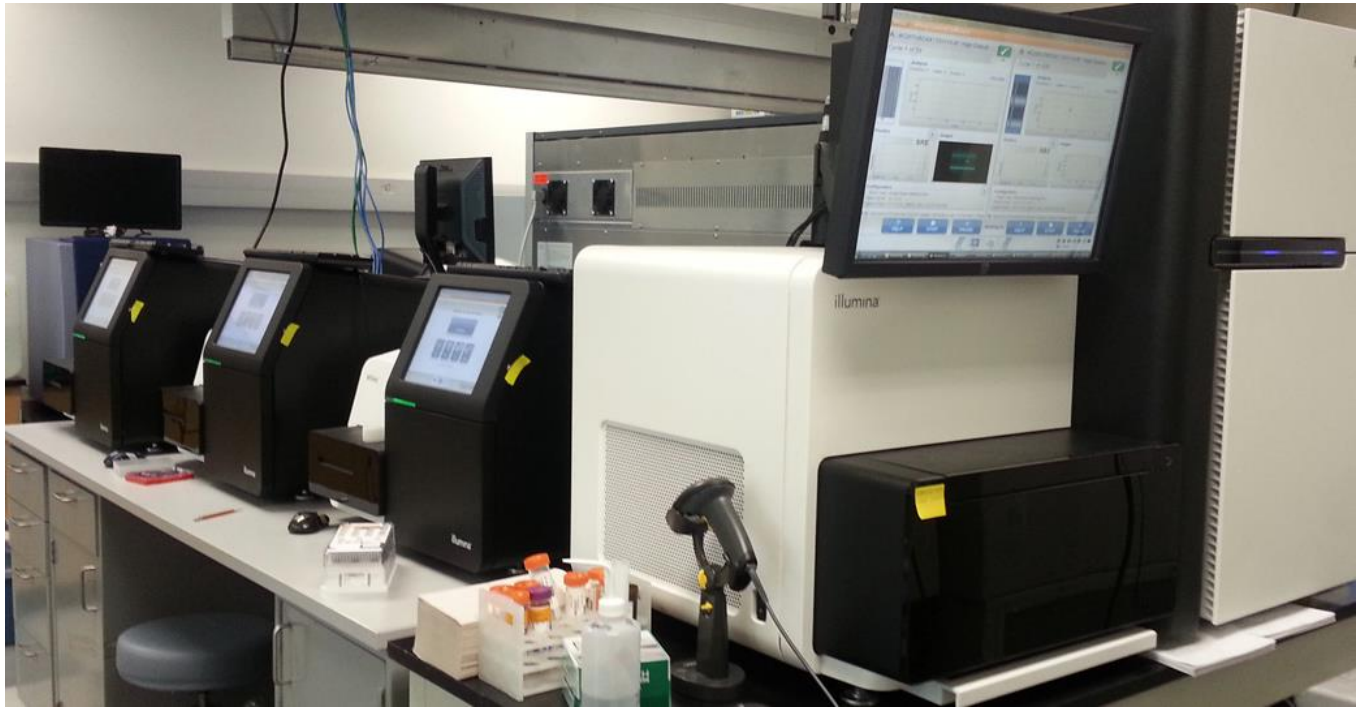


High-Throughput Sequencing for Variant Analyses

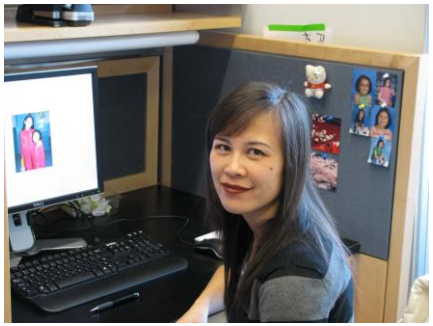
Lutz Froenicke
DNA Technologies & Expression Analysis Cores
UC Davis Genome Center
August 2017

DNA Technologies & Expression Analysis Cores

- HT Sequencing (Illumina & PacBio)
- Illumina microarray (~~expression analysis~~, genotyping)
- consultations
- introducing new technologies to the campus
- shared equipment
- teaching (workshops)

The DNA Tech Core Team

Emily Kumimoto
library preps



Oanh Nguyen
PacBio Seq.

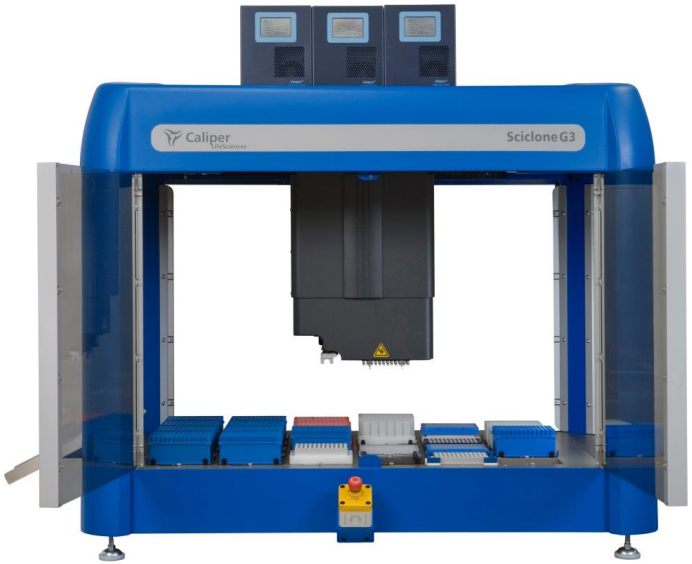
Diana Burkardt-Waco
10X Genomics, HiSeq

Siranoosh Ashtari
all Illumina Seq.

Vanessa Rashbrook
Miseq, Bead Array, Fludigm

Ruta Sahasrabudhe
HMW DNA , Nanopore

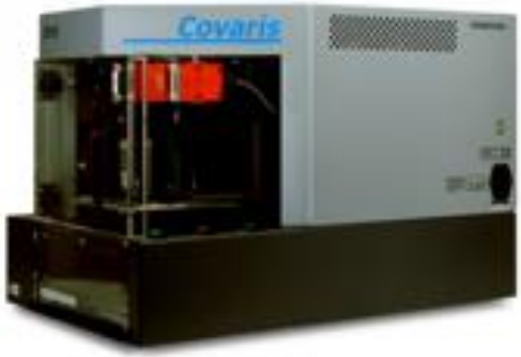
Shared Instruments at the DNA Tech Core



Caliper Sciclone NGS G3

- Plate reader
- Blue Pippin & Pippin HT
- QuantStudio RT-qPCR
- Nanodrop

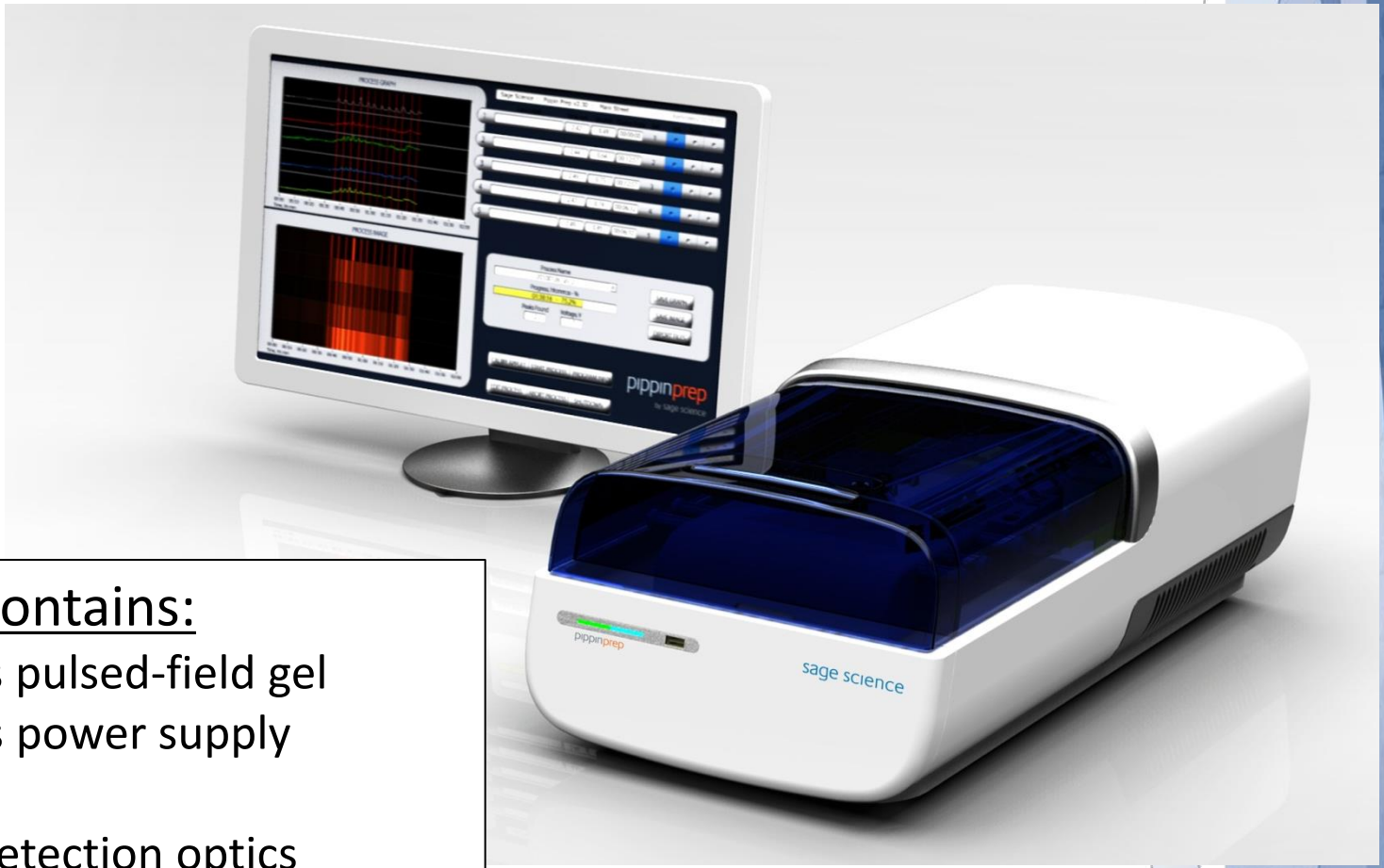
Caliper LabChip GX



Covaris E220 focused ultrasonicator

The Blue Pippin Prep System

Automated Preparative Gel Electrophoresis for NGS



Instrument contains:

Electrophoresis pulsed-field gel electrophoresis power supply
Electrode array
Fluorescence detection optics
Single-board PC with control software

Illumina Infinium arrays

- up to 3 million markers/SNPs
- $100 to $200 per sample (or more)
- custom content pricey
- large scale required

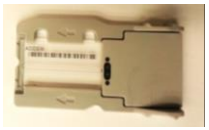
- restricted to known variants

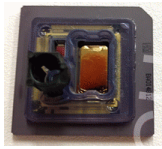
HTS Platform Features









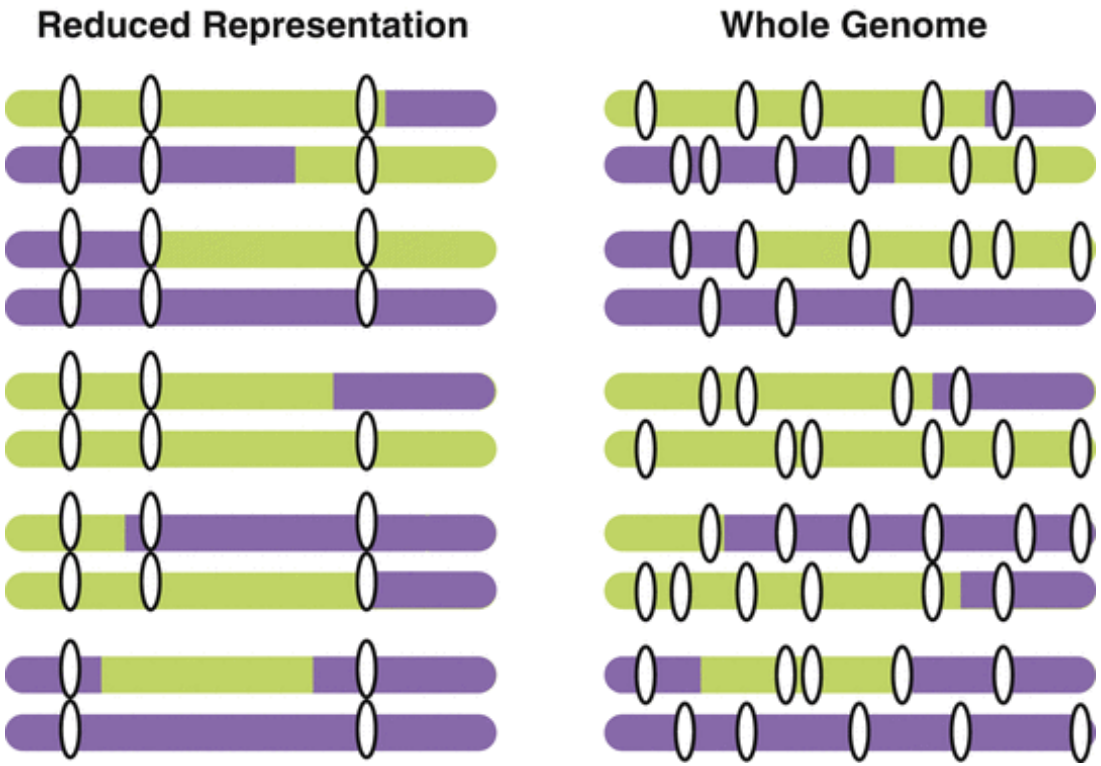
	HiSeq 4000	NextSeq	MiSeq	PacBio RSII	PacBio Sequel
Number of reads	300-400M/lane	300-500M/run	12-15M (v2) 20-25M (v3)	50-80K / SMRT cell	250-400K / SMRT cell
Max. Read Length	2 x 150 bp	2 x 150 bp	2 x 300 bp (v3)	~ 10-60 kb	~ 10-60 kb
Yield per lane (PF data)	up to 100 Gb	up to 150Gb	up to 15 Gb	up to 1.2 Gb	up to 6 Gb
Instrument Time	~1-4 days	~2 days	~2 days	~4.5-6.5 hours	10 hours
Pricing per Gb	$27 (PE150)	$34 to $44 (PE40, PE150)	$130 (PE300)	$350	$180

Studying historic Bean varieties from herbarium samples

- GBS (Genotyping-By-Sequencing)
- 60 year old herbarium samples

Sarah Dohle,
Gepts Lab

Population screening:
As little sequencing as possible



Rowan et al. 2016

- RR: higher confidence calls (multiple reads)

vs.

- WGR: even distribution
- reference available?

Plant Biotechnology Journal

Plant Biotechnology Journal (2017) **15**, pp. 149–161

doi: 10.1111/pbi.12645

Review

Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application

Armin Scheben, Jacqueline Batley and David Edwards*

School of Plant Biology and Institute of Agriculture, University of Western Australia, Perth, WA, Australia

HTS Variant analyses scenarios

- mapping populations
- diversity panels; core collections
- population genetics & evolution
- tumor samples, somatic samples (low frequency variants)

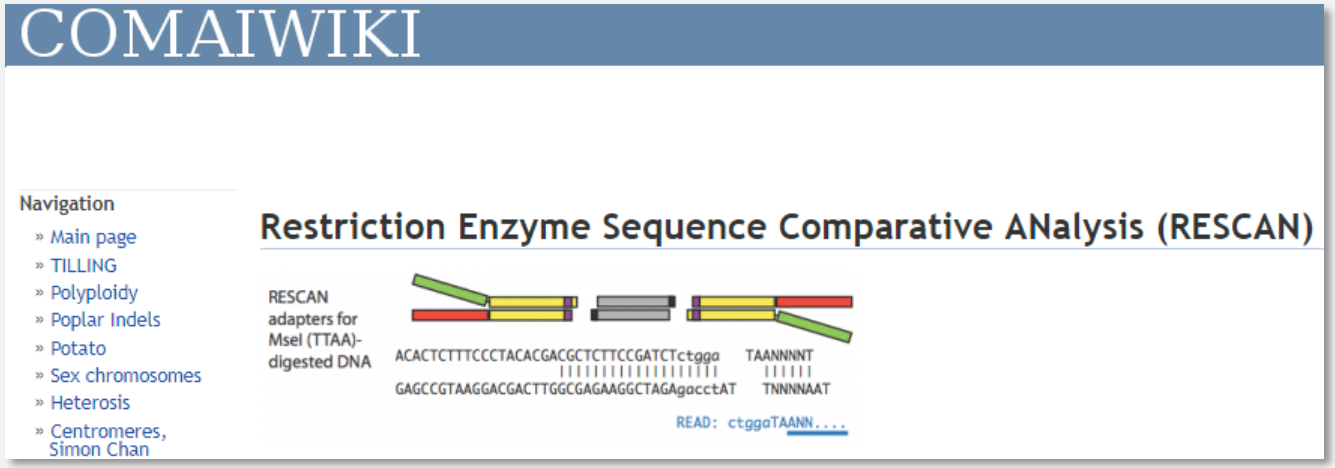
- markers known?
- reference genome or reference-free?
- simple (SNPs) or structural variants?

- genome size?
- high or low genetic diversity?
- size and distribution of haplotype blocks (LD)

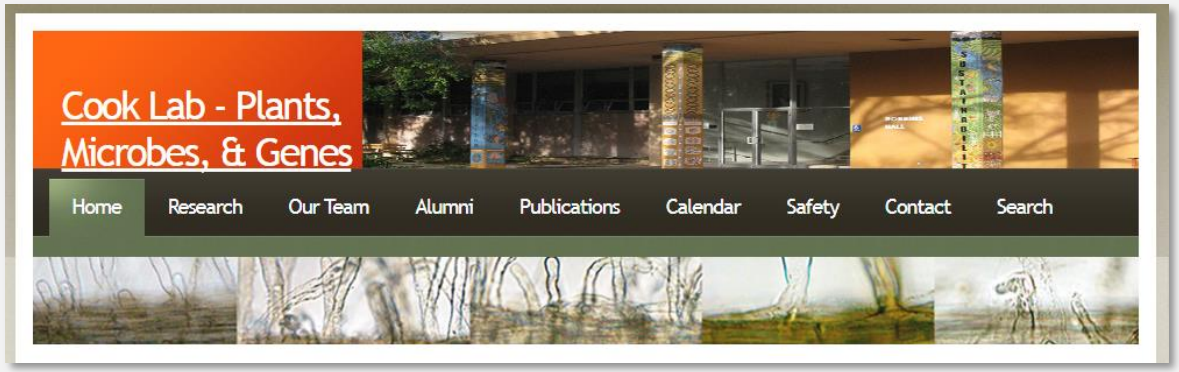
- genetic mapping
- ordering genome assemblies
- GWAS
- MAS

RAD-SEQ & GBS at UC Davis

Miller lab:
RAD-Seq, BEST-RAD, Rapture

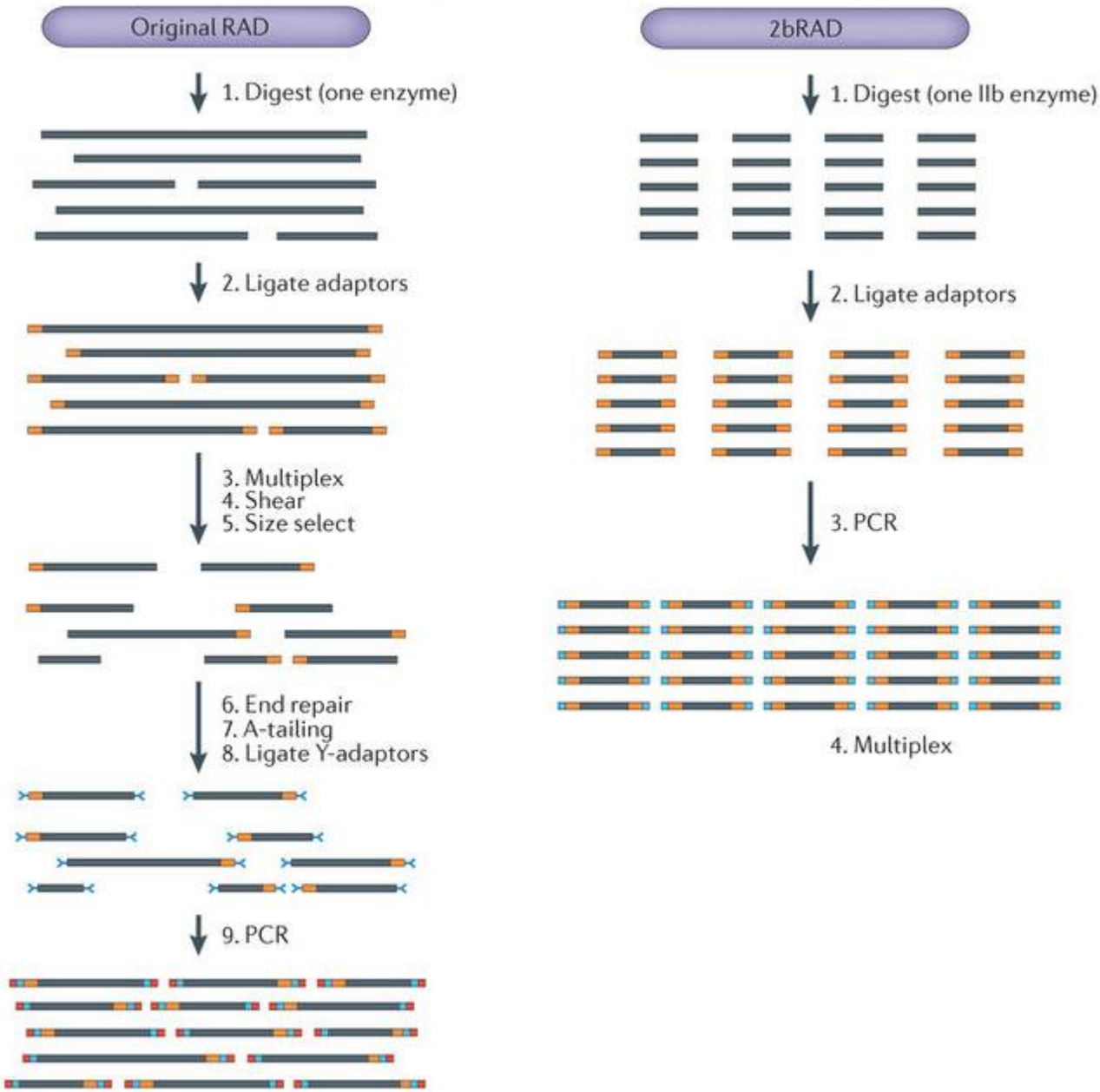


Comai lab:
RESCAN - GBS



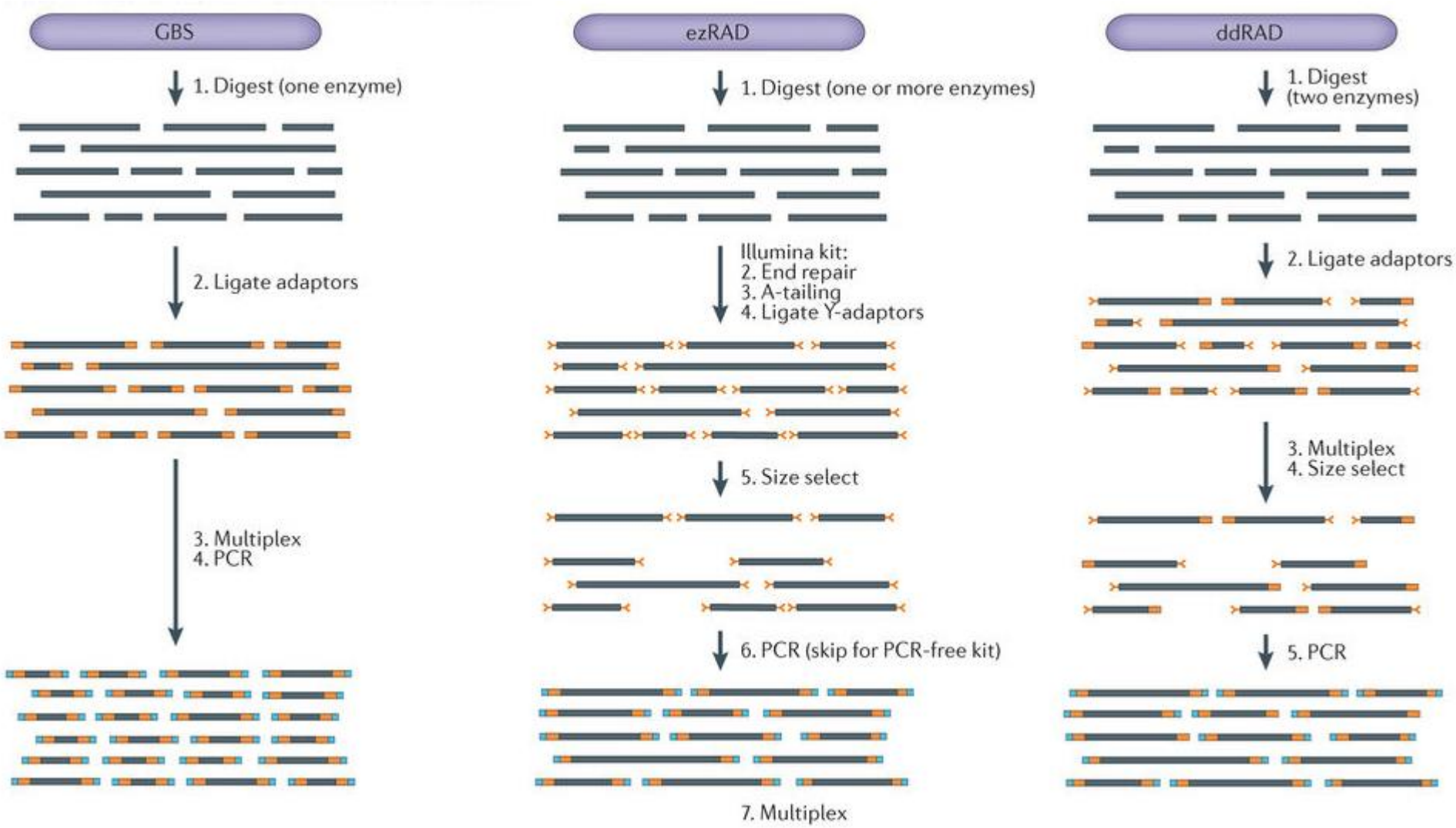
Cook lab:
ddRAD -- legumes

RAD-Seq



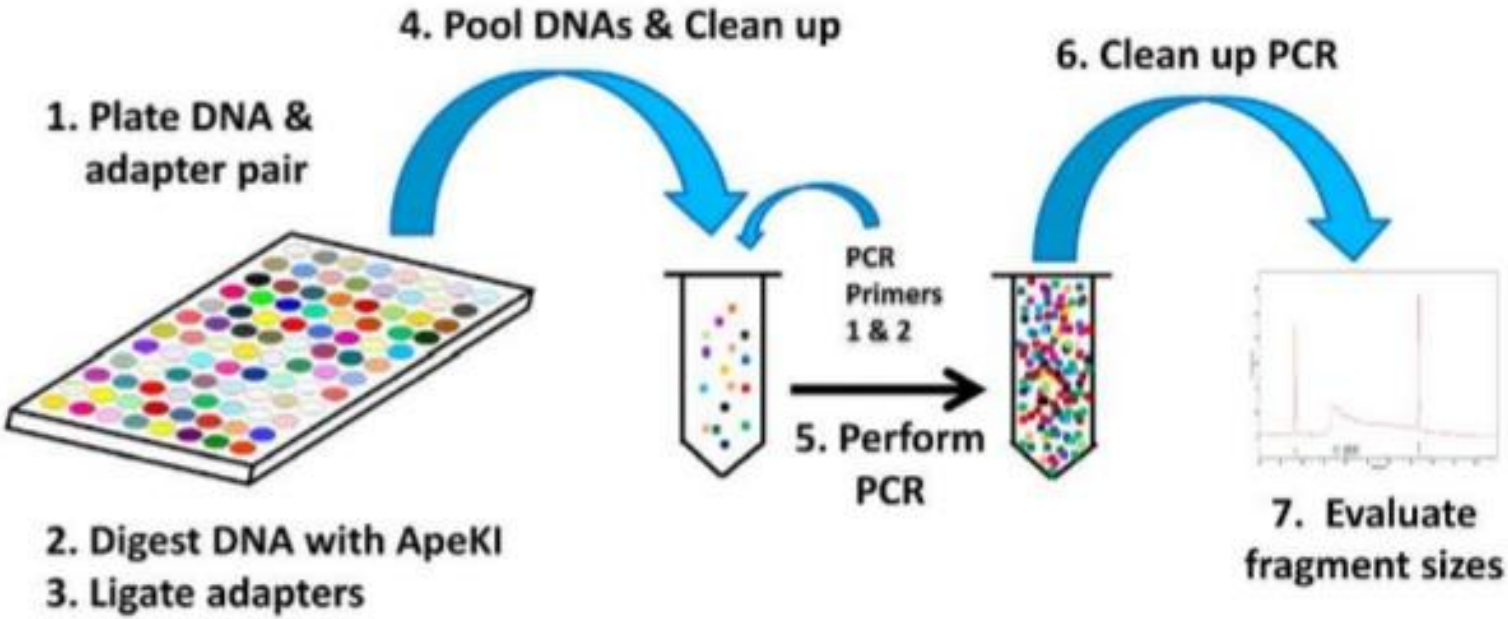
Andrews et al. 2016

RAD-Seq & GBS



Andrews et al. 2016

GBS library construction



(Elshire et al., 2011 PLOS One)

picking the optimal restriction enzyme

MOLECULAR ECOLOGY
RESOURCES

Explore this journal >

Resource Article

DDRADSEQTOOLS: a software package for in silico simulation and testing of double-digest RADseq experiments

F. Mora-Márquez, V. García-Olivares, B. C. Emerson,
U. López de Heredia

First published: 12 July 2016 Full publication history

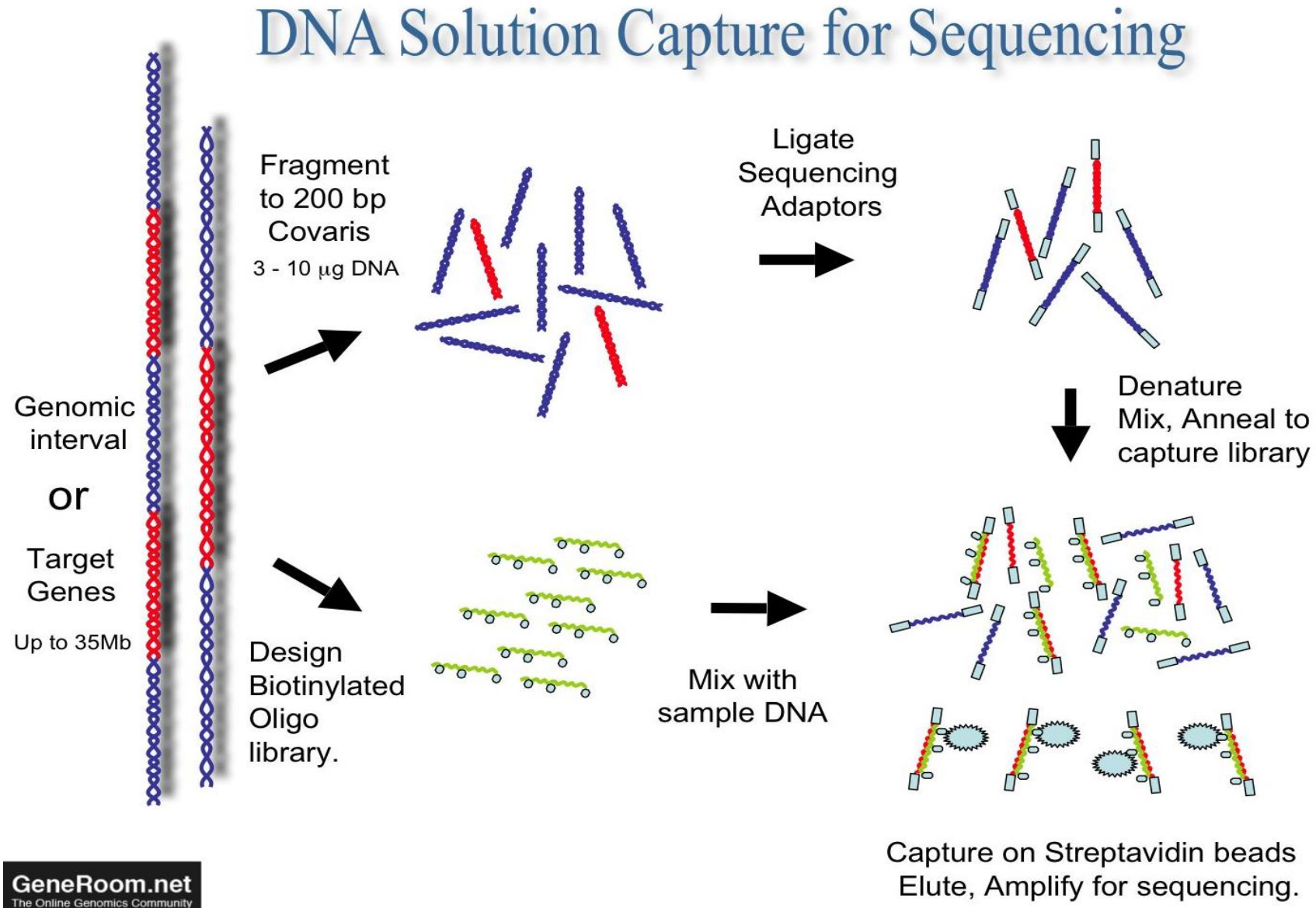
Targeted DNA sequencing

- “targeted capture sequencing”

- exomes
- cancer gene panels (often amplicons)
- any non-repeat ROI
- 2 HiSeq lanes / genome
- 20 exomes / lane

- Combining GBS/RAD with capture:
  → sequencing hundreds of samples per lane (“rapture”)

exome sequencing



comparison of genotyping approaches (Scheben *et al.* 2017)

	Cost per sample[a]	Cost per marker data point[a]	SNP discovery rate	Analysis complexity	Prior genomic knowledge	Preferred population type	Drawbacks	Applications
RADseq	Low	Moderate	Low to moderate	Moderate	No	All	Labour-intensive library preparation; high read depth variation	*De novo* SNP discovery, genome improvement, genetic mapping
Elshire GBS	Low	Moderate	Low	Moderate	No	All	High levels of missing data	*De novo* SNP discovery in simple genomes, genome improvement, genetic mapping
ddRAD	Low	Moderate	Low to moderate	Moderate	No	All	Sensitive to allele dropout; high-quality sample required	*De novo* SNP discovery, genome improvement, genetic mapping
Parental inference WGR	High	Low	High	High	No	Biparental cross	High cost; inference is error-prone	*De novo* SNP discovery, high-resolution mapping of (complex) plant genomes, genome improvement
SkimGBS	High	Low	High	High	Yes	Biparental cross	High cost; need for prior genomic information	SNP discovery and high-resolution mapping of (complex) plant genomes, genome improvement
SNP array	Moderate	High	High	Low	Yes	All	Ascertainment bias; need for prior genomic information	SNP discovery and high-resolution mapping, genetic mapping
Exome sequencing	Moderate	High	Low	Moderate	Yes	All	Need for prior genomic information	SNP discovery in complex genomes, genetic mapping
RNA-seq	Moderate	High	Low	Moderate	No	All	Biases in transcript abundances	SNP discovery in complex genomes, genetic mapping, expression analysis

comparison of genotyping approaches (Scheben *et al.* 2017)

$$$$ est. minimal costs – very dependent on scale, genome size, marker numbers, etc.

	Cost per sample[a]	Cost per marker data point[a]	SNP discovery rate	Analysis complexity	Prior genomic knowledge	Preferred population type	Drawbacks	Applications
RADseq	Low $30	Moderate	Low to moderate	Moderate	No	All	Labour-intensive library preparation; high read depth variation	*De novo* SNP discovery, genome improvement, genetic mapping
Elshire GBS	Low $25	Moderate	Low	Moderate	No	All	High levels of missing data	*De novo* SNP discovery in simple genomes, genome improvement, genetic mapping
ddRAD	Low $35	Moderate	Low to moderate	Moderate	No	All	Sensitive to allele dropout; high-quality sample required	*De novo* SNP discovery, genome improvement, genetic mapping
Parental inference WGR	High	Low	High	High	No	Biparental cross	High cost; inference is error-prone	*De novo* SNP discovery, high-resolution mapping of (complex) plant genomes, genome improvement
SkimGBS	High $200	Low	High	High	Yes	Biparental cross	High cost; need for prior genomic information	SNP discovery and high-resolution mapping of (complex) plant genomes, genome improvement
SNP array	Moderate $80	High	High	Low	Yes	All	Ascertainment bias; need for prior genomic information	SNP discovery and high-resolution mapping, genetic mapping
Exome sequencing	Moderate $300	High	Low	Moderate	Yes	All	Need for prior genomic information	SNP discovery in complex genomes, genetic mapping
RNA-seq	Moderate $180	High	Low	Moderate	No	All	Biases in transcript abundances	SNP discovery in complex genomes, genetic mapping, expression analysis

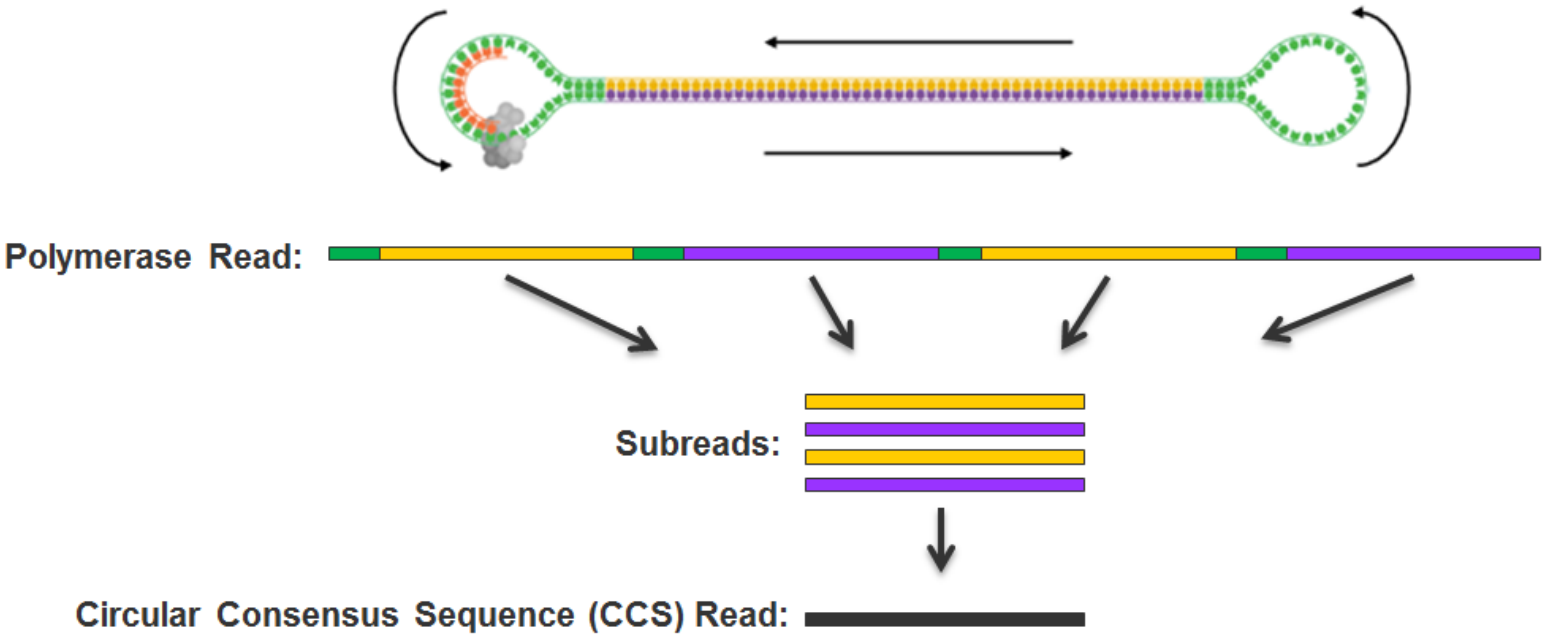
- Exome sequencing & RNA-seq:
  → variants of biological interest

- GBS with methylation sensitive RE:
  → plants: avoid repetitive elements

- Small numbers of markers:
  → amplicon sequencing

- Phasing of variants in specific regions:
  → Pacbio sequencing

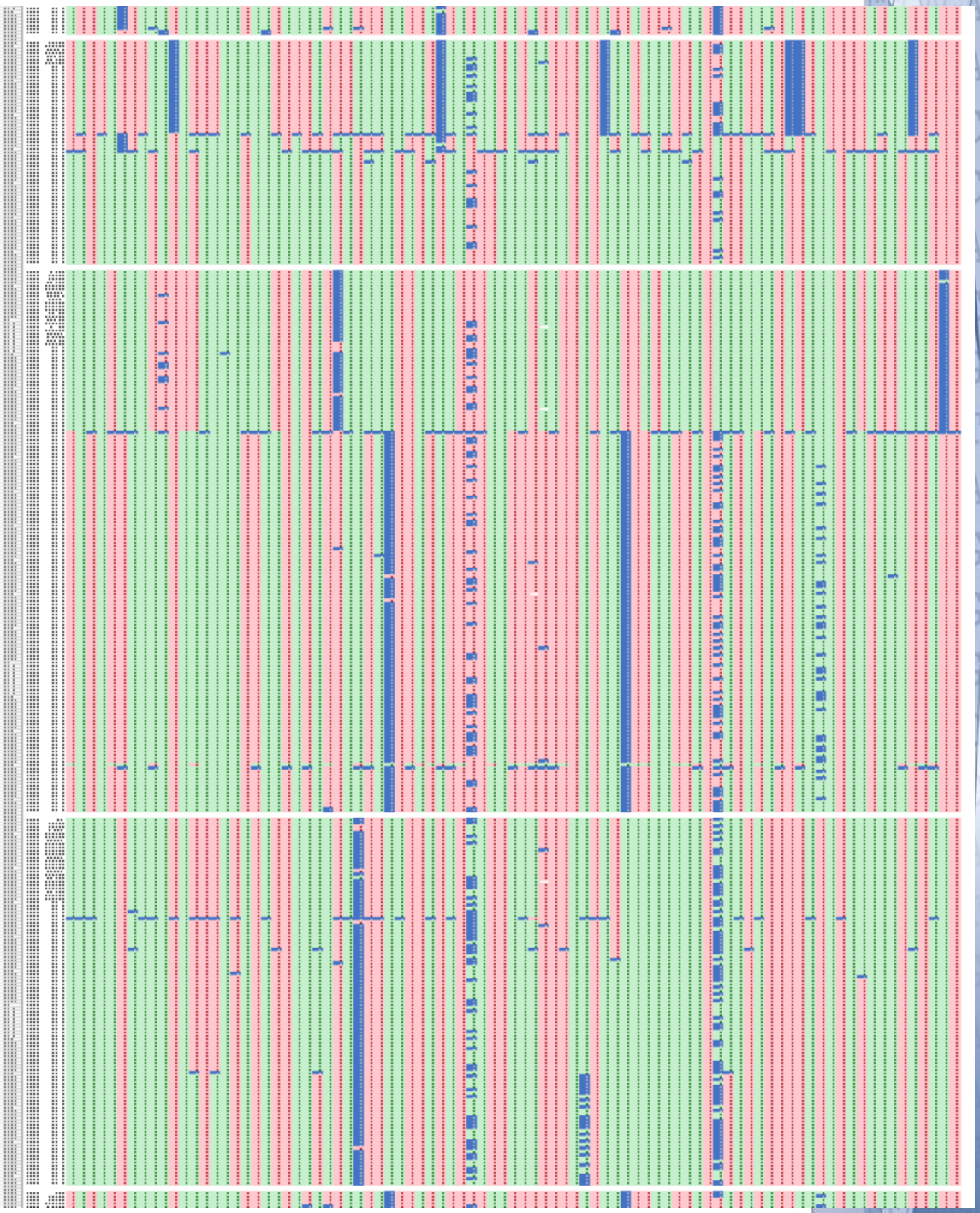
PacBio SMRT-bell adapters circular sequencing



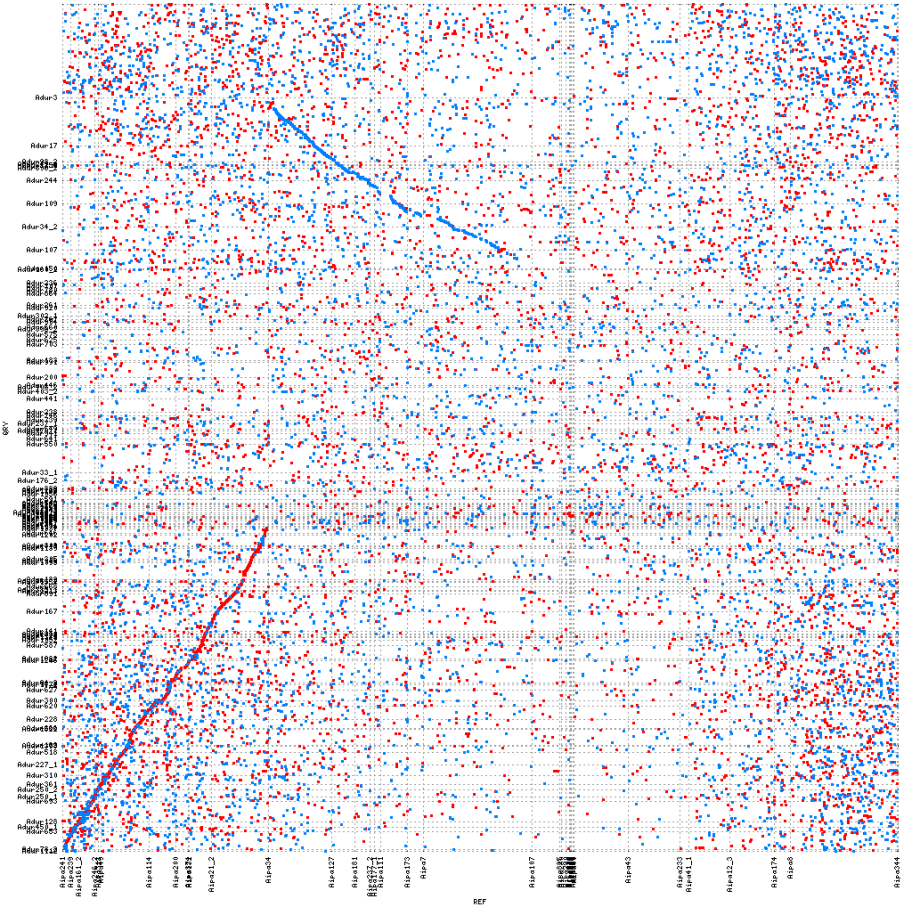
Skim-seq low coverage sequencing

Haplotyping of groups of 20 to 100 SNPs

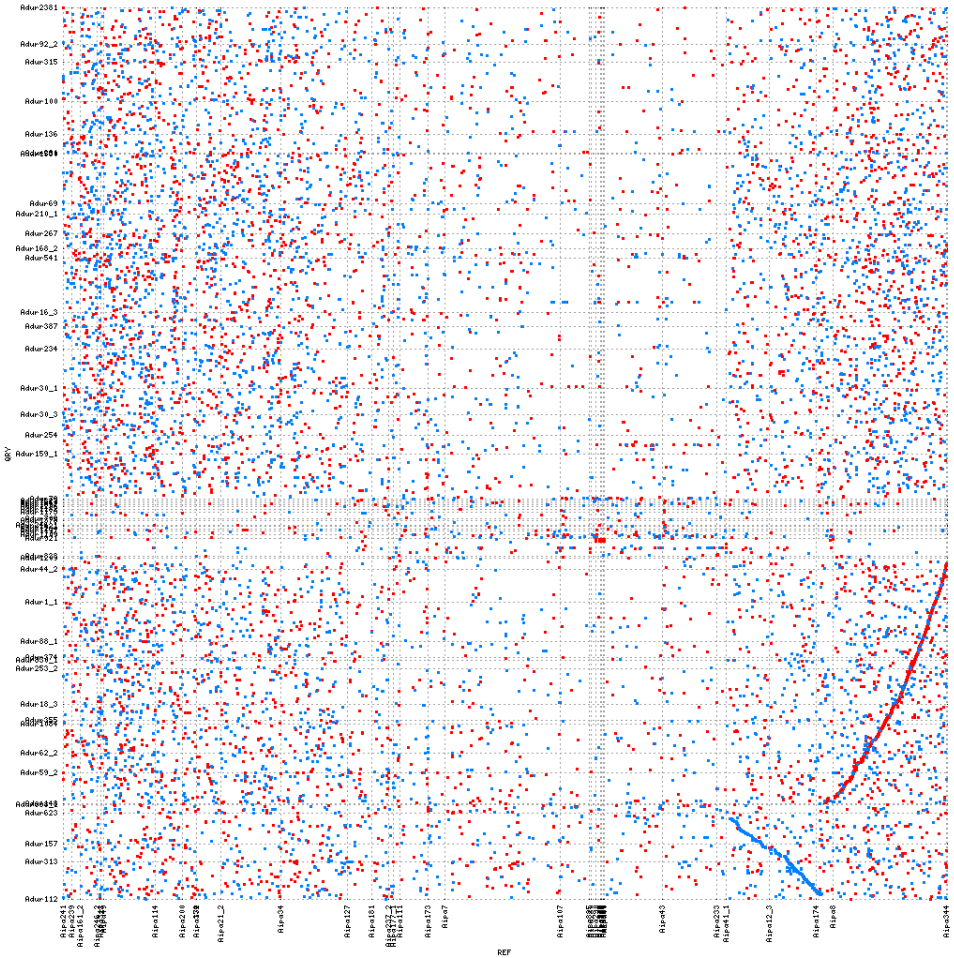
Assembly errors can be easily spotted



altogether: 7 inv, 2 transl.

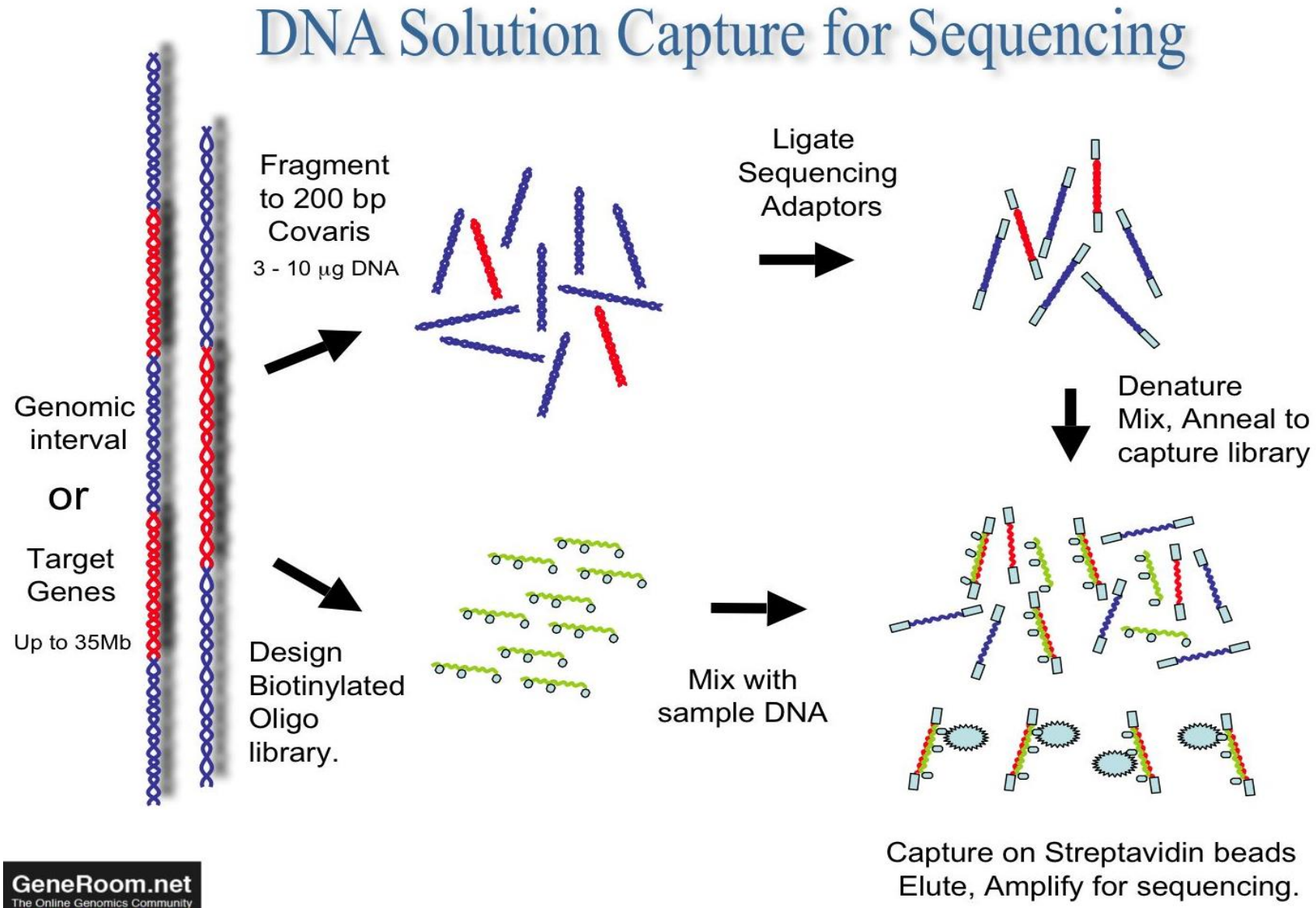


B7/A7



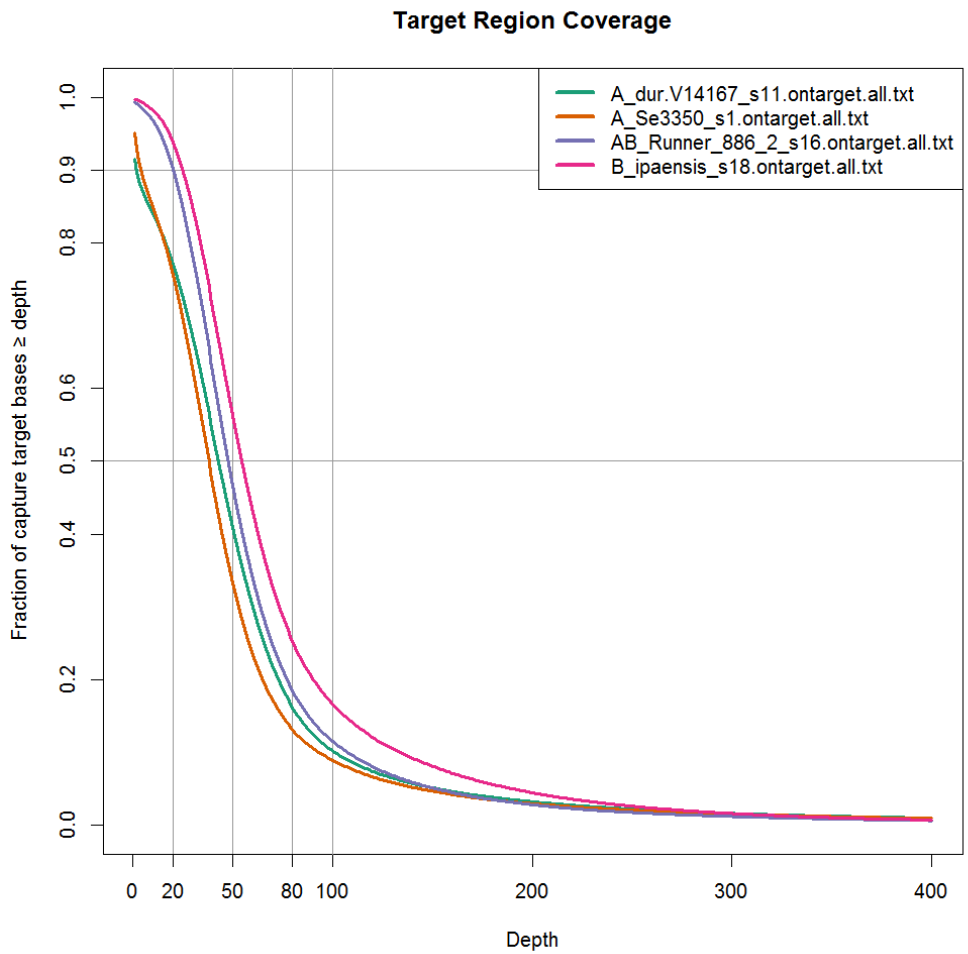
B7/A8

exome sequencing



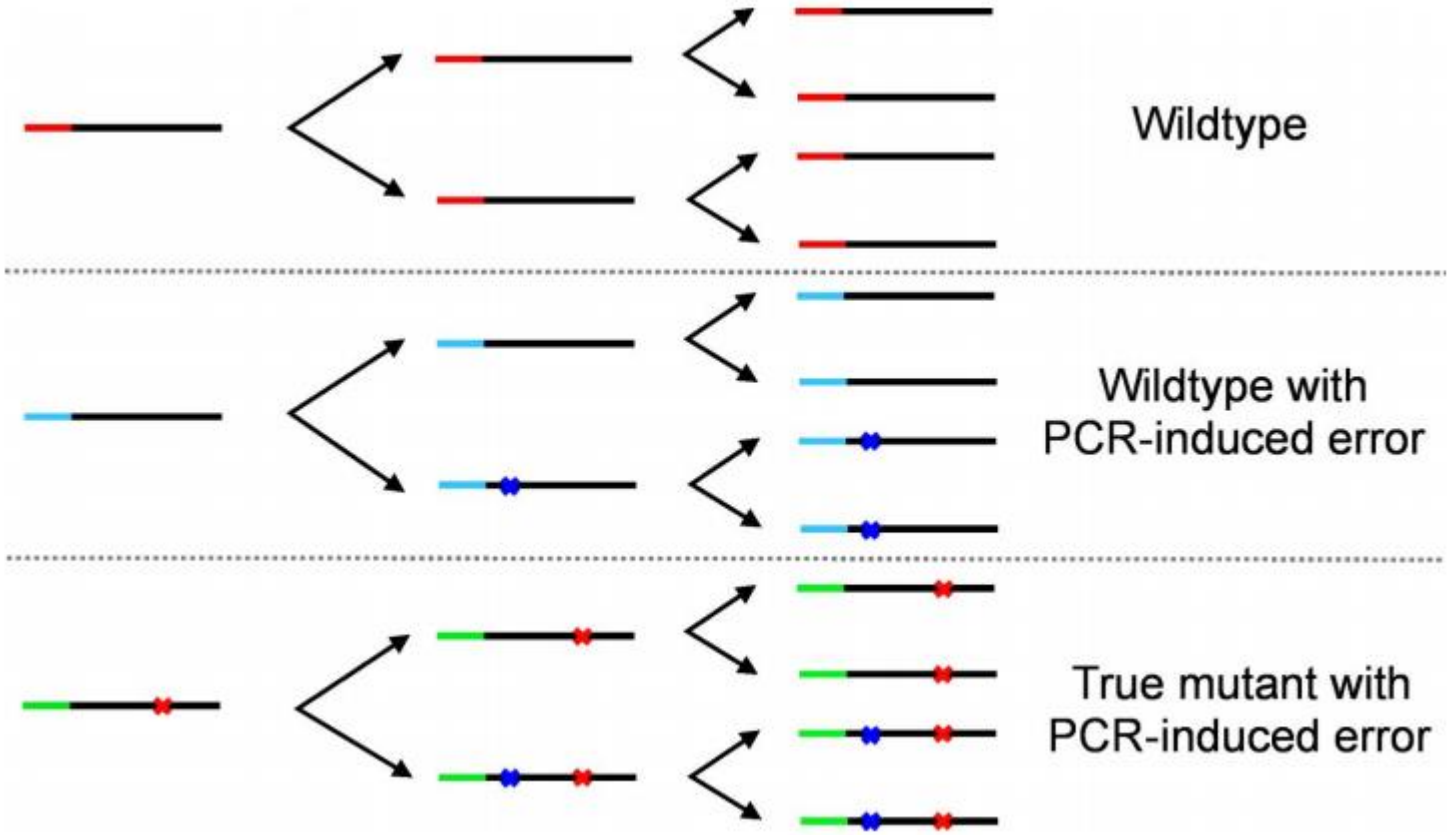
exome capture coverage distribution

- Peanut exome sequencing



Eliminating sequencing errors with UMIs
→ low frequency variants & high coverage sequencing

- random and unique barcodes as Unique Molecular Identifiers



Stahlberg et al . 2017

Eliminating sequencing errors with UMIs for Amplicons

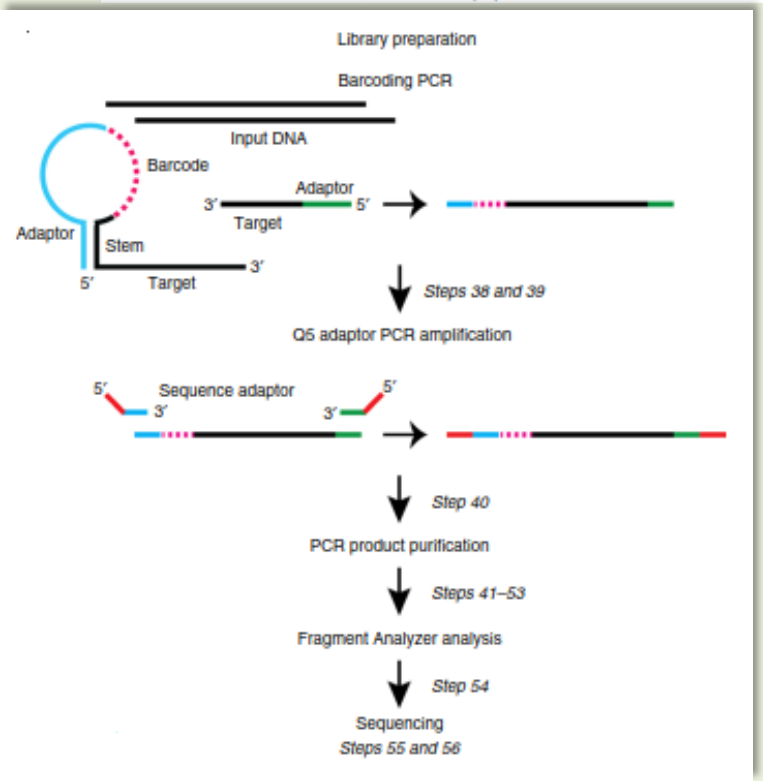
PROTOCOL

Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing

Anders Ståhlberg[1,4], Paul M Krzyzanowski[2,4], Matthew Egyud[3], Stefan Filges[1], Lincoln Stein[2] & Tony E Godfrey[3]

[1]Department of Pathology and Genetics, Sahlgrenska Cancer Center, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden. [2]Ontario Institute for Cancer Research, MaRS Centre, Toronto, Ontario, Canada. [3]Department of Surgery, Boston University School of Medicine, Boston, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to A.S. (anders.stahlberg@gu.se) or T.E.G. (godfreyt@bu.edu).

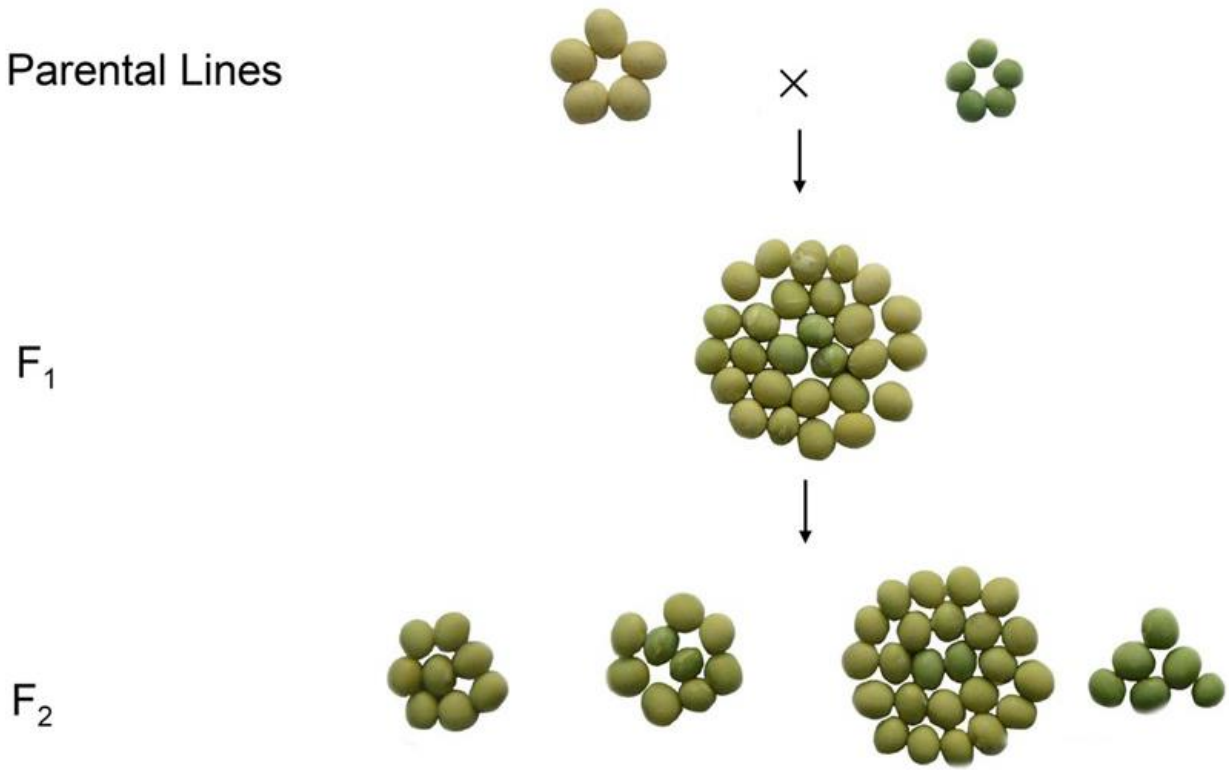
Published online 2 March 2017; doi:10.1038/nprot.2017.006



- **SiMSen-Seq**
- Hairpin PCR primer with **UMI** for first two cycles
- 2 PCR steps ; the second at very high annealing temperature

BSA-sequencing

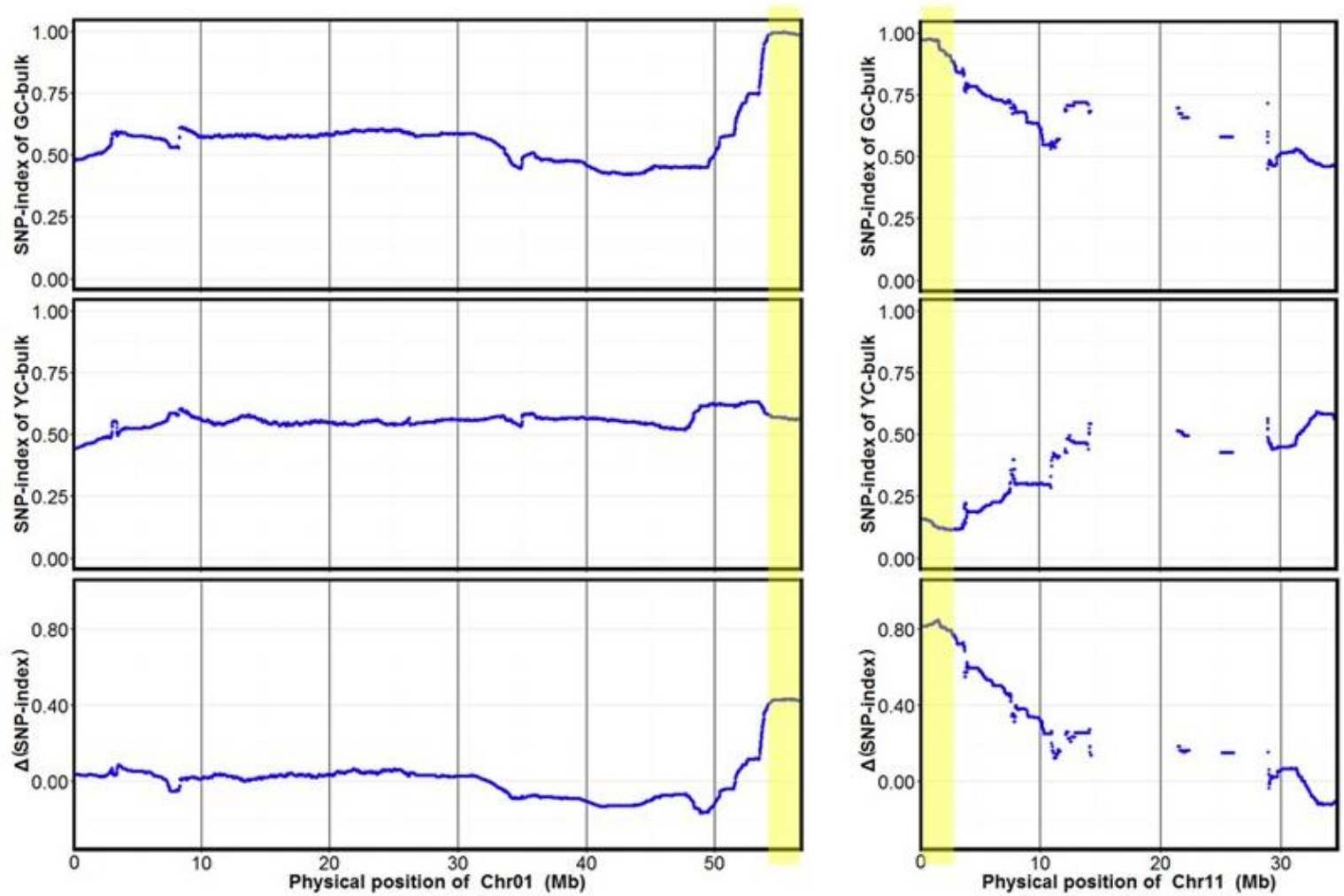
- Finding causal genes (simple traits)
- Segregating population



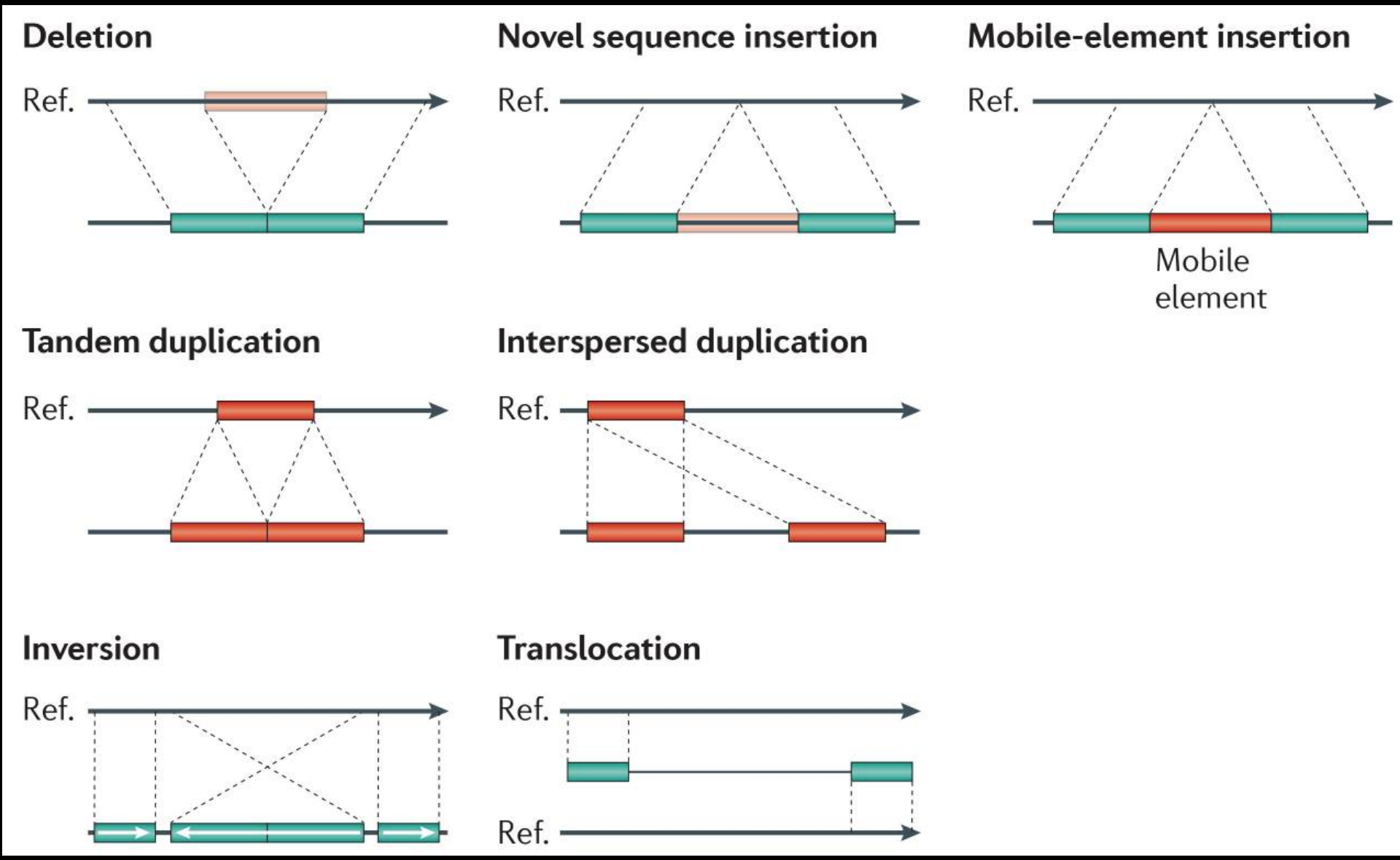
Song et al. 2017 Soy bean BSA for yellow or green cotyledon

BSA-sequencing

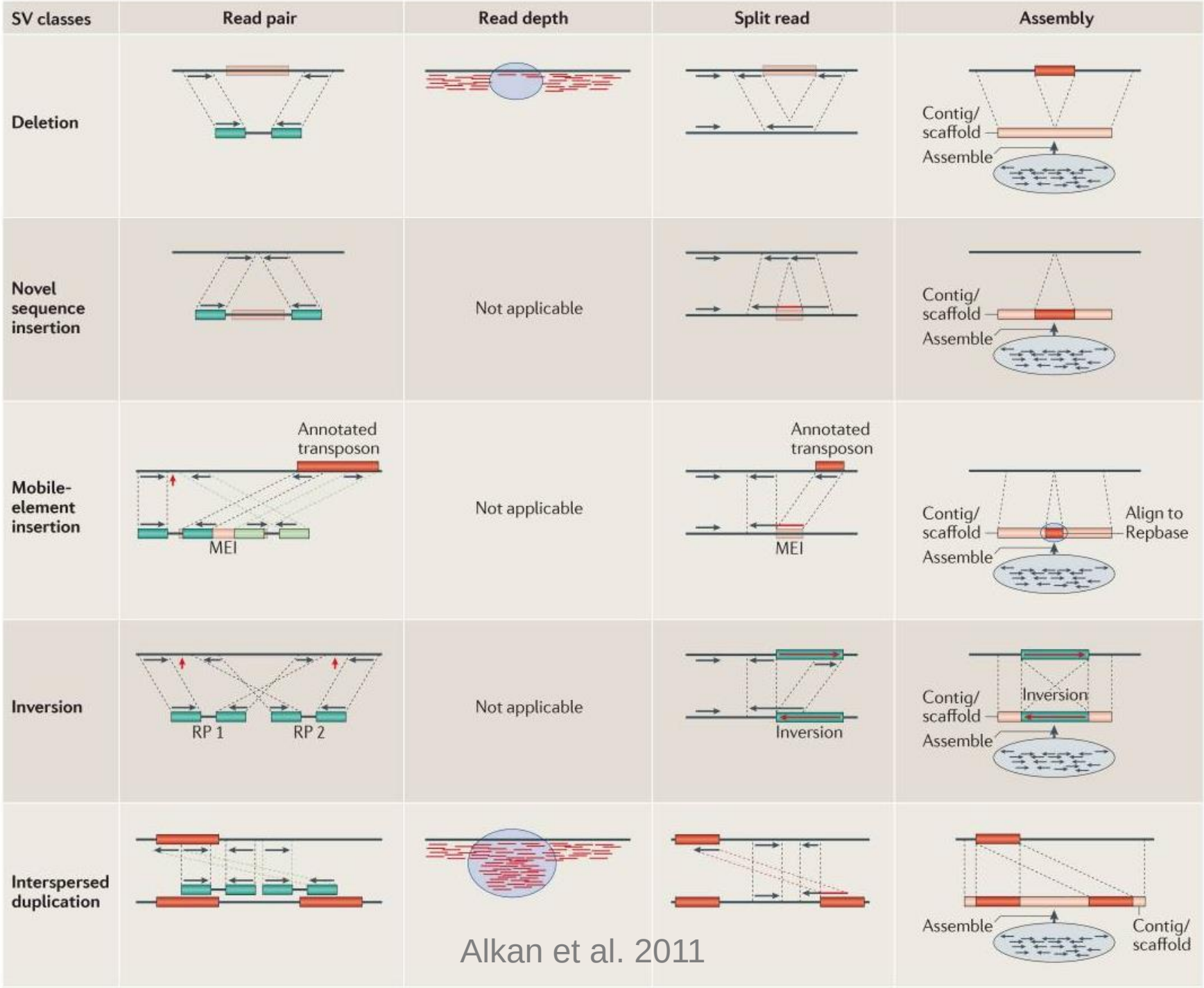
- high SNP-index indicates candidate regions



Song et al. 2017

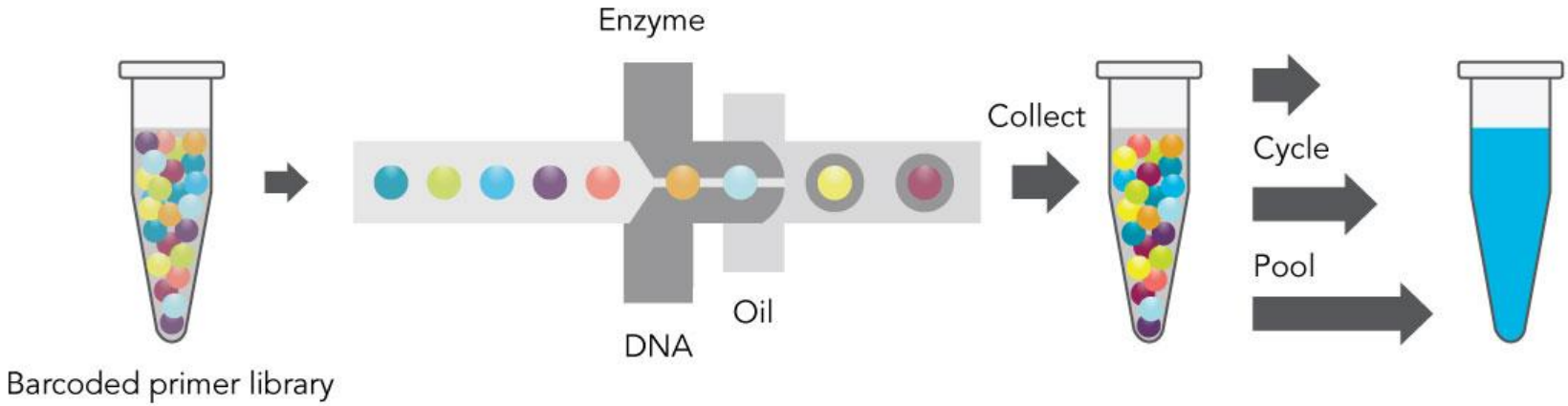


Alkan et al. 2011

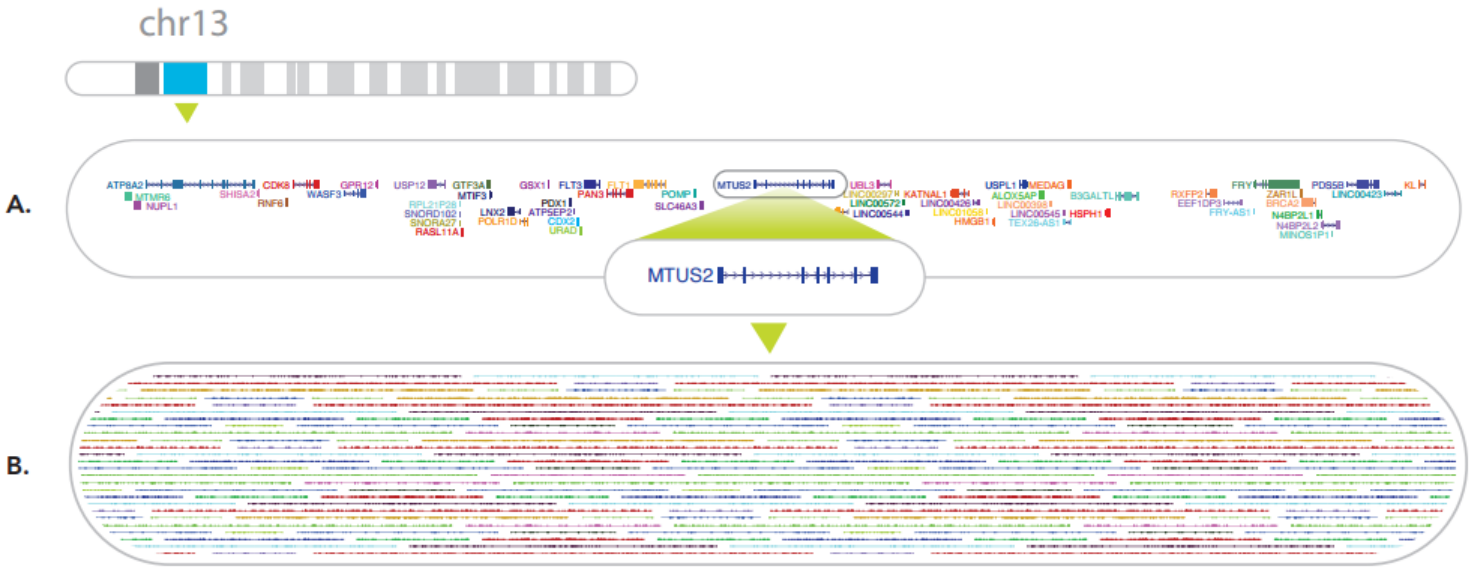


Alkan et al. 2011

10X Genomics
(genomic DNA analysis and single cell RNA)

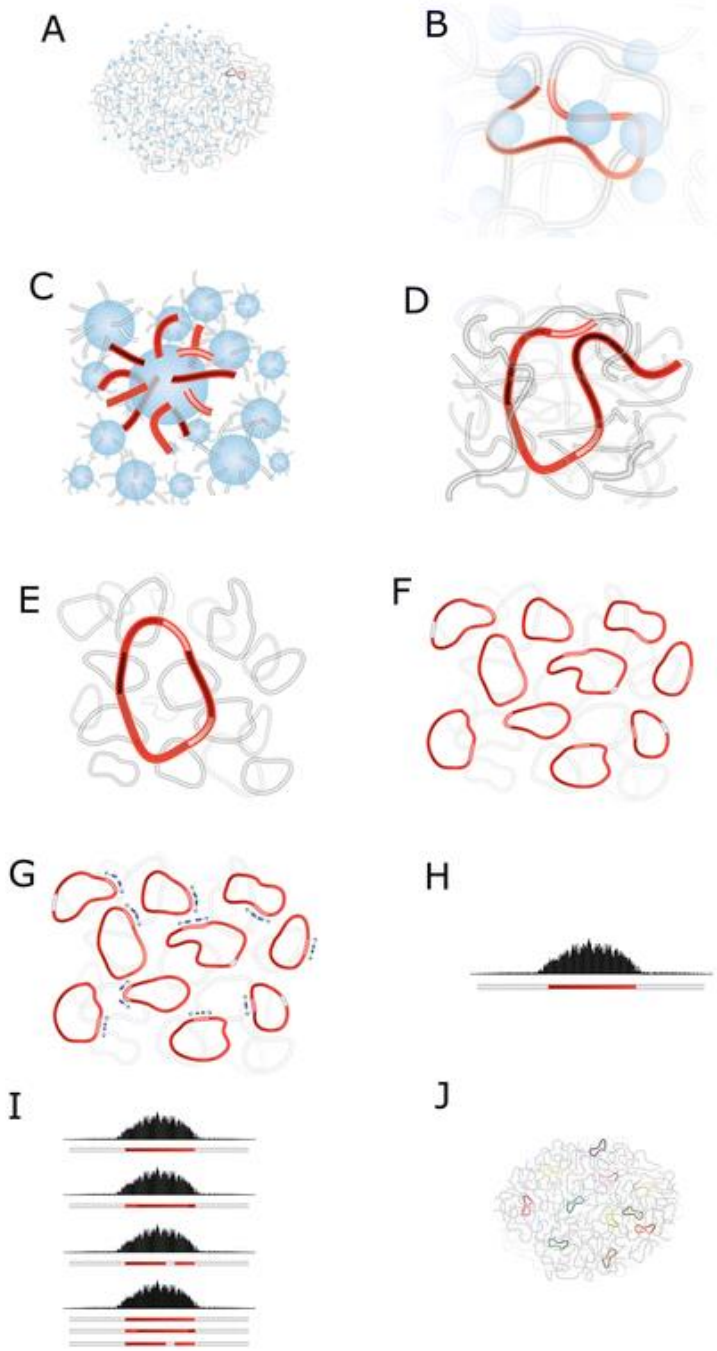

Enzyme
Collect
Cycle
Pool
Oil
DNA
Barcoded primer library

25x linked –read coverage



- 60 kb deletion

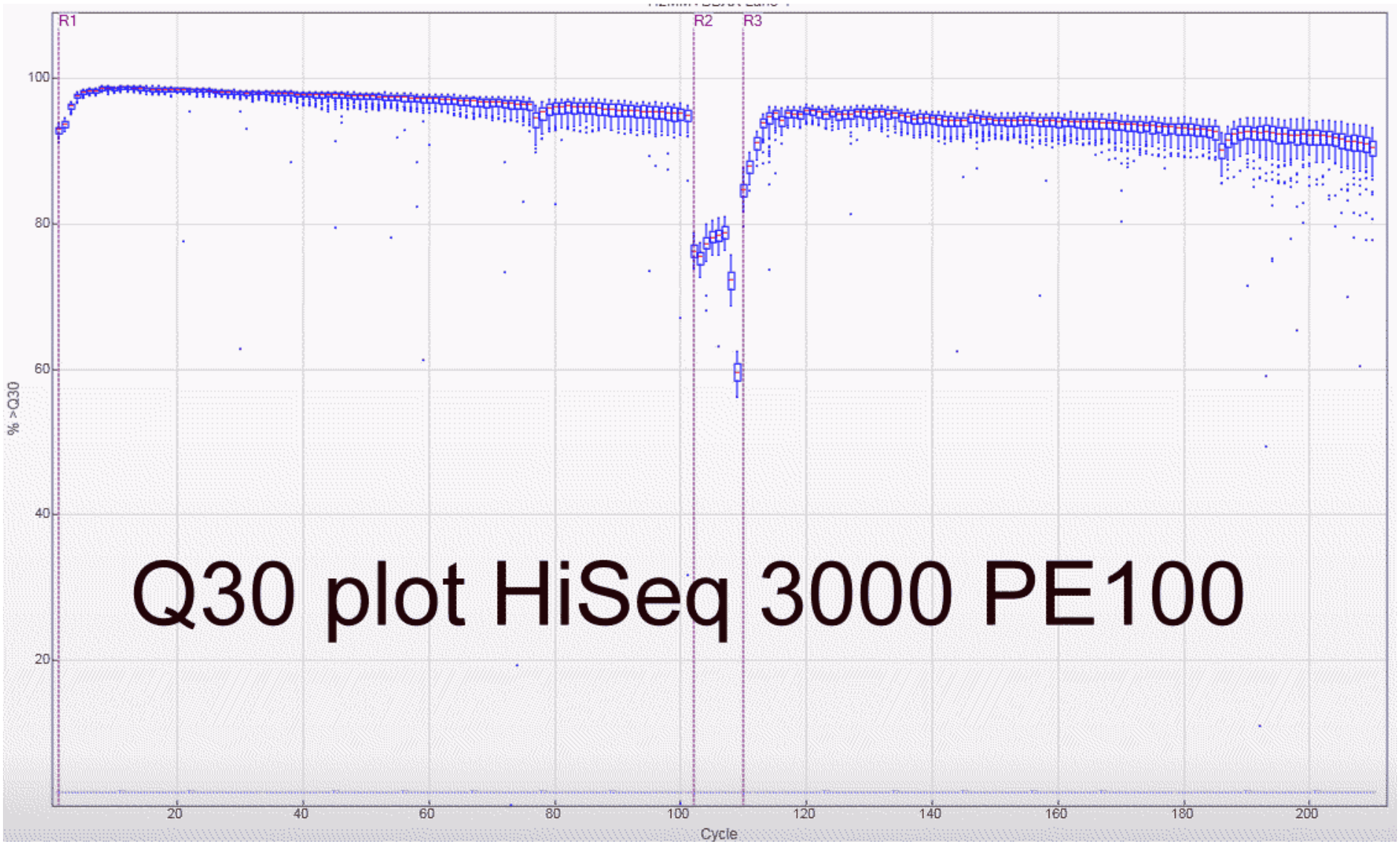

10X GENOMICS
Loupe · Sample 5728
File
SUMMARY
PHASING
STRUCTURAL VARIANTS
chr2+29989813-chr2+30039813,chr2+42493841-chr2+42543841
chr2
29,995,000
30,000,000
30,005,000
30,010,000
30,015,000
30,020,000
30,025,000
30,030,000
30,035,000
HAPLOTYPE 1
UNPHASED
HAPLOTYPE 2
GENES
ALK
EXONS
COVERAGE
BREAKPOINTS
SV #61
chr2
42,495,000
42,500,000
42,505,000
42,510,000
42,515,000
42,520,000
42,525,000
42,530,000
42,535,000
42,540,000
HAPLOTYPE 1
UNPHASED
HAPLOTYPE 2
GENES
EML4
EXONS
COVERAGE
BREAKPOINTS
23
SV #61
SV #71
SV #279



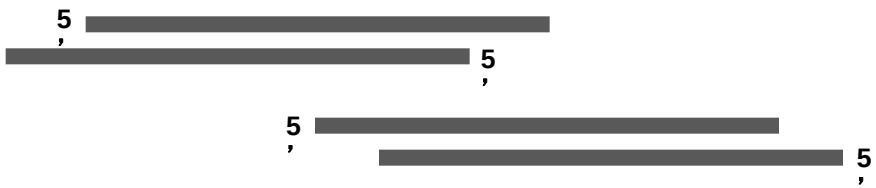
TLA

- Targeted Locus Amplification
- Hottentot et al . 2016
- up to 100 kb regions

Illumina SAV viewer



DNA library construction



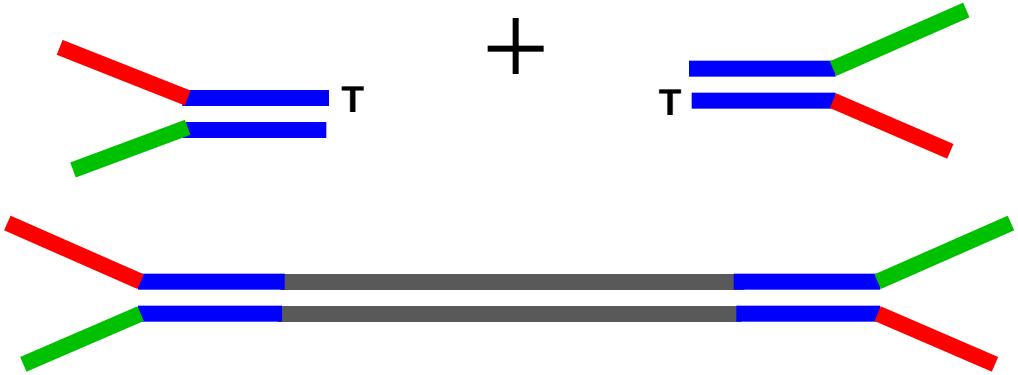
Fragmented DNA

↓ End Repair
(ignore variants at ends?)

Blunt End Fragments

↓ "A" Tailing

Single Overhang Fragments



↓ Adapter Ligation

DNA Fragments with Adapter Ends

PACIFIC BIOSCIENCES™

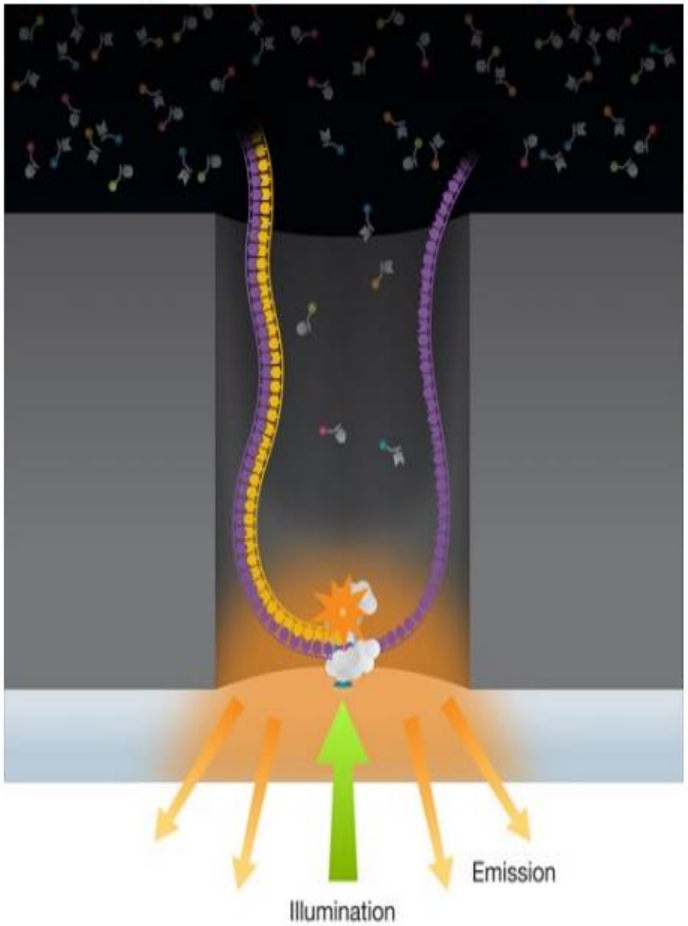
http://pacificbiosciences.com

THIRD GENERATION DNA SEQUENCING

Single Molecule Real Time (SMRT™) sequencing
Sequencing of single DNA molecule by single polymerase
Very long reads: average reads over 15 kb, up to 60 kb
High error rate (~15%).
Complementary to short accurate reads of Illumina

Third Generation Sequencing : Single Molecule Sequencing

Pacific Biosciences

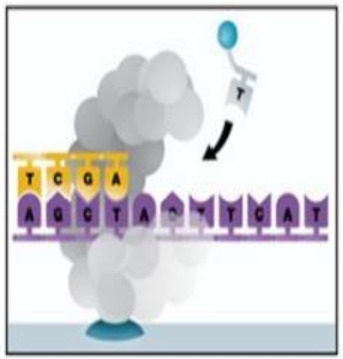


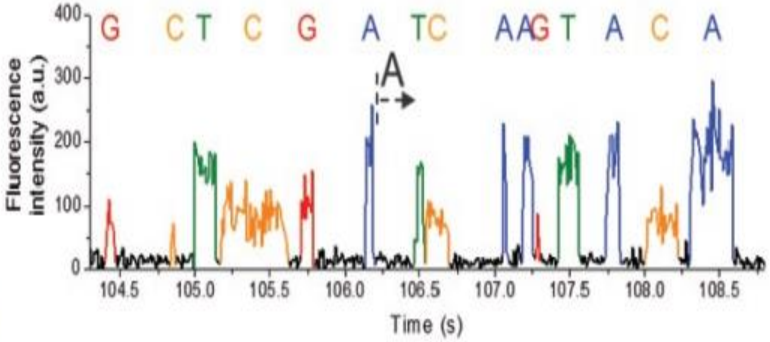
4 nucleotides with different fluorescent dye simultaneous present

2-3 nucleotides/sec
2-3 Kb (up to 50) read length
6 TB data in 30 minutes

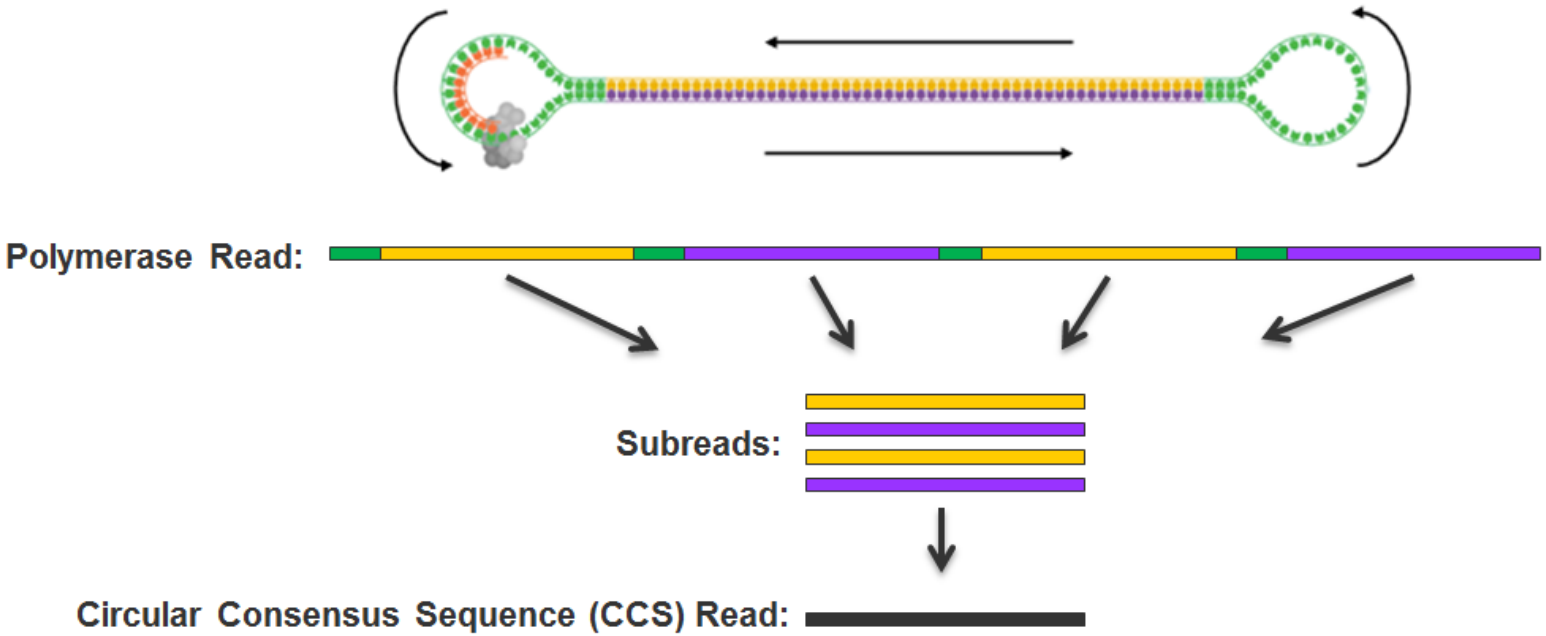
laser damages polymerase

70 nm aperture
“Zero Mode Waveguide”

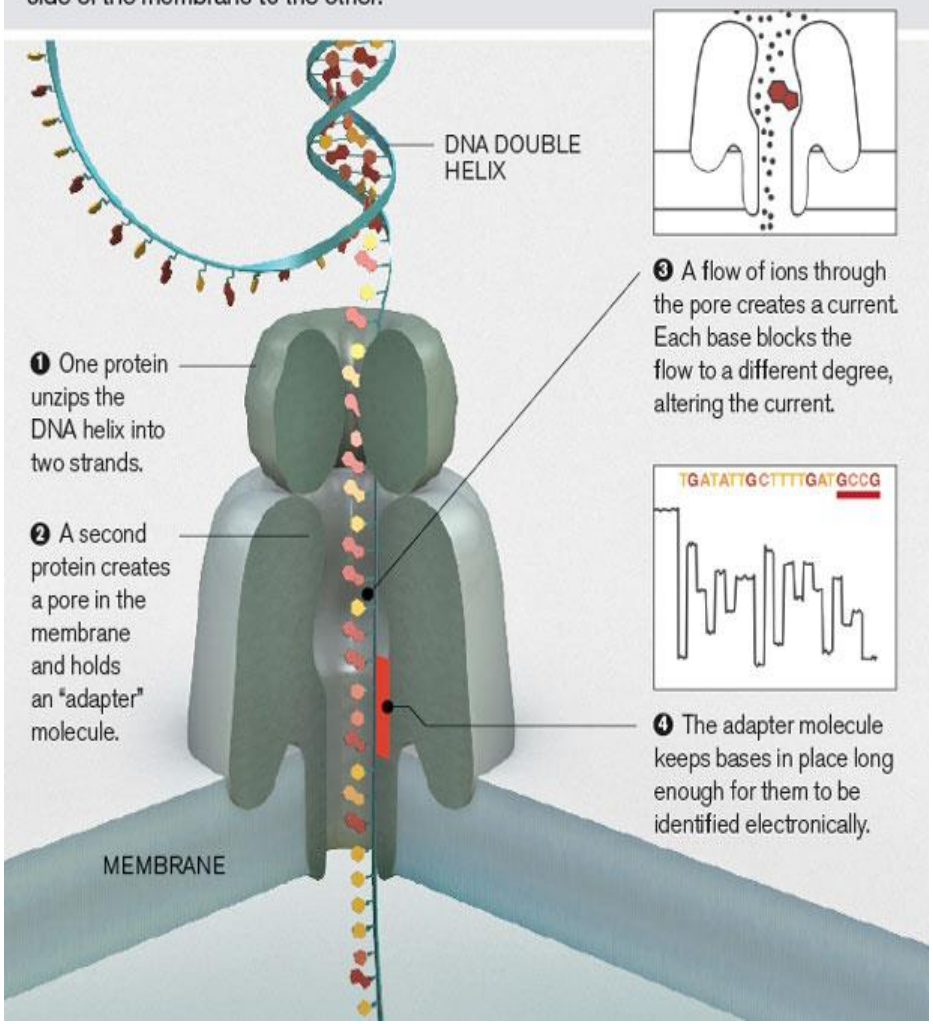




SMRT-bell adapters circular sequencing



DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.



Future's so bright

Thank you!

Let’s get started!