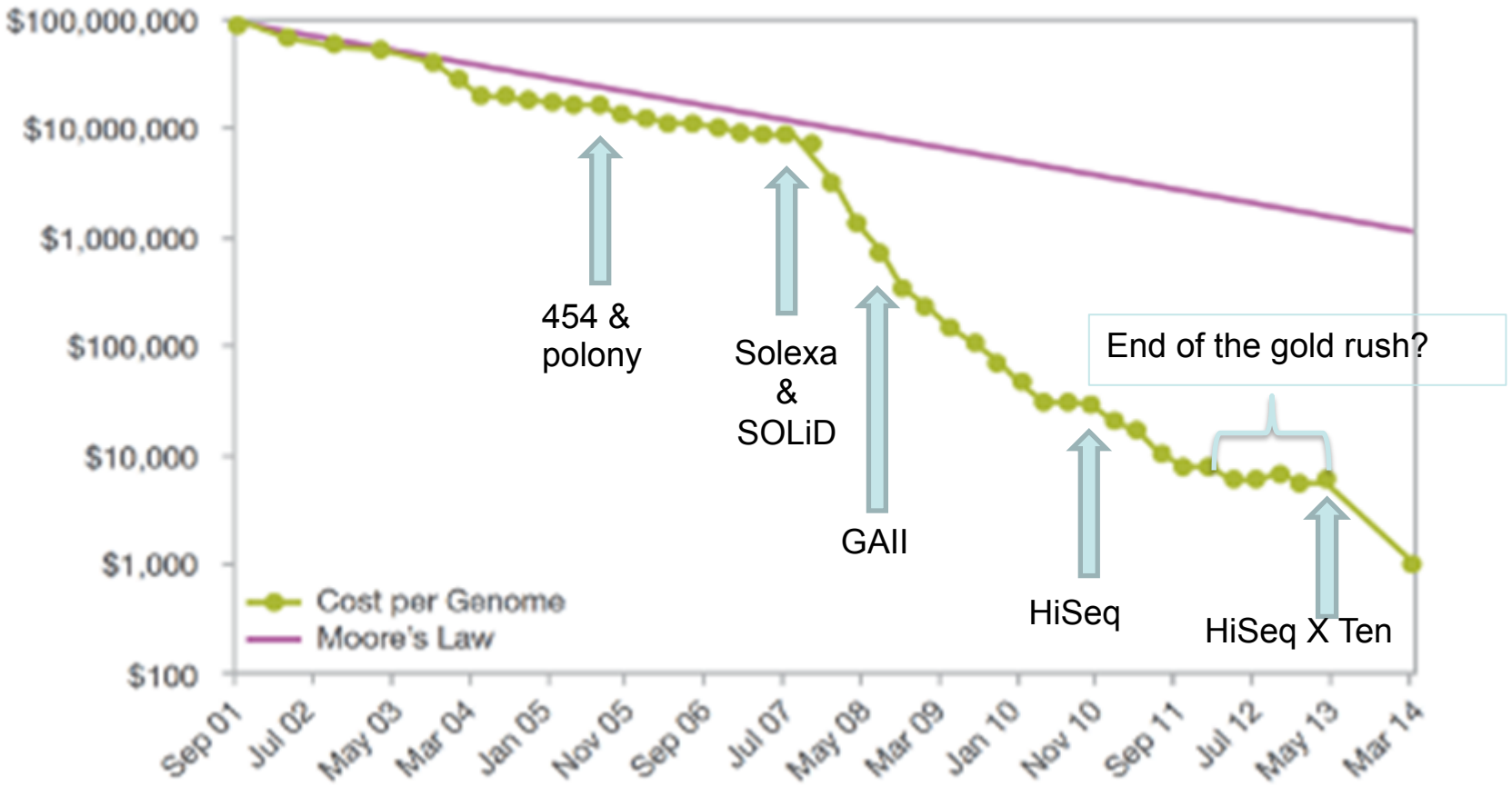
The Open Door Workshop

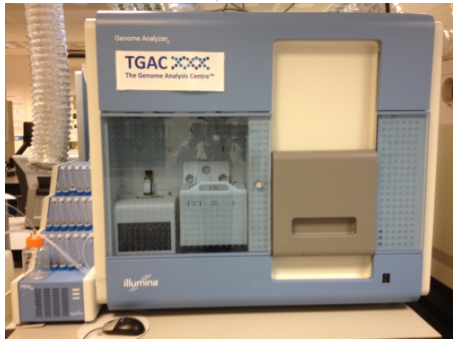
Next Generation Sequencing

Matthew D. Clark PhD
Group leader
Genomics, The Genome Analysis Centre
Norwich, UK

Costs & disruptive technologies



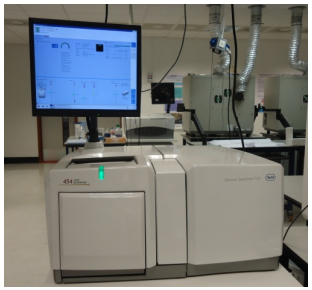
Illumina



GAII x 1

HiSeq2500 x 4
HiSeq2000 X 1

MiSeq x 3



Roche 454FLX x 2

PacBio RSII x 1

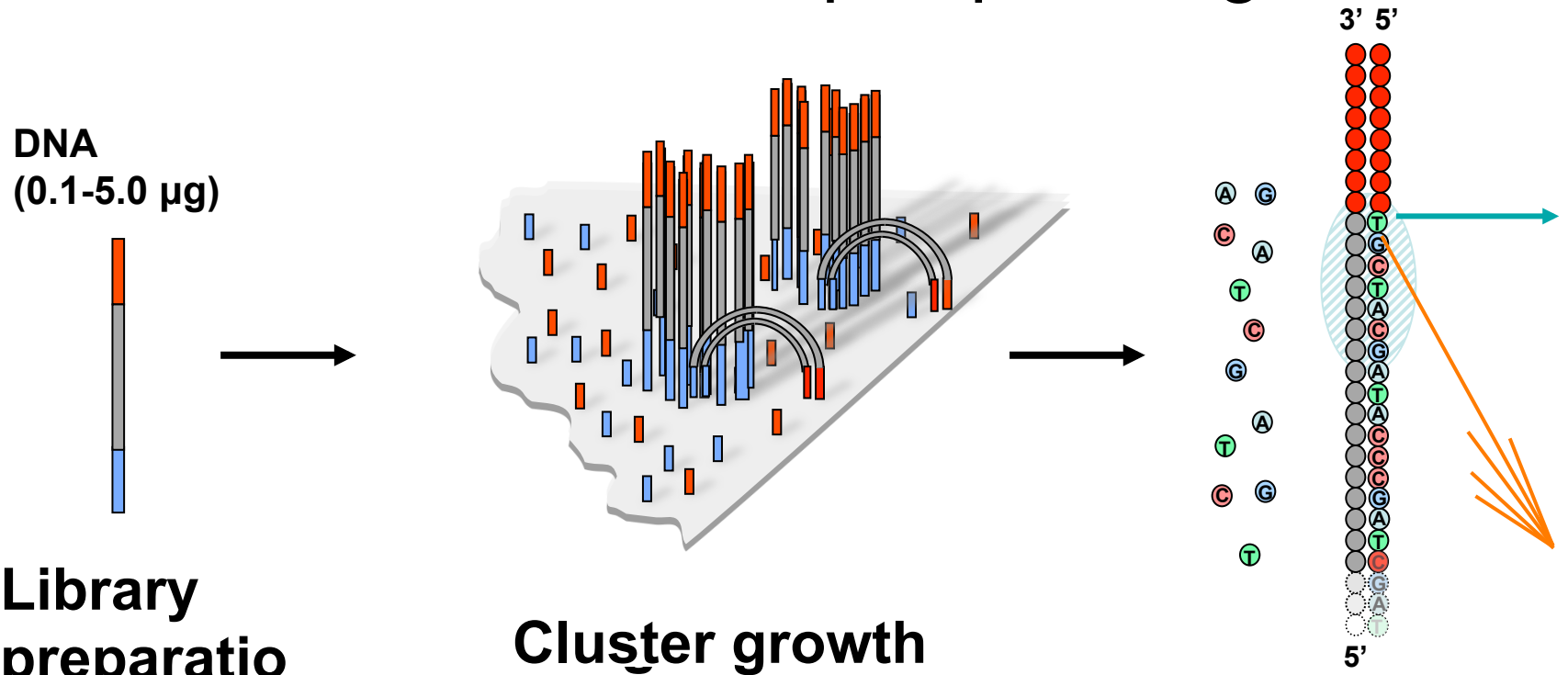
Ion Proton x 1

MinIon x 2

Sequencer comparison

	METHOD	READ	NUMBER OF READS	Yield	TIME
Illumina	SBS	100-250bp PE	<200 million/lane	30-80 Gb	1-11 days
454	Pyrosequencing	4-800bp	1 million /plate	4-800 Mb	10 hours
PacBio	Single molecule	10-15,000 bp	>50,000 /SMRT cell	500+ Mb	1-4 hours
Ion Proton	Pyrosequencing	~170bp SE	~70 million	9 Gbp	3 hours
MinIon	Nanopore	300-50,000bp	1-50,000	0.1-1Gbp	1-48hrs

Illumina HiSeq sequencing



Library preparation

Cluster growth array

Sequencing

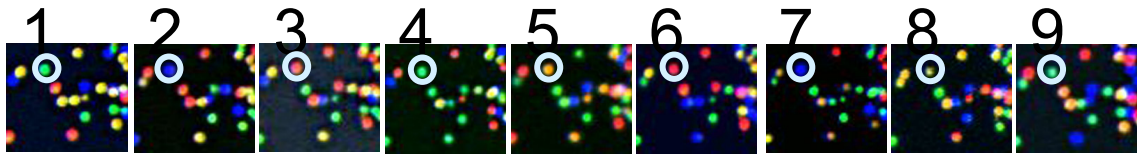
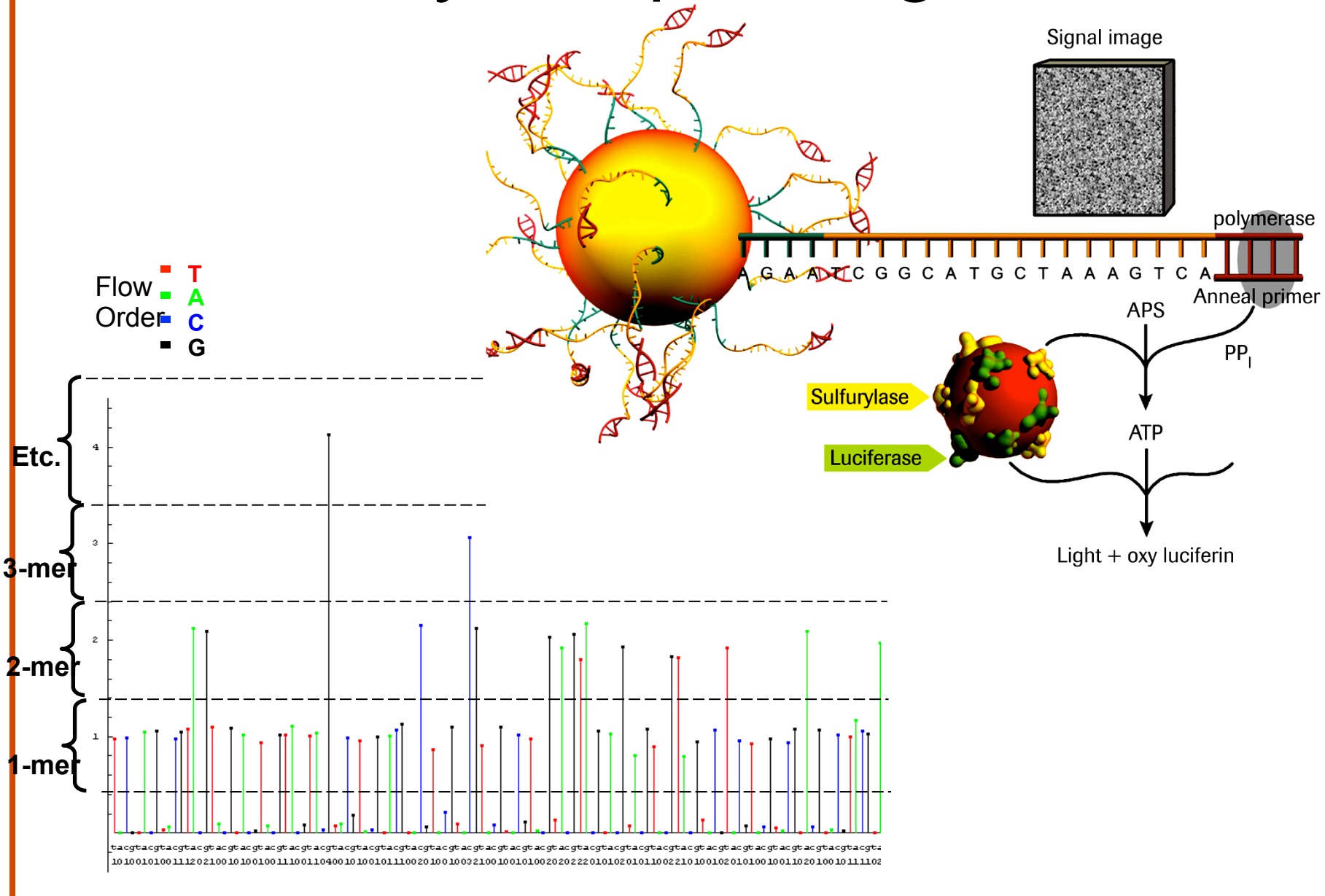


Image acquisition

Base calling

Leading tech. BUT…

Pyrosequencing



Sequencer comparison

	METHOD	Strengths	Weakness
Illumina	SBS	High throughput Low cost/Gb Protocol support Software support	Short reads Low complexity issue Some GC/AT rich errors
454	Pyrosequencing	Long & accurate reads Software support	Low throughput Very high cost Homopolymer errors
PacBio	Single molecule	Very long reads Random errors DNA modification calls Software support	Low accuracy Low throughput High cost
Ion Proton	Pyrosequencing	Cheaper machine High throughput	emPCR needed Homopolymer errors Software support
MinIon	Nanopore	Small & portable Long reads	Low accuracy Low throughput Cost??

Why sequence?

1. ***de novo***

   Genomic sequence empowers genetic analysis.

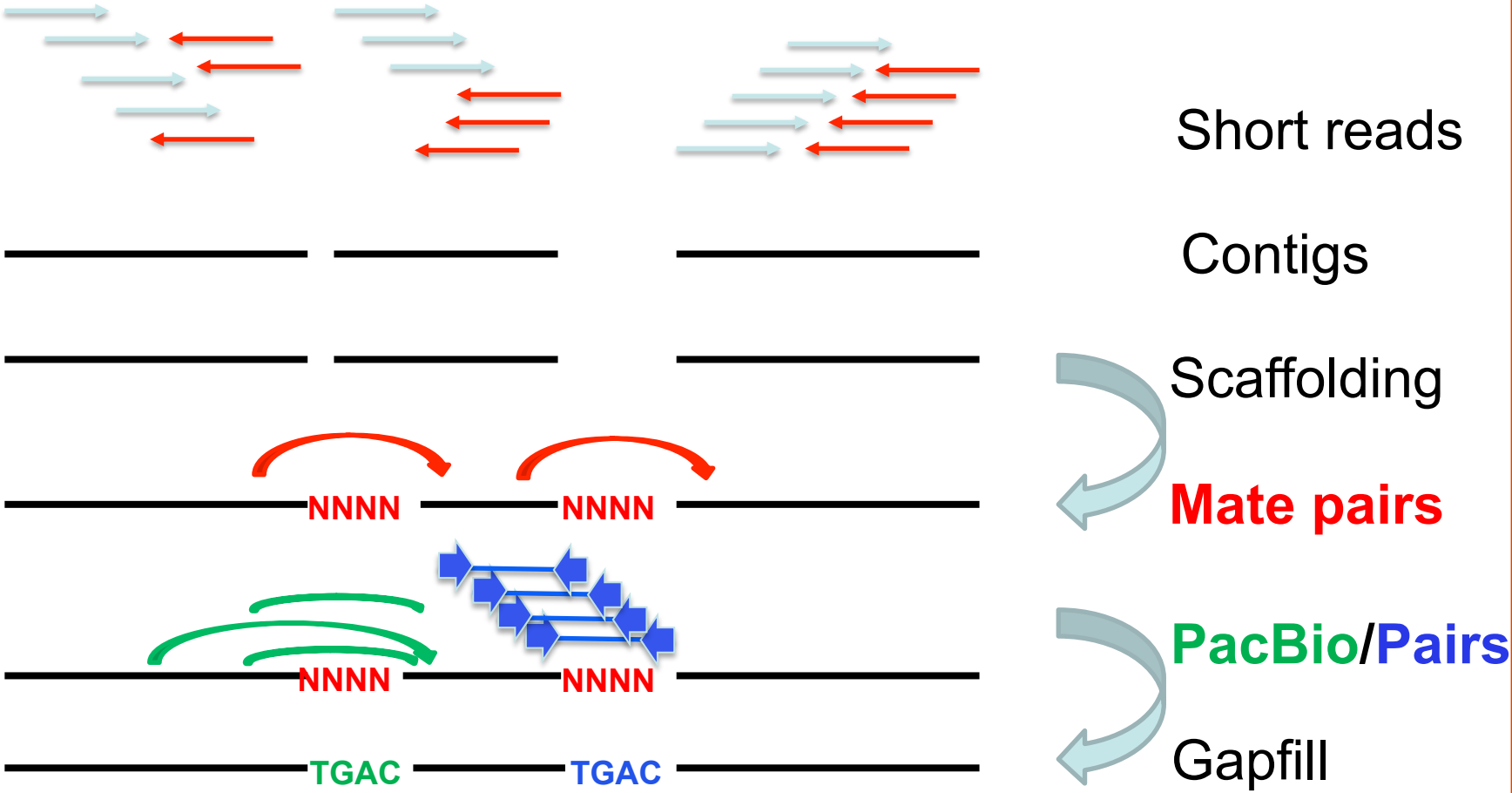
2. **Resequencing**

   Understand the nature of genetic variation.

3. **Counting**

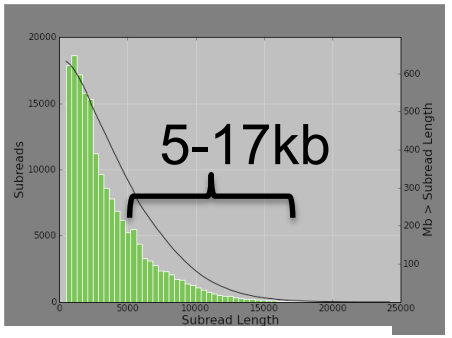
   Functional genomics e.g. gene expression etc.

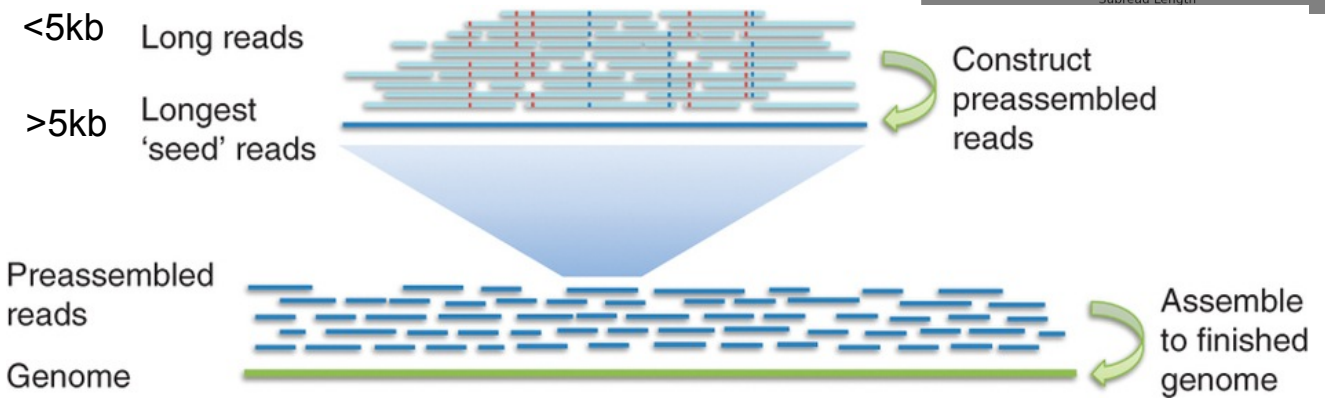
de novo genome assembly

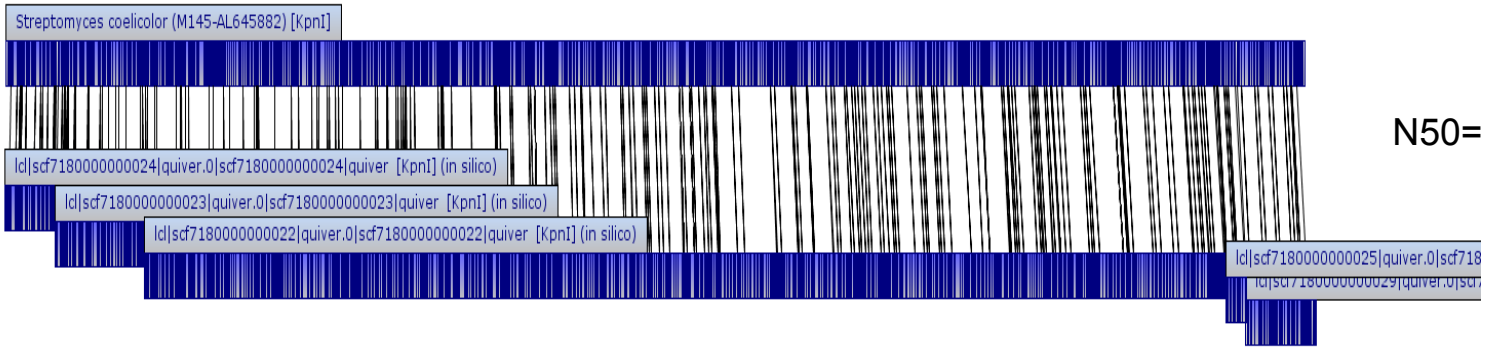


de novo genome assembly – long reads

PacBio
500Mb/cell

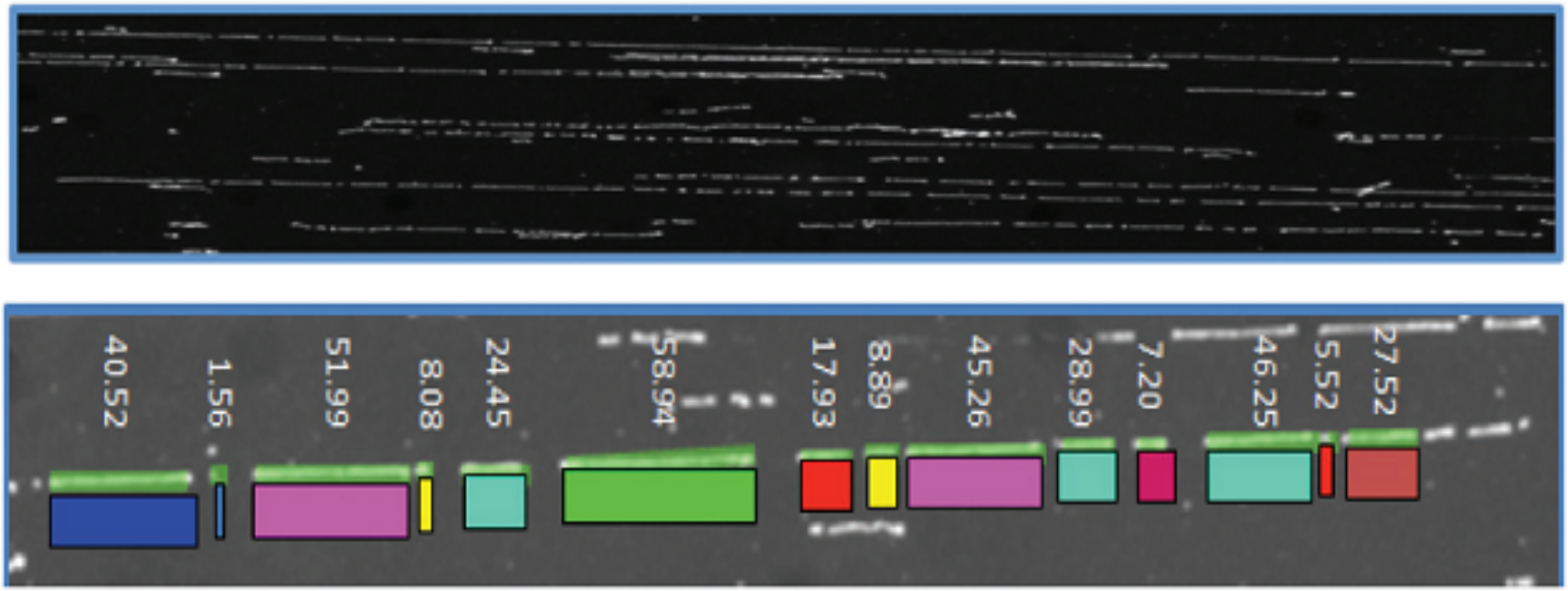




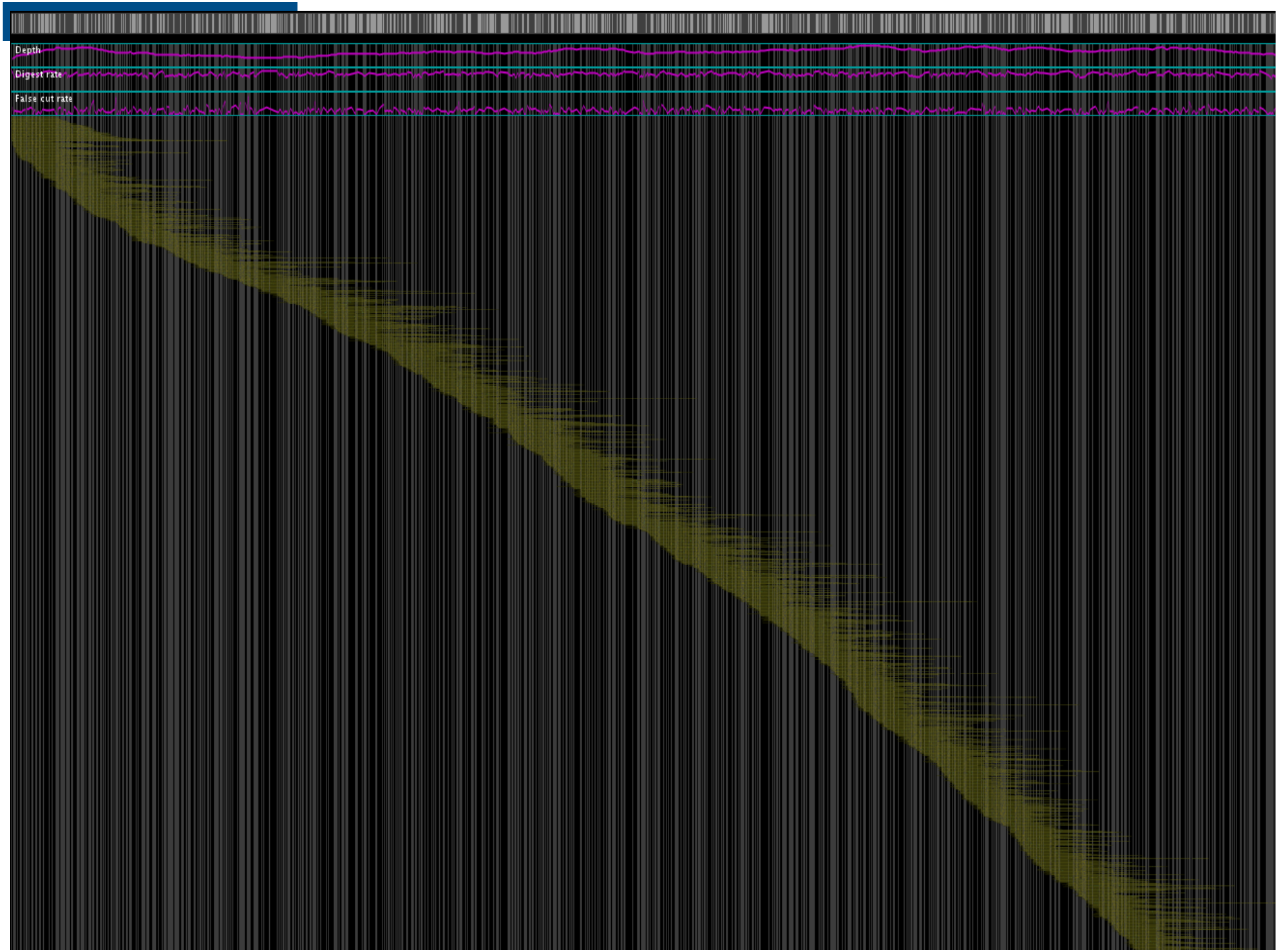


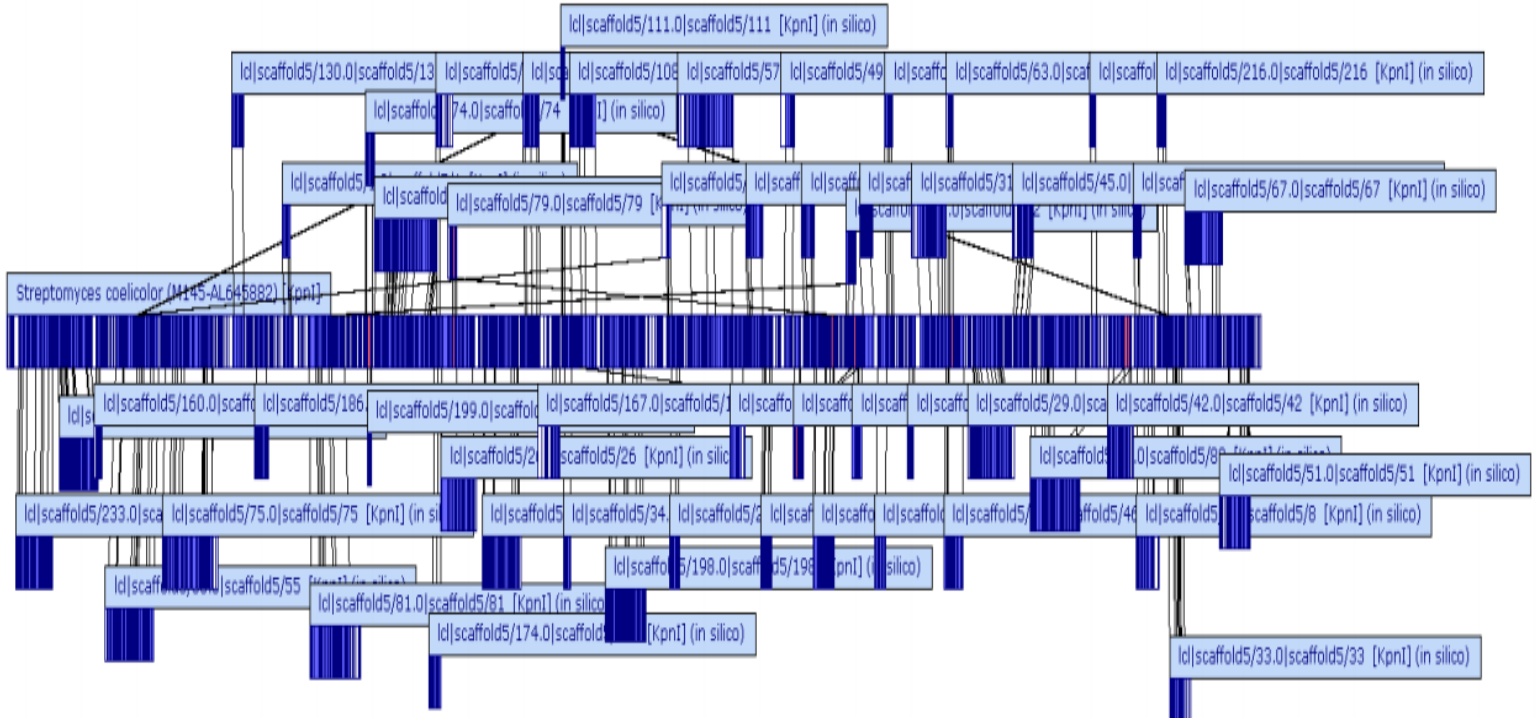
N50= 7.1Mbp

Optical mapping

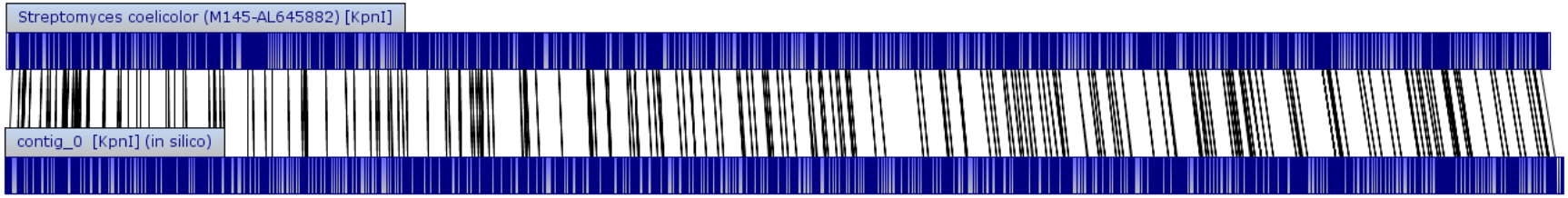


Long range information 150kb-2Mb


Depth
Digest rate
False cut rate



Comparison of assembly to OpGen map (N50= 177,135bp)

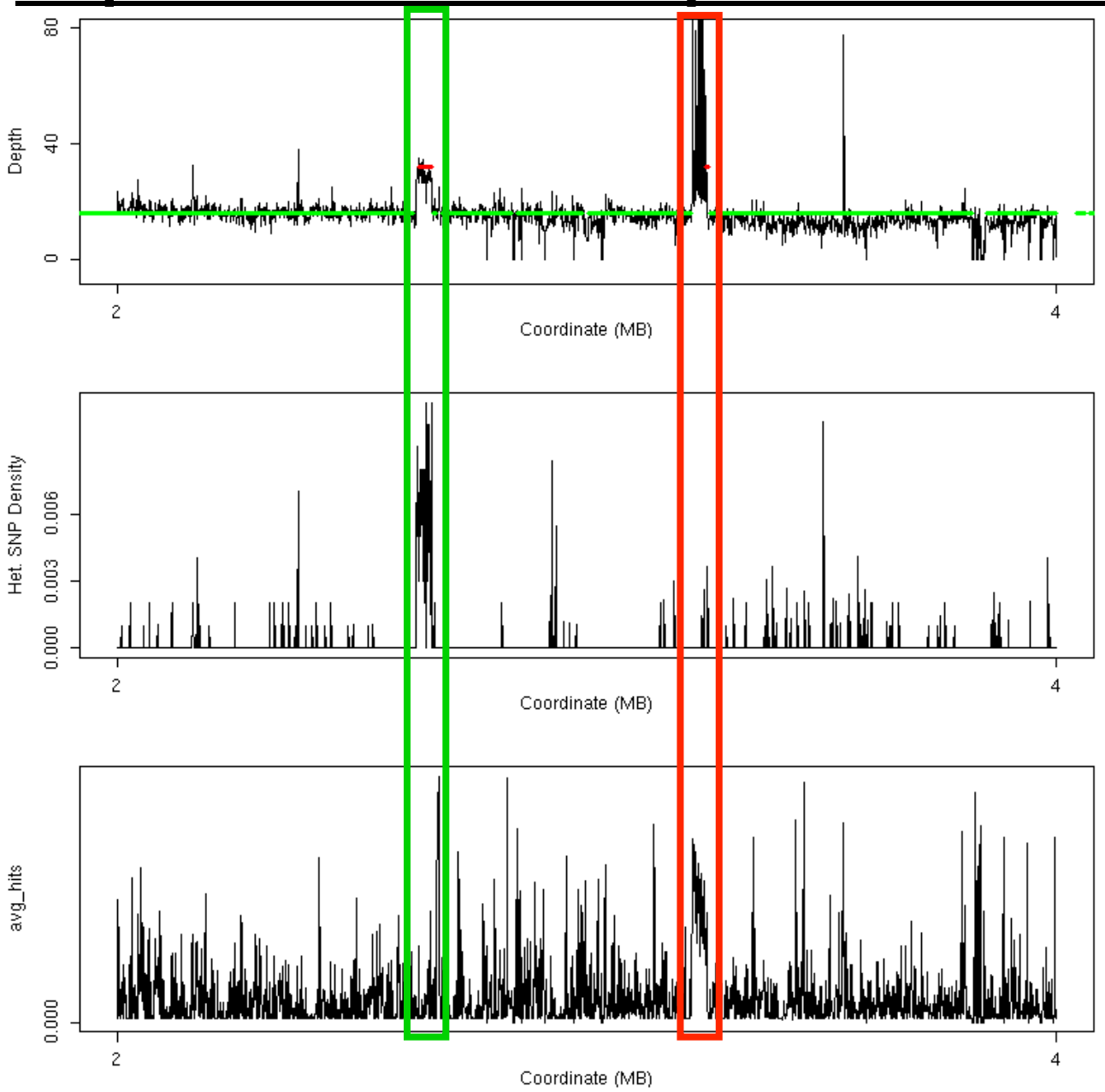


Comparison of assembly to OpGen map (N50= 8,632,393bp)

Resequencing variation Types

- Cytological level:
  - Chromosome numbers
  - Segmental duplications, rearrangements, and deletions
- Molecular level:
  - Transposable Elements
  - Short Deletions/Insertions, Tandem Repeats
- Sequence level (WGS/Exomes):
  - Single Nucleotide Polymorphisms (SNPs)
  - Small Nucleotide Insertions and Deletions (Indels)

Duplications/CNV - depth and heterozygosity

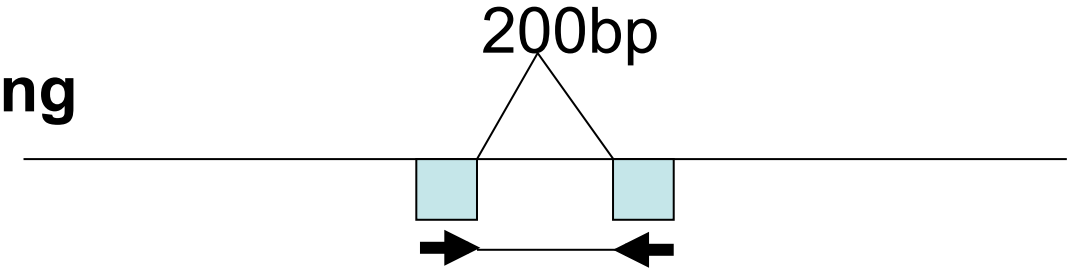


Jared Simpson
(pers. comm)

CNV-seq,
CNVator etc.

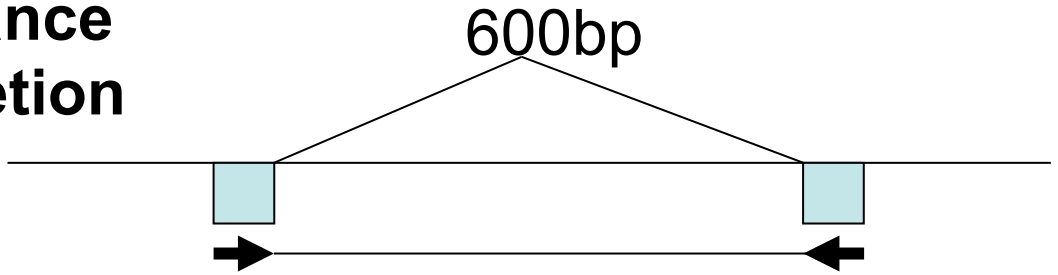
Molecular level variation (10s-1000s bp)

Normal mapping

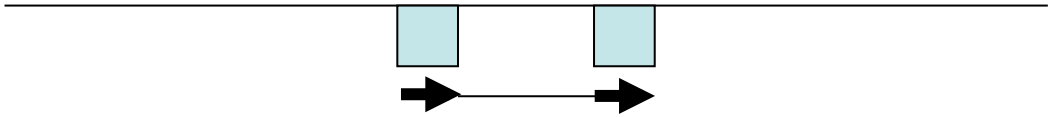


Reference sequence

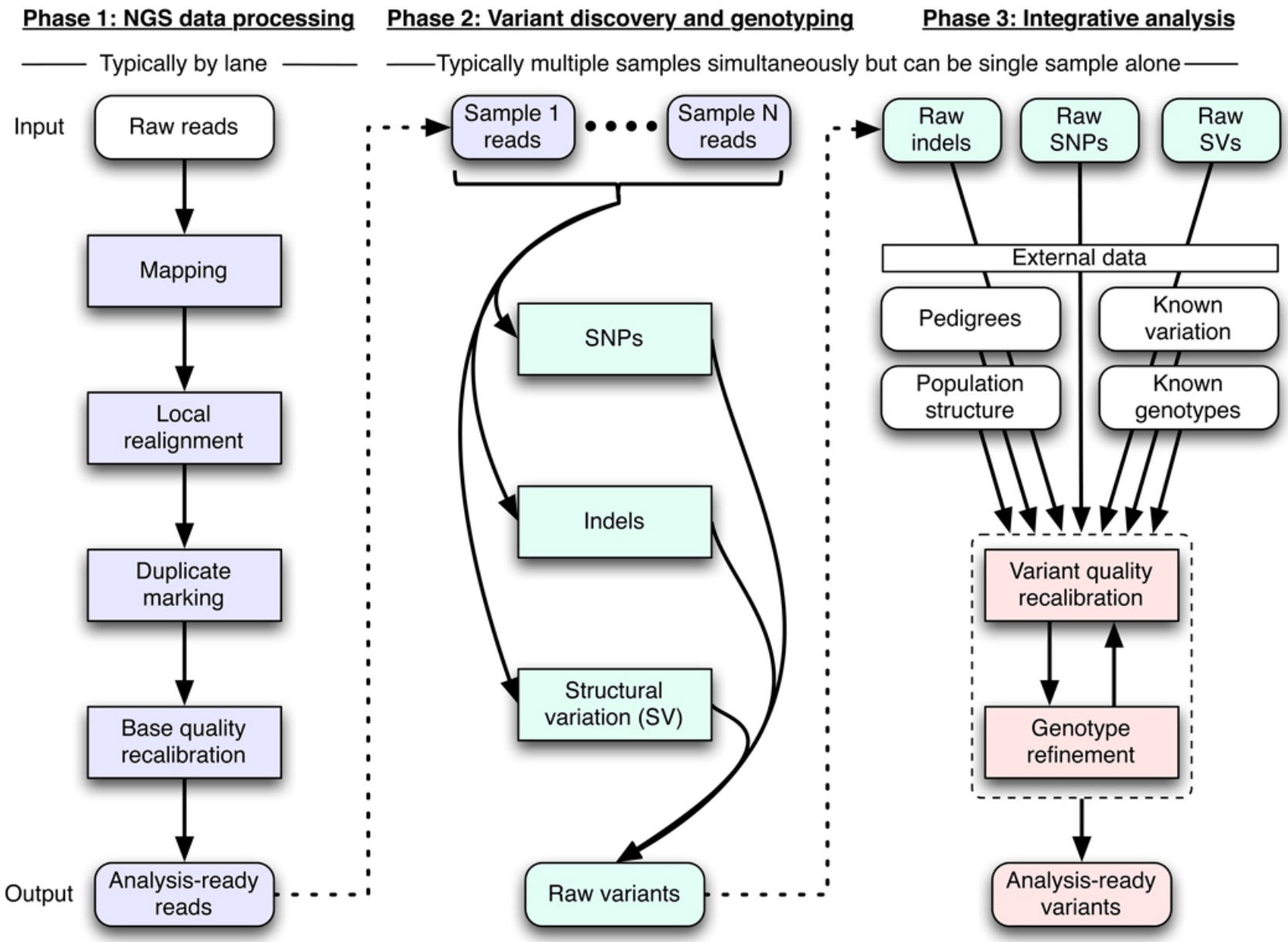
Abnormal distance
= insertion/deletion



Abnormal direction
= inversion

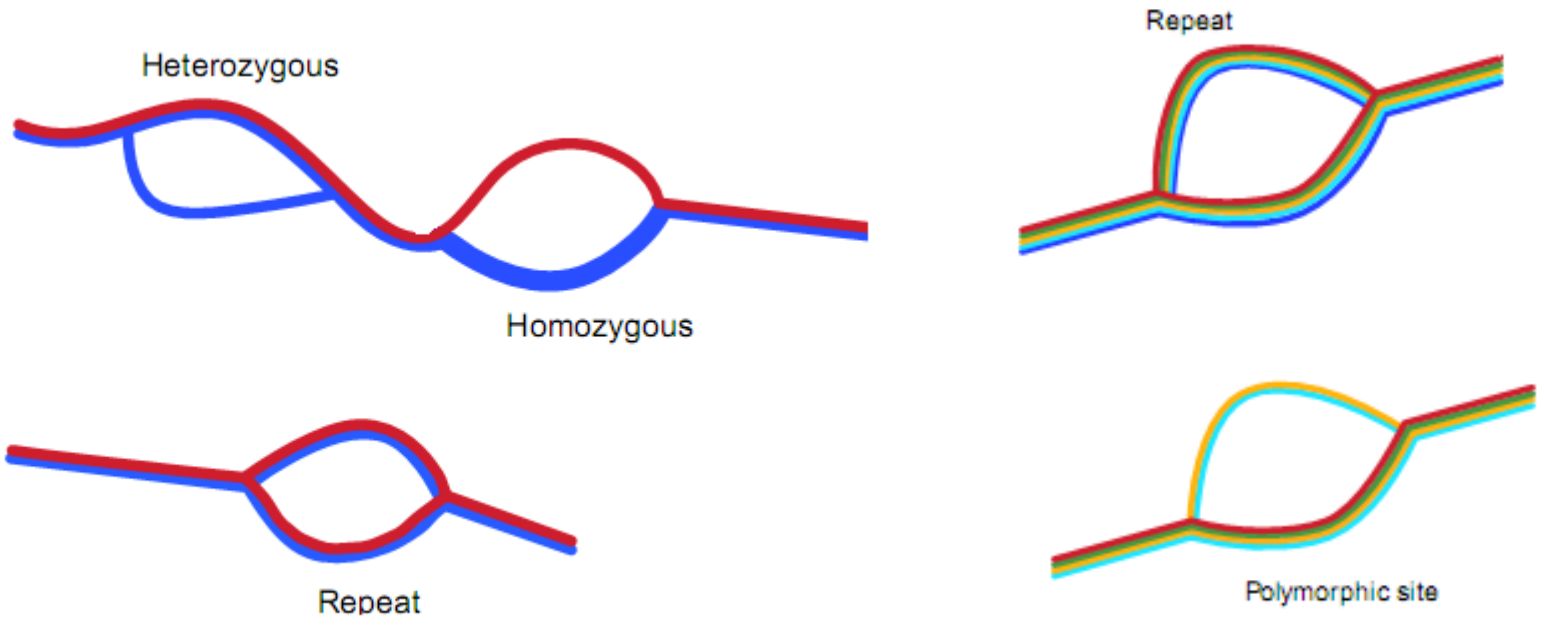


Pindel, Breakdancer, GATK - confirmation by local assembly



DePristo et al. Nature Genetics 2011 (GATK)

De novo assembly *for* genotyping



Cortex: Iqbal, Caccamo et al. Nat. Gen. 2012
DISCOVAR: Weisenfeld et al. Nat. Gen. 2014

Alignment based tools have reference bias
Cortex & DISCOVAR better at calling indels & SVs

Optical mapping for SV, CNV and maybe haplotypes?

Whole Genome Sequencing

+ Possible to call all variants
- High cost 1 sample per lane @25x

Few samples = lower statistical power

Whole Exome Sequencing

+ All coding, and some flanking genic variants
- Medium cost 8-24 samples per lane @ 50x

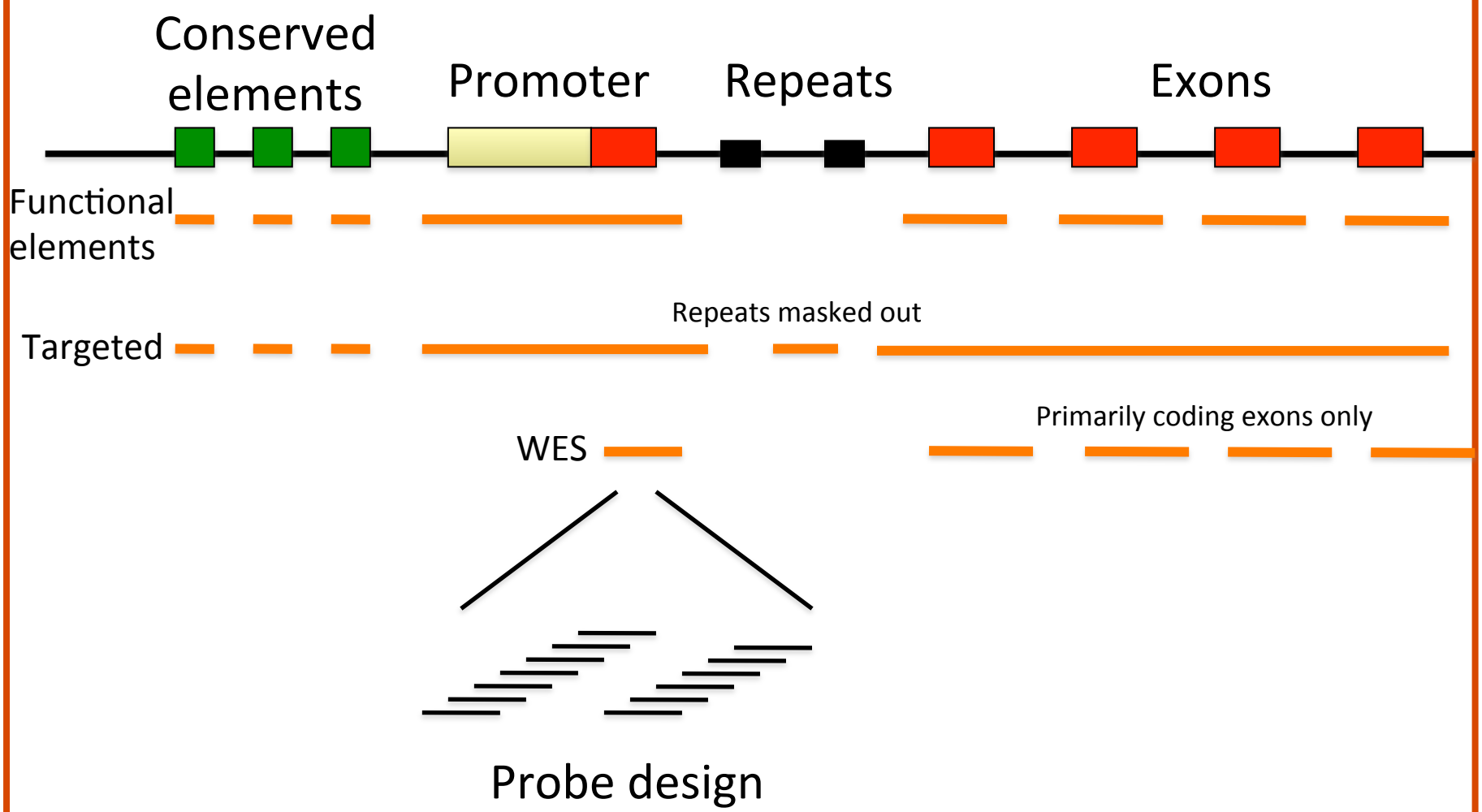
More samples per $XXX = more statistical power

Targeted Sequencing

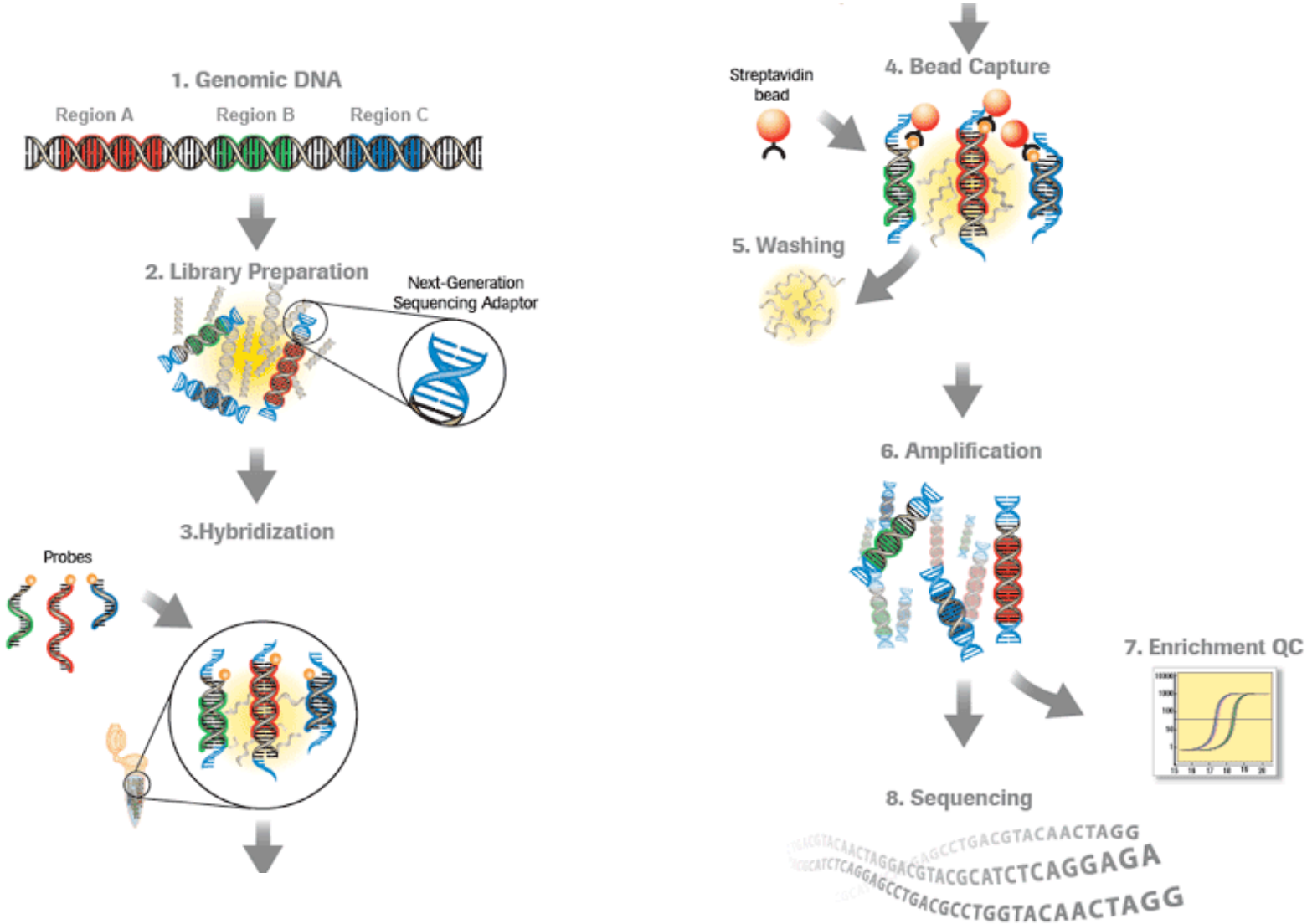
+ Variants in small regions
- Low cost 96+ samples per lane

Most samples per $XXX = highest statistical power

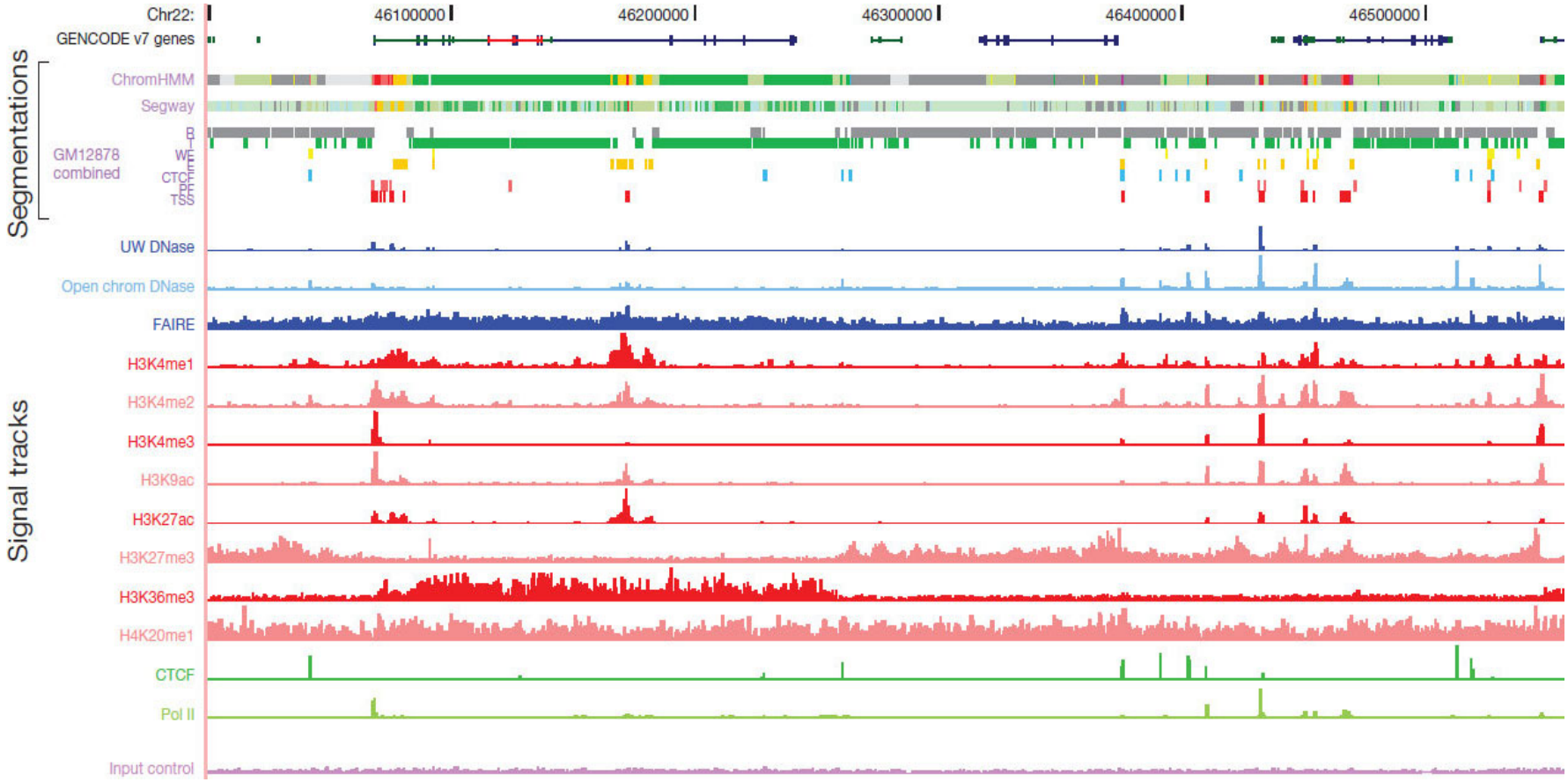
Targeted re-sequencing vs Whole exome sequencing (WES)



DNA Enrichment (Nimblegen)



Counting tags



ENCODE data tracks

Future developments

Longer reads

- Better *de novo* assemblies
- Haplotypes calling
- Structural variants

Epigenetics

- Single molecule sequencing direct detection

Low input sequencing

- Few or single cells

Cloud based computing

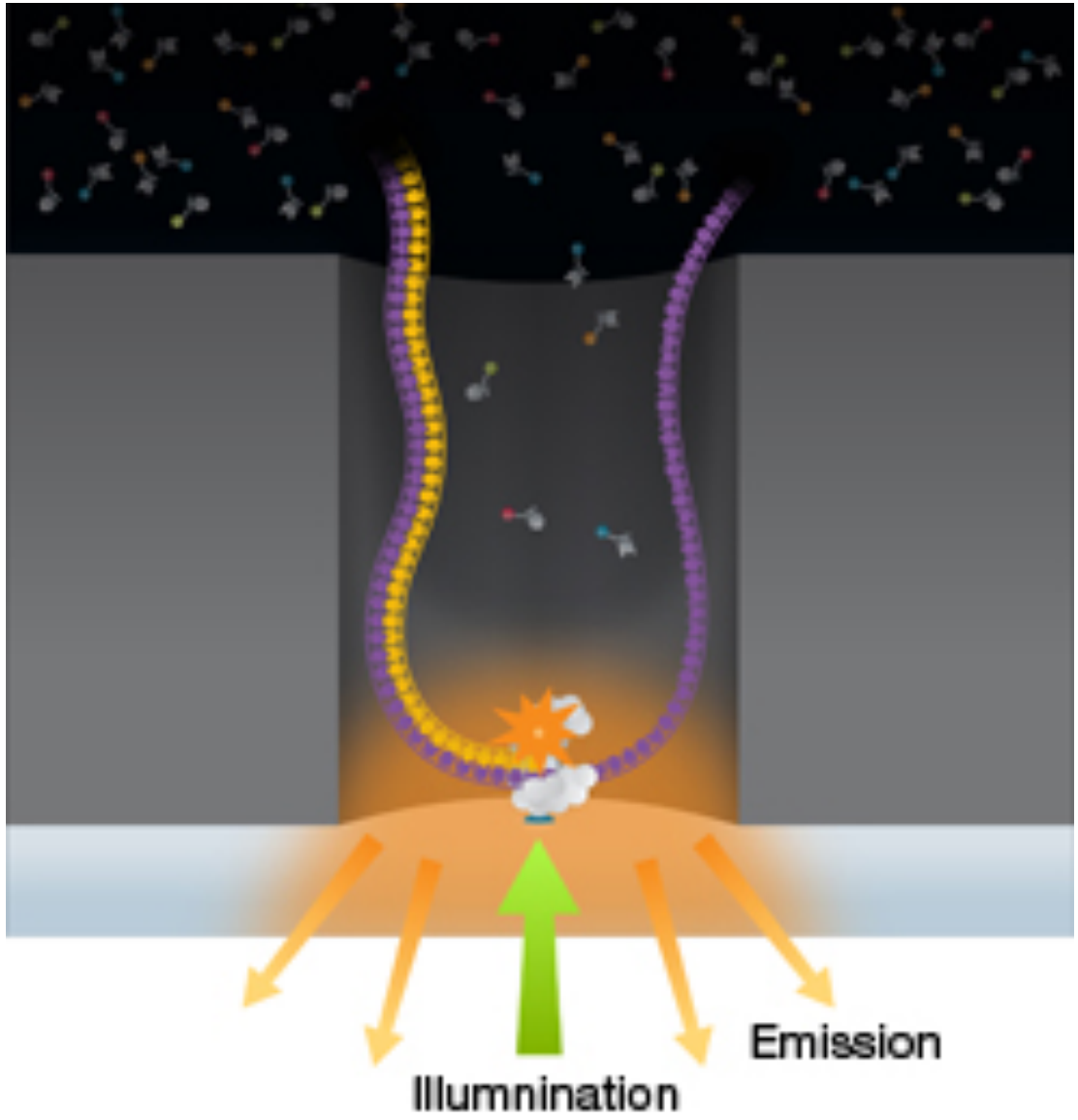
- Corporation: Amazon, Google cloud etc.
- Public: Galaxy, iPlant etc.

3rd generation sequencers

	2nd Generation	3rd Generation
Template	Amplified	Single molecule
Read length	Short	Long
Accuracy	~99%	70-85%
Speed	1 day> 2 week	Minutes
Epigenetics	Indirect	Direct
Cost/Gb	Low	High but falling

Pacific Biosciences RS

Zero Mode Waveguide



500+Mb 240min run

PacBio Base Modification sequencing

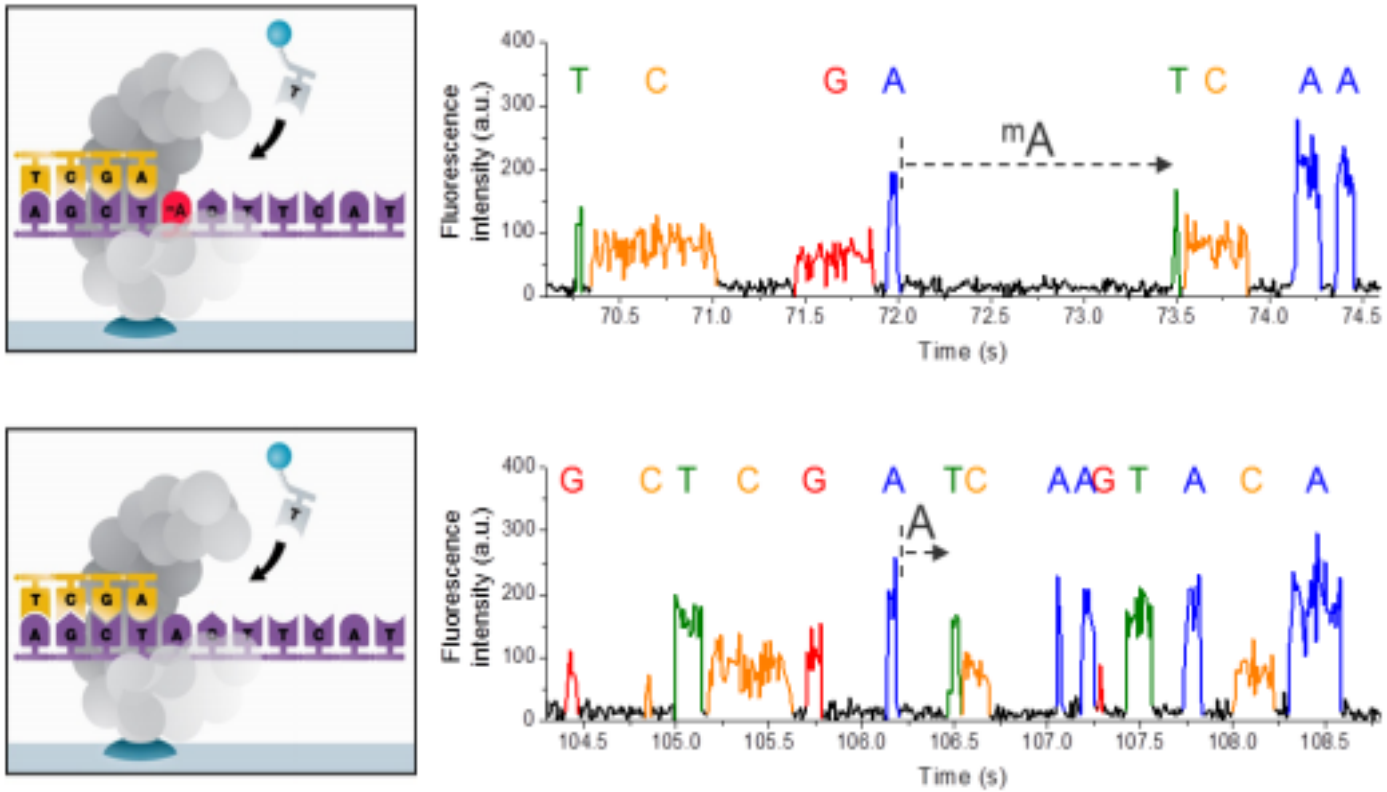


Figure 2. Principle of detecting modified DNA bases during SMRT sequencing. The presence of the modified base in the DNA template (top), shown here for 6-methyladenine, results in a delayed incorporation of the corresponding T nucleotide, i.e. longer interpulse duration (IPD), compared to a control DNA template lacking the modification (bottom).[3]

DNA modifications

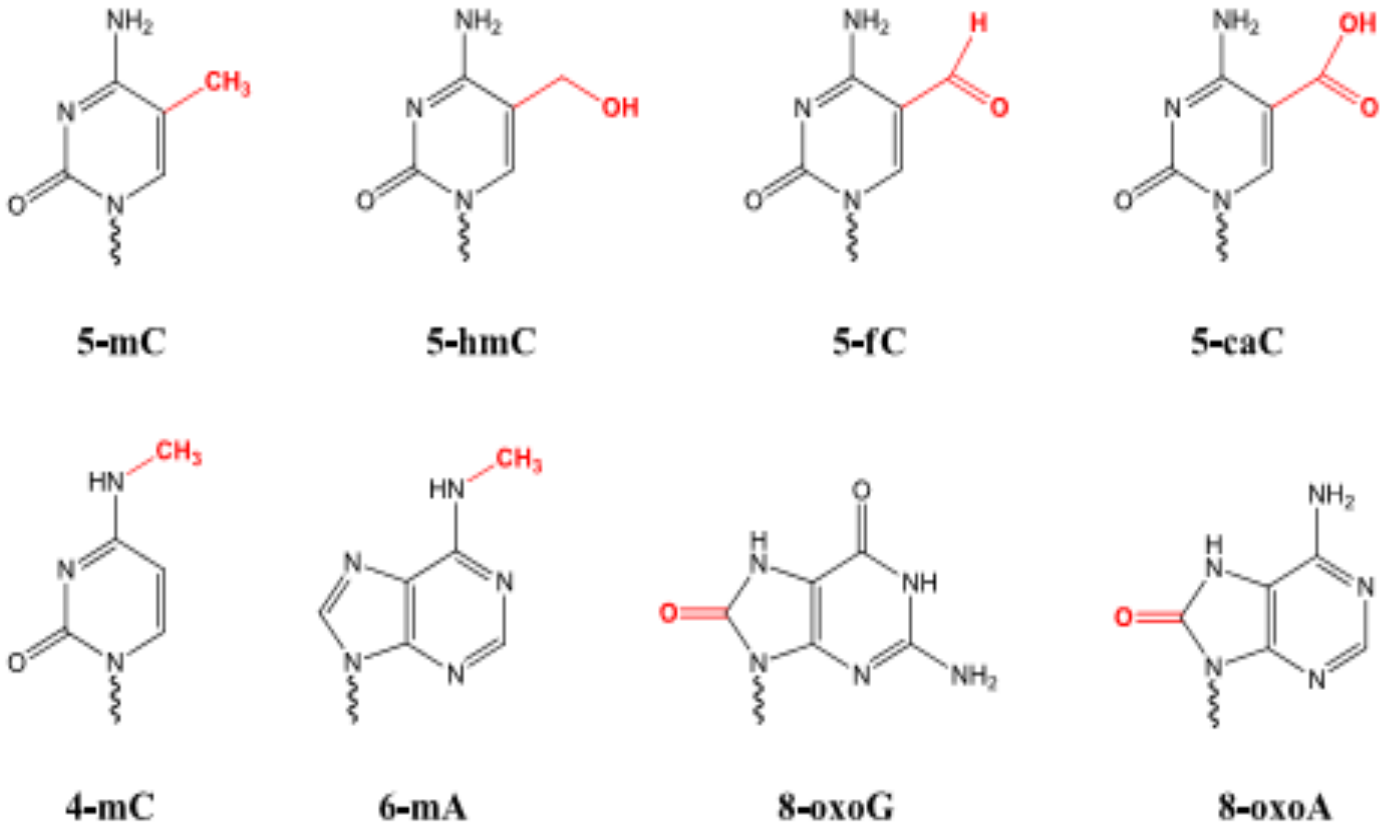
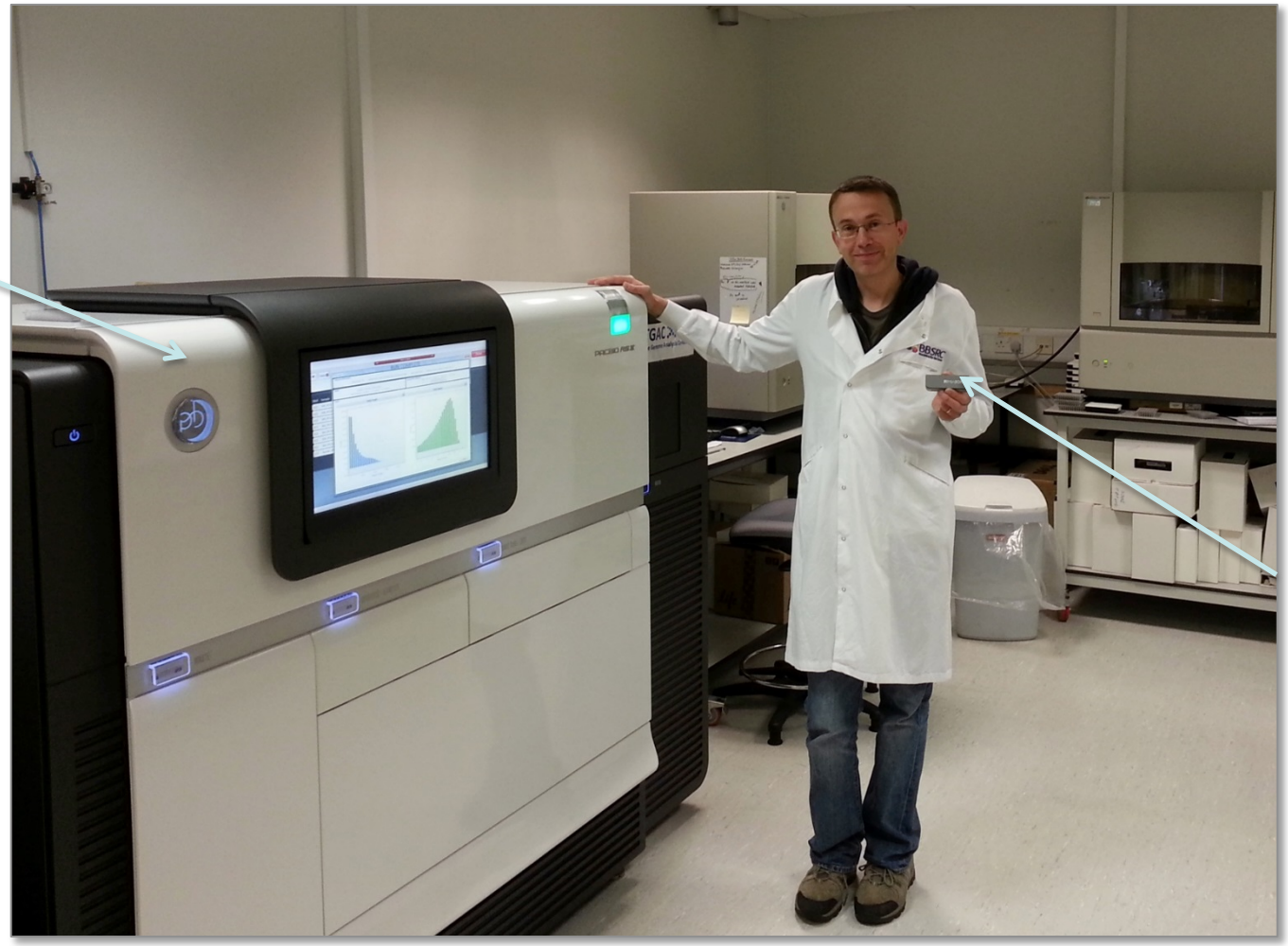


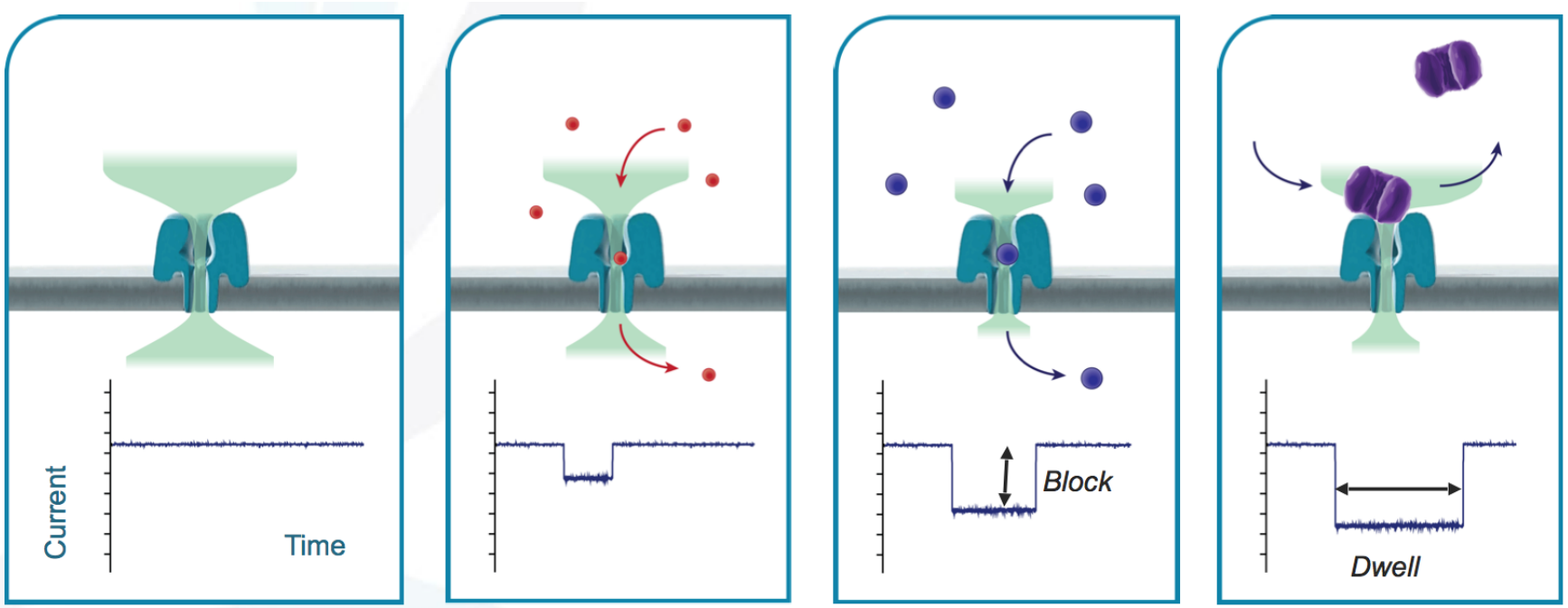
Figure 1. Molecular structures of base modifications including 5-methylcytosine (5-mC), 5-hydroxymethylcytosine (5-hmC), 5-formylcytosine (5-fC), 5-carboxylcytosine (5-caC), 4-methylcytosine (4-mC), 6-methyladenine (6-mA), 8-oxoguanine (8-oxoG), and 8-oxoadenine (8-oxoA).

Pacific Biosciences v Oxford nanopore



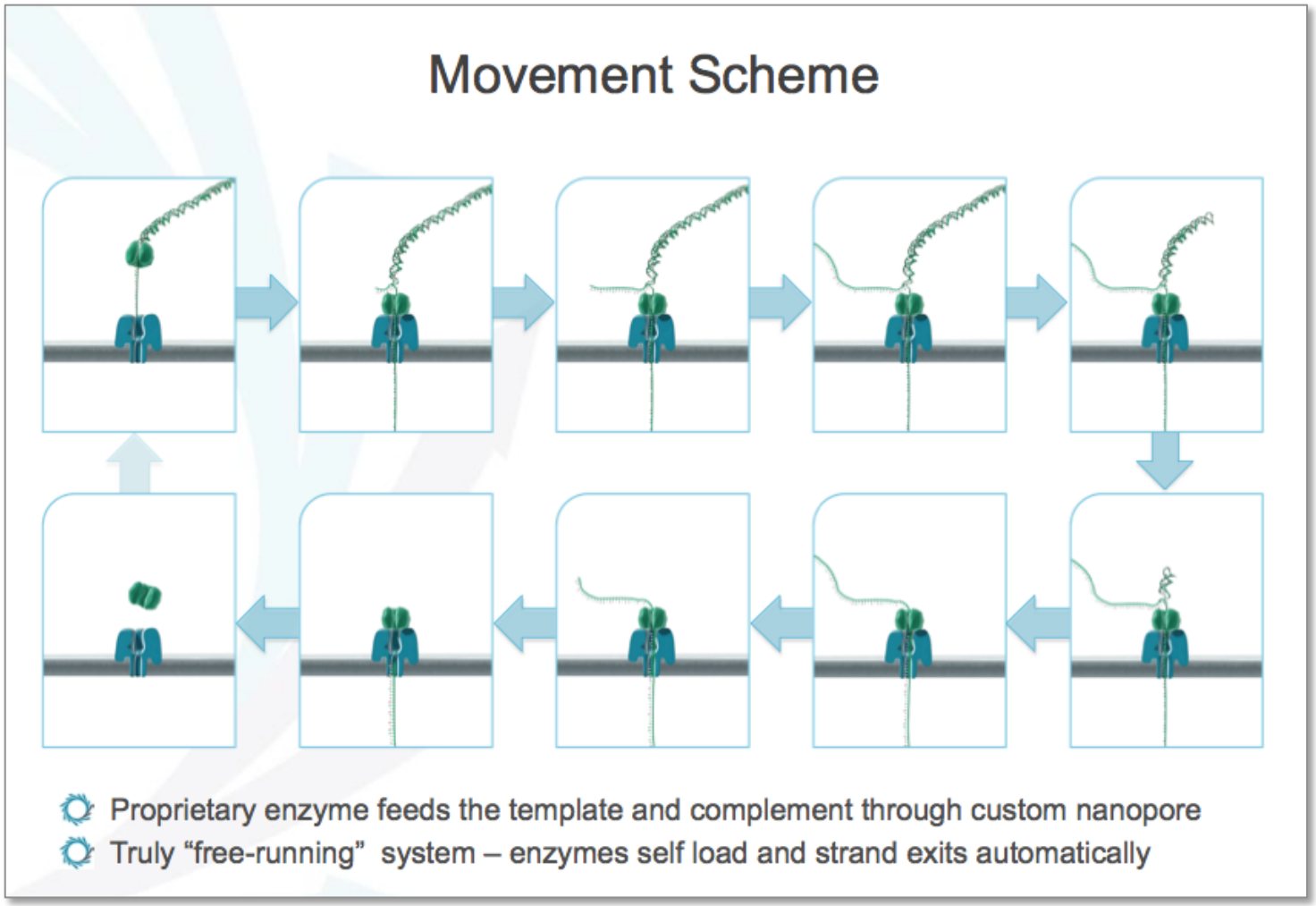
Oxford Nanopore Technologies

- Nanopore sensing based on disruption of ion flow through nanopore.



Oxford Nanopore Technologies

- Template, Complement and 2D reads.

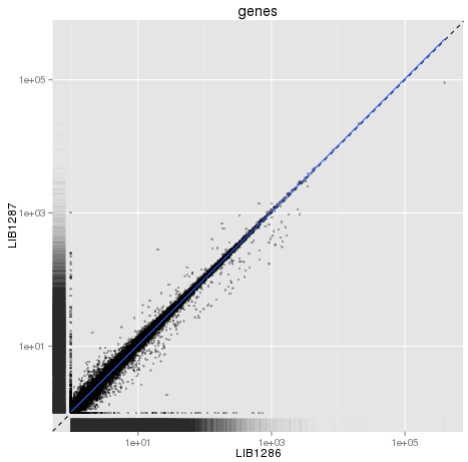


Oxford Nanopore Technologies

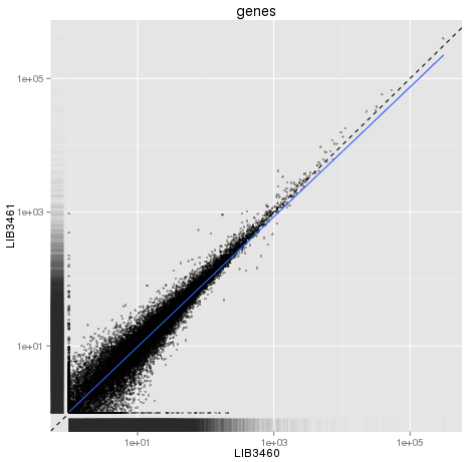
Reference

MinION read (mismatchs)

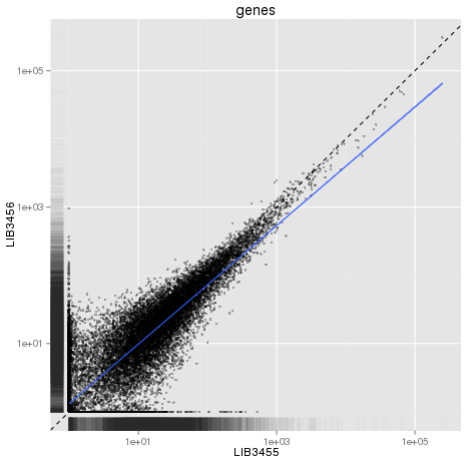
Low input sequencing (RNA-seq)



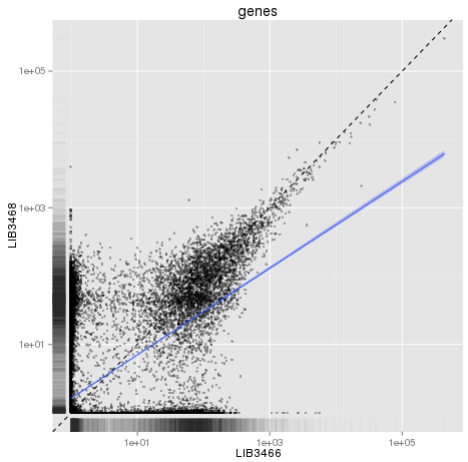
1µg



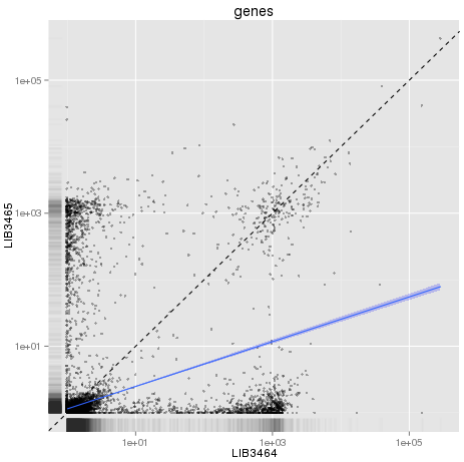
10ng (1,000 cells)



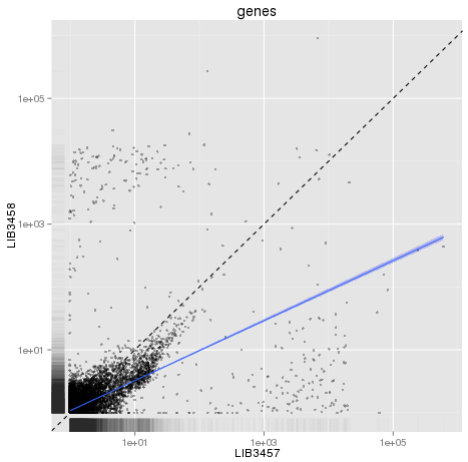
1ng (100 cells)



100pg (10 cells)

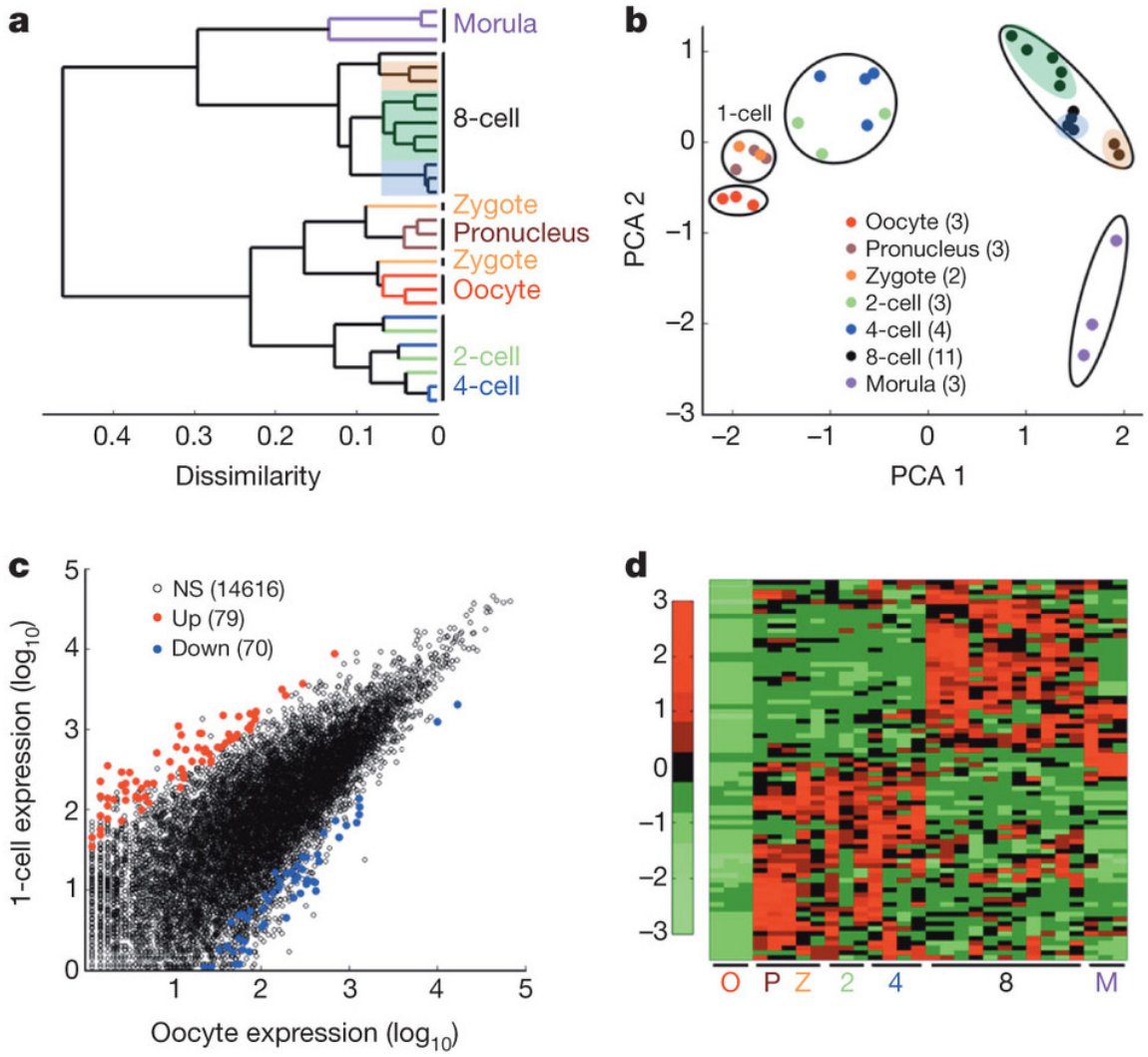


10pg (1 cell)



1pg (0.1 cell)

Low input sequencing (RNA-seq)

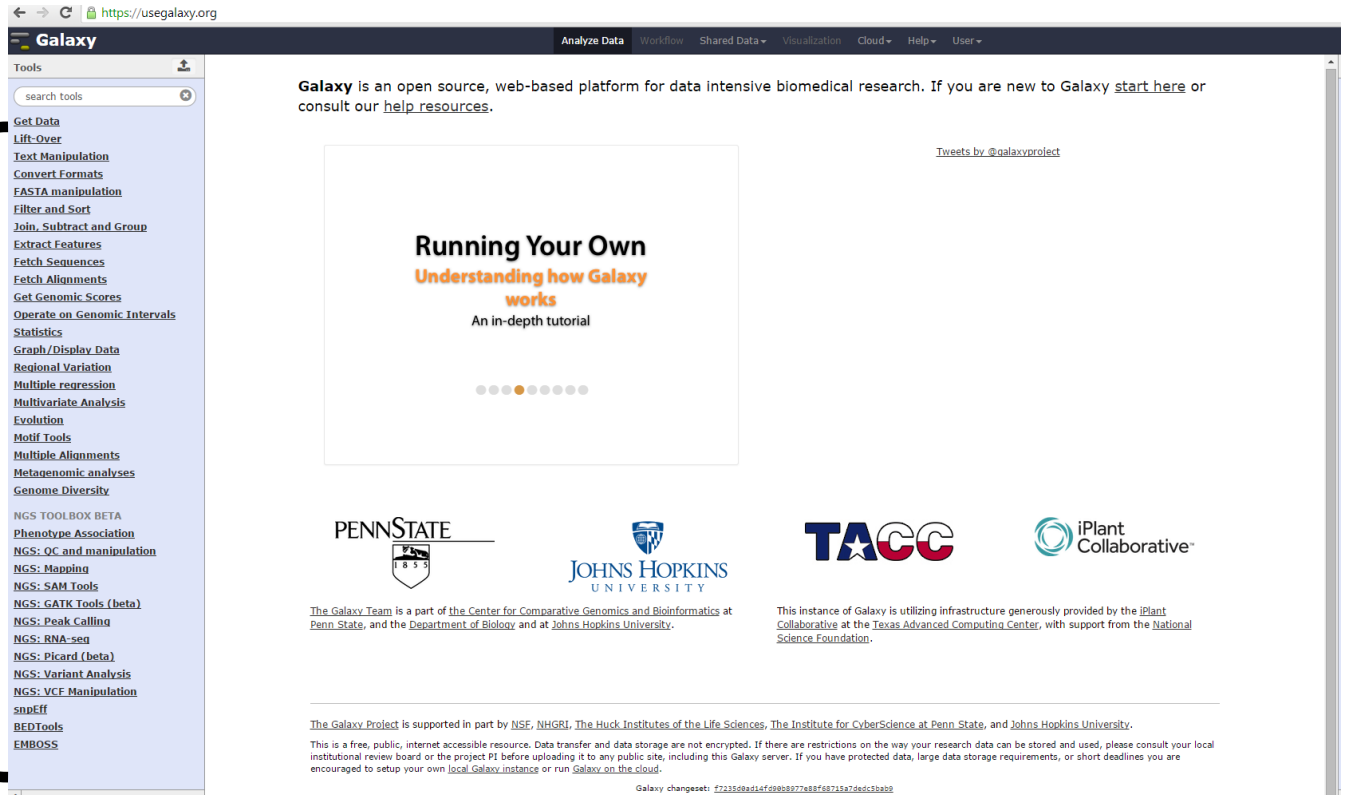


Zhigang Xue et al. Nature 2013

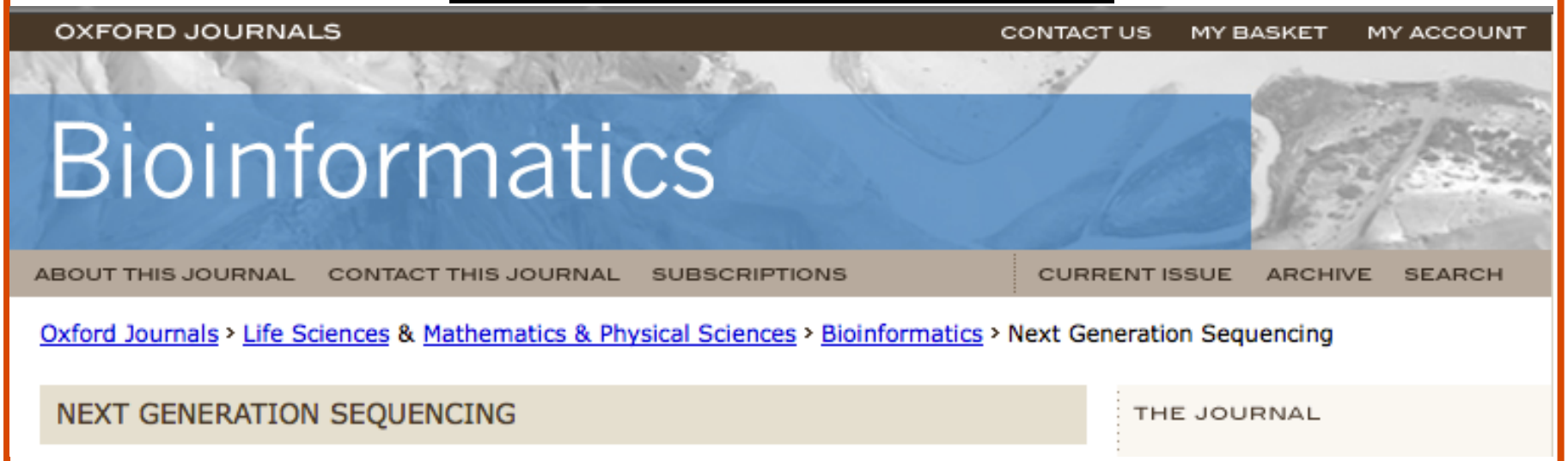
Cloud computing - Galaxy

- Galaxy is an open, web-based platform for data intensive biomedical research https://usegalaxy.org/

Premade Bioinformatics workflows



Additional resources



- http://www.oxfordjournals.org/our_journals/bioinformatics/nextgenerationsequencing.html

- Training: http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course
- Forums: https://www.biostars.org/ http://seqanswers.com/