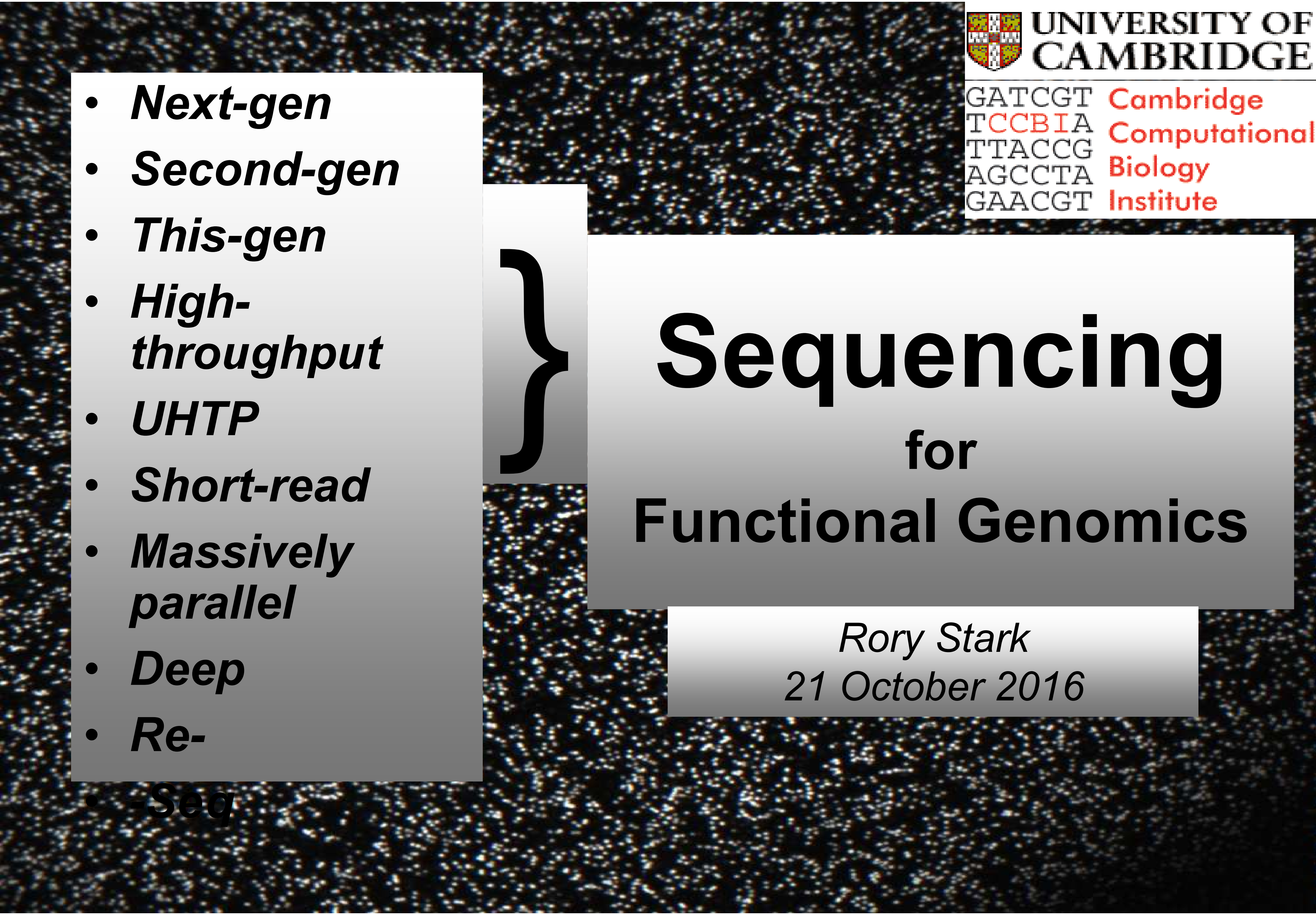


UNIVERSITY OF CAMBRIDGE

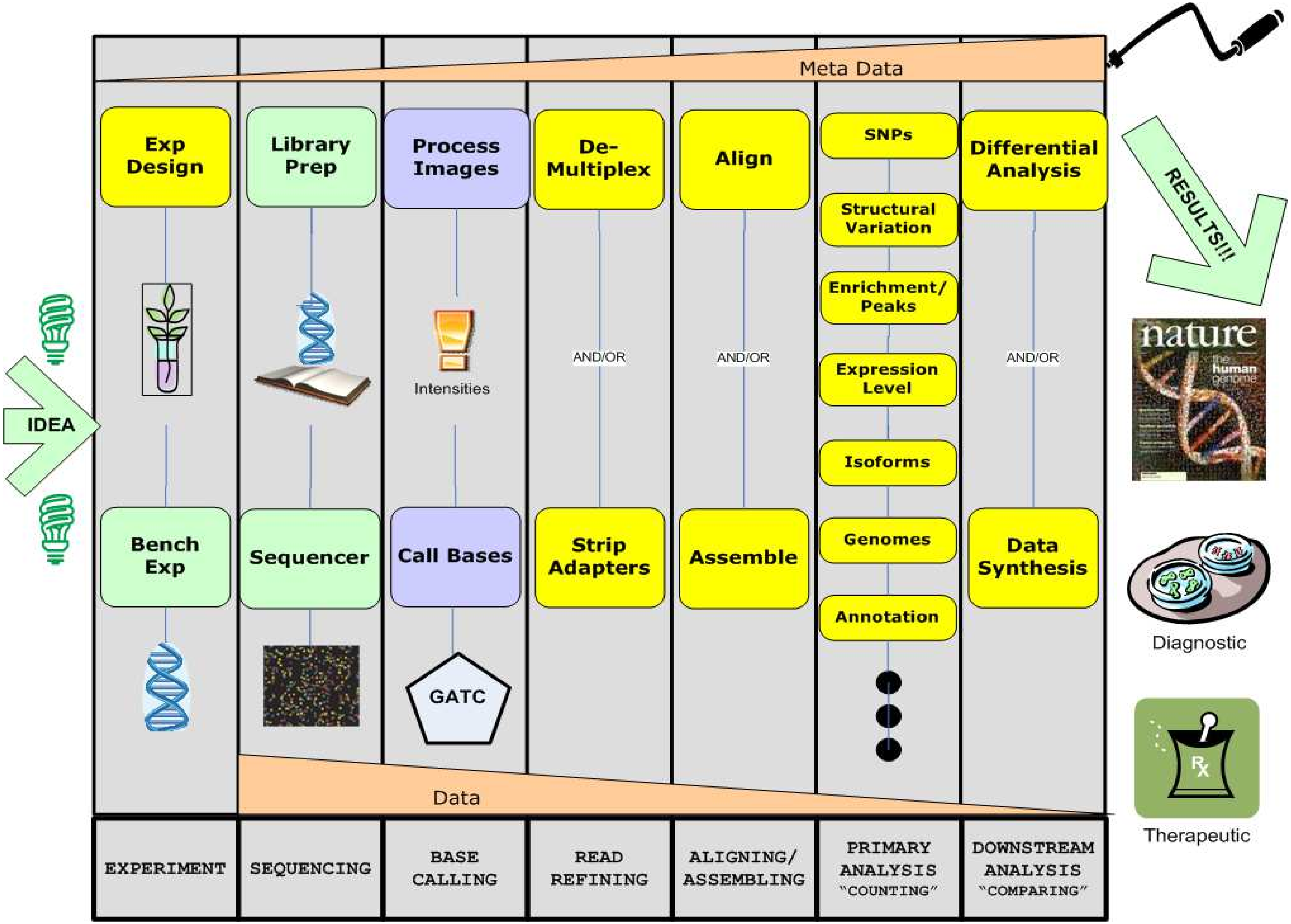
Cambridge Computational Biology Institute

- ***Next-gen***
- ***Second-gen***
- ***This-gen***
- ***High-throughput***
- ***UHTP***
- ***Short-read***
- ***Massively parallel***
- ***Deep***
- ***Re-***
- ***-Seq***

Sequencing for Functional Genomics

Rory Stark

21 October 2016



Meta Data

IDEA

Exp Design

Bench Exp

EXPERIMENT

Library Prep

Sequencer

SEQUENCING

Process Images

Intensities

Call Bases

GATC

BASE CALLING

De-Multiplex

AND/OR

Strip Adapters

READ REFINING

Align

AND/OR

Assemble

ALIGNING/ ASSEMBLING

SNPs

Structural Variation

Enrichment/ Peaks

Expression Level

Isoforms

Genomes

Annotation

PRIMARY ANALYSIS "COUNTING"

Differential Analysis

AND/OR

Data Synthesis

DOWNSTREAM ANALYSIS "COMPARING"

Data

RESULTS!!!

Diagnostic

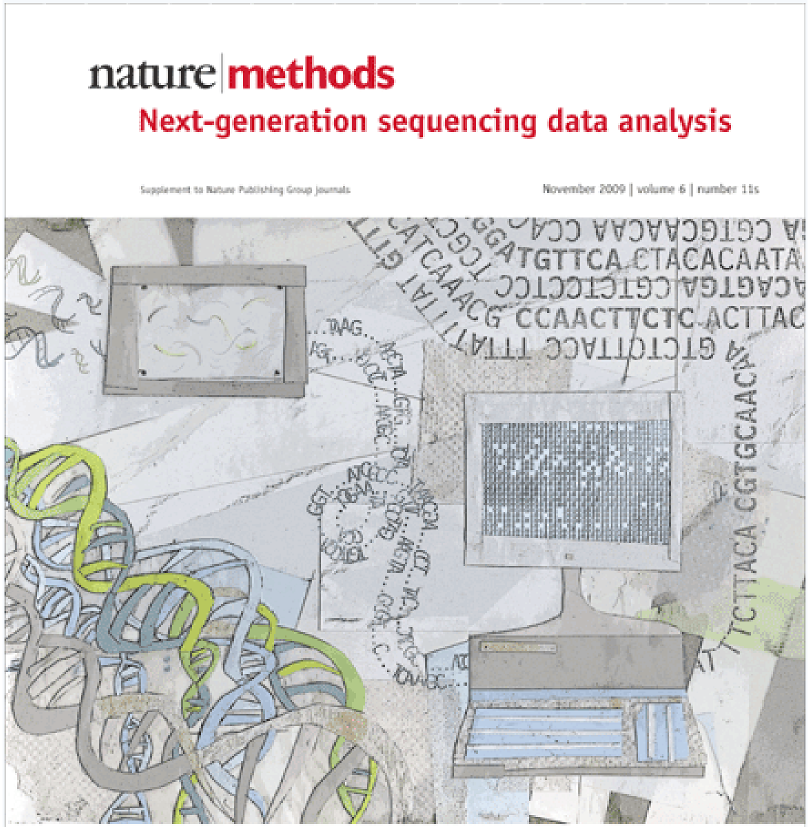
Therapeutic

Plan of Lectures

- Lecture 5: Intro to Sequencing
- Lecture 6: RNA-Seq I: Mapping Strategies (+ Practical)
- Lecture 7: RNA-Seq II: Counting and Estimation
- Lecture 8: RNA-Seq III: Normalisation and Differential Expression (+ Practical)
- Lecture 9: ChIP-seq I: Design, QC, and Peak Calling
- Lecture 10: ChIP-seq II: Differential Binding Analysis (+Practical)

Nature Methods supplement November 2009

http://www.nature.com/nmeth/journal/v6/n11s/index.html



nature methods

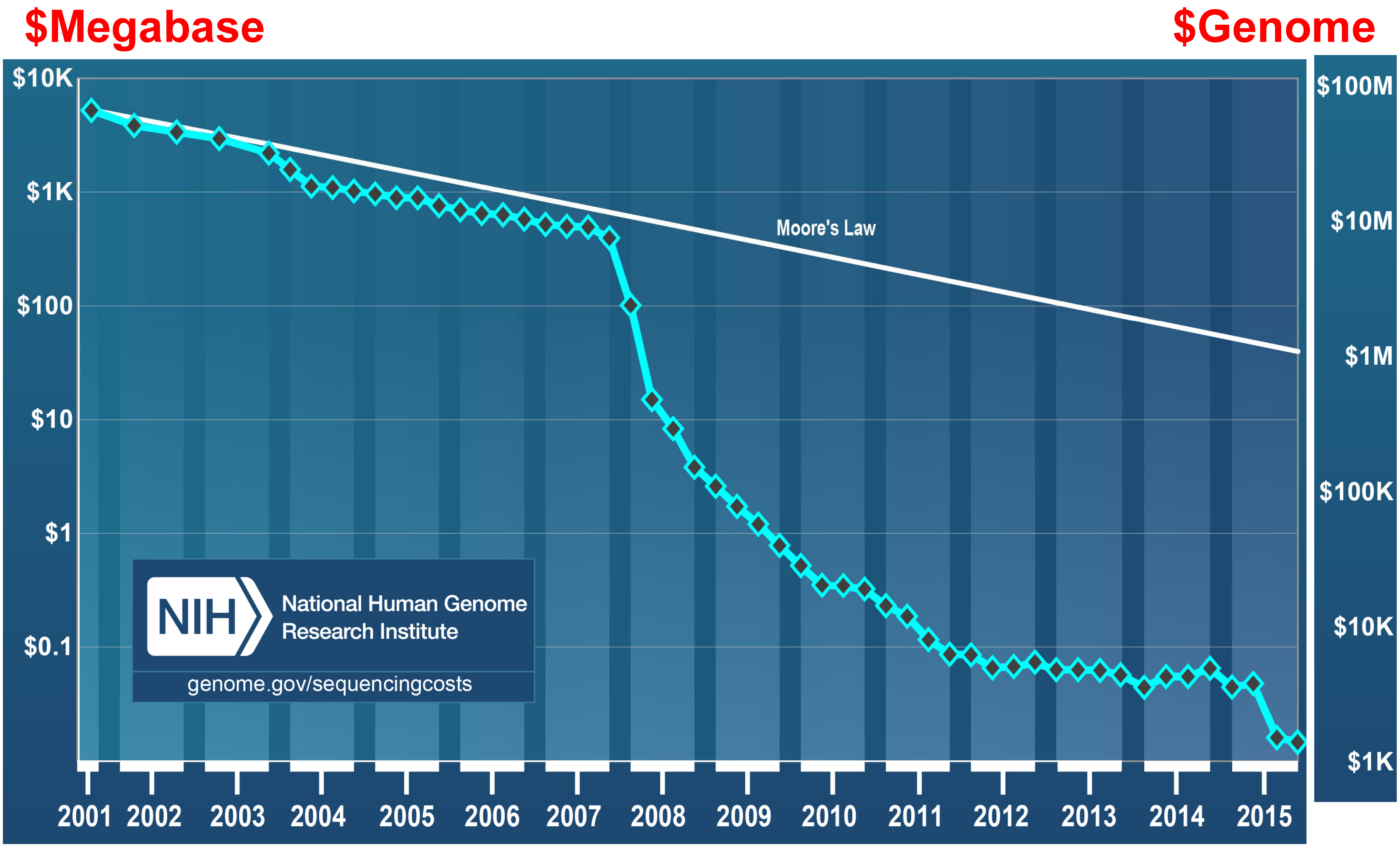
Next-generation sequencing data analysis

Supplement to Nature Publishing Group journals

November 2009 | volume 6 | number 11s

- Forward: Focus on next-generation sequencing data analysis
- Commentary: Next-generation gap
- Reviews:
  - Sense from sequence reads: methods for **alignment** and **assembly**
  - Computational methods for discovering **structural variation** with next-generation sequencing
  - Computation for **ChIP-seq** and **RNA-seq** studies

"There is a growing gap between the generation of massively parallel sequencing output and the ability to process and analyze the resulting data. New users are left to navigate a bewildering maze of base calling, alignment, assembly and analysis tools with often incomplete documentation and no idea how to compare and validate their outputs. Bridging this gap is essential, or the coveted $1,000 genome will come with a $20,000 analysis price tag."


$Megabase
$Genome
$10K
$1K
$100
$10
$1
$0.1
$100M
$10M
$1M
$100K
$10K
$1K
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

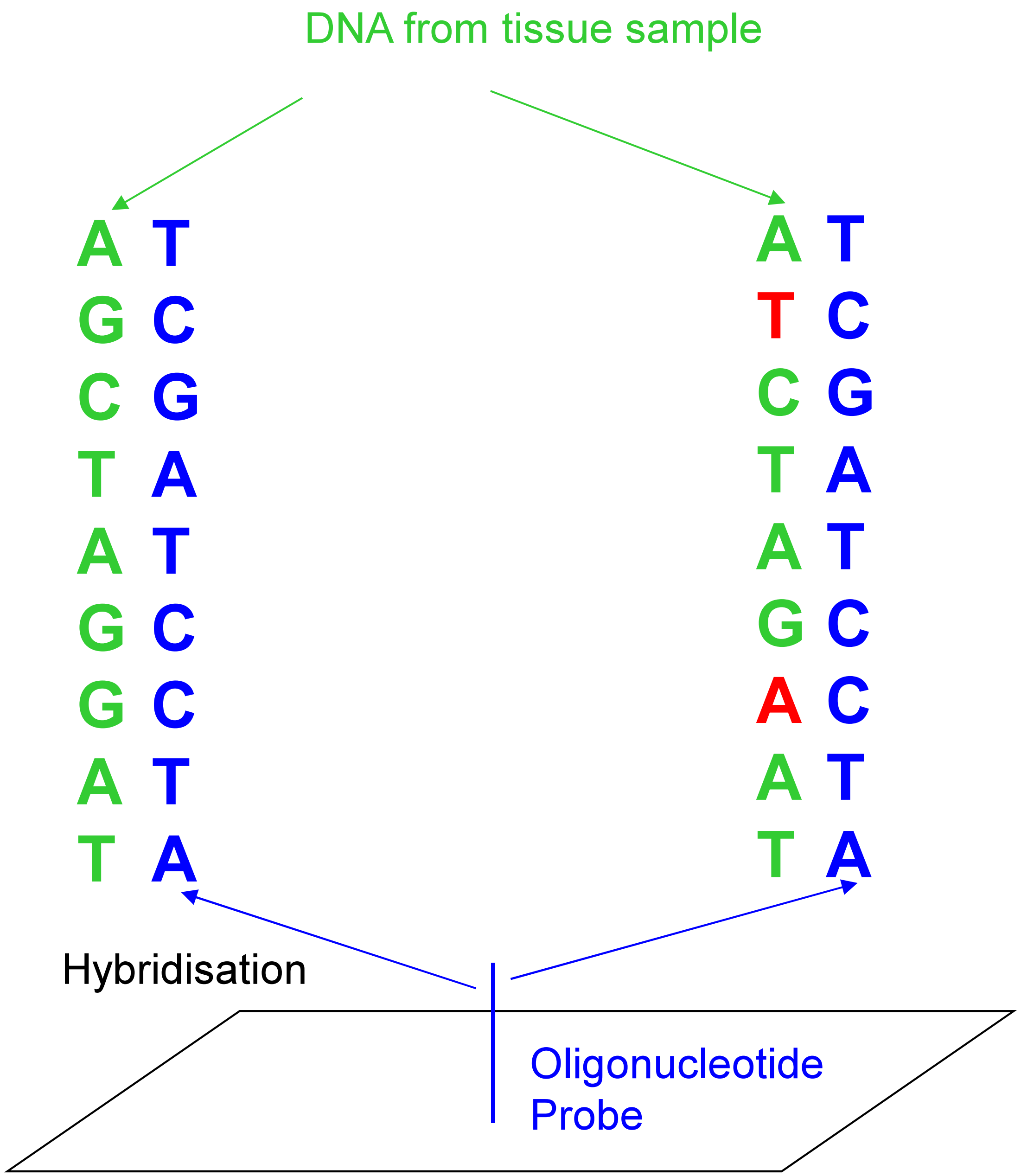
Microarrays vs. Sequencing

UNIVERSITY OF CAMBRIDGE

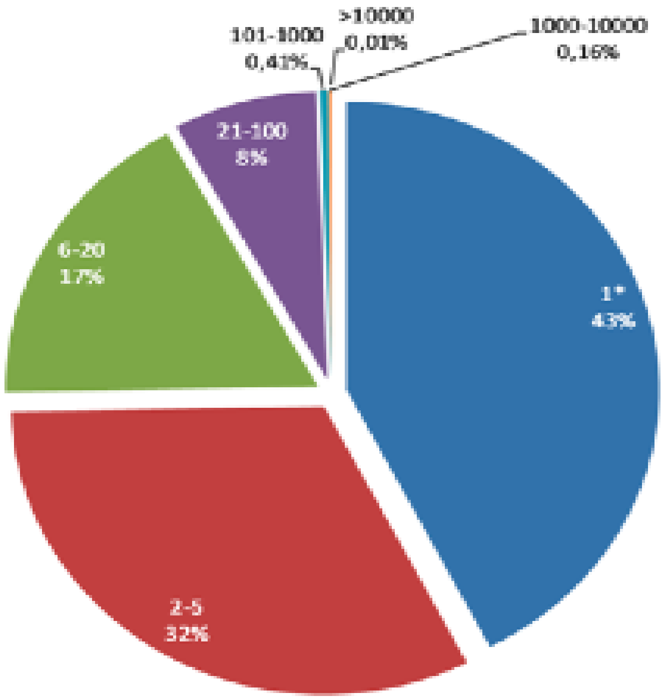
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Why sequencing instead of microarrays for functional genomics ?

- **Cross-hybridisation problem**
  - ***Limits sensitivity***
  - ***Limits range***
  - ***Limits probe coverage***



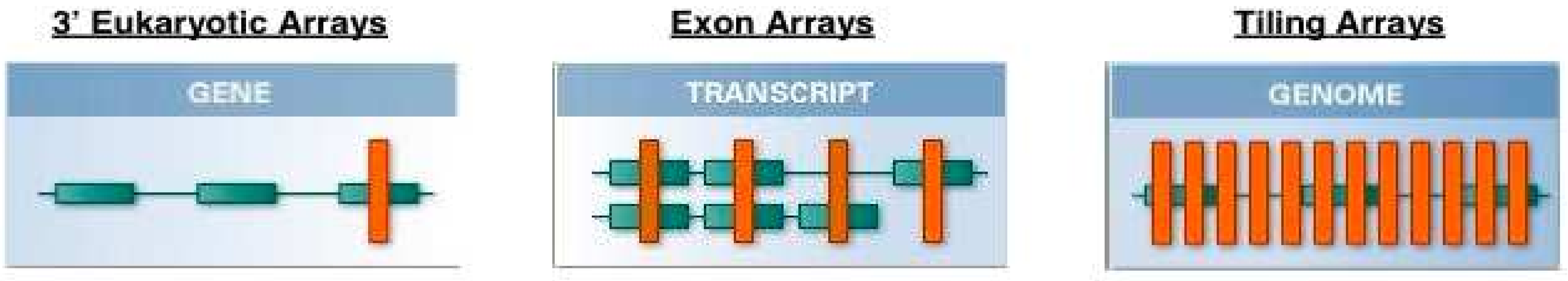
Limits on Sensitivity and Range



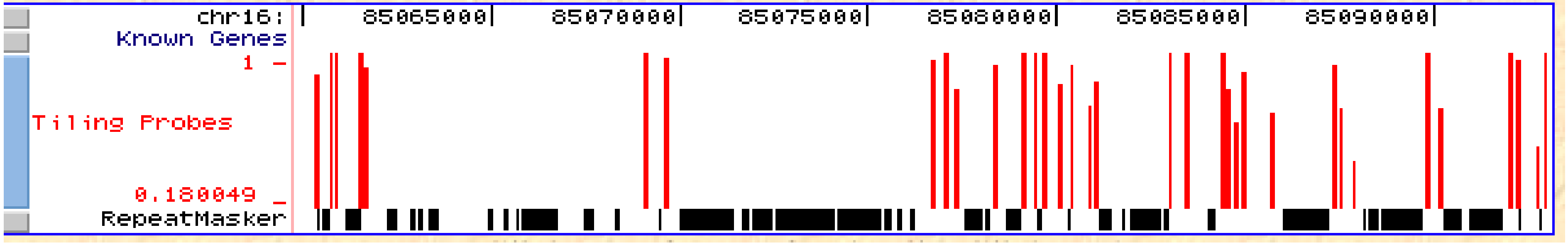
75% of transcripts 1-5 copies per cell

- **Sensitivity (Low abundance transcripts)**
  - ***DE between 1 transcript/cell and 100 transcript/cell***
- **Dynamic range (Very high abundance transcripts)**
  - ***DE between 1K transcript/cell and 100K transcript/cell?***
  - ***10K transcript/cell and 1M transcript/cell?***
  - ***Saturation levels***

Limits on probes



- Repeat regions (cross-hyb again)



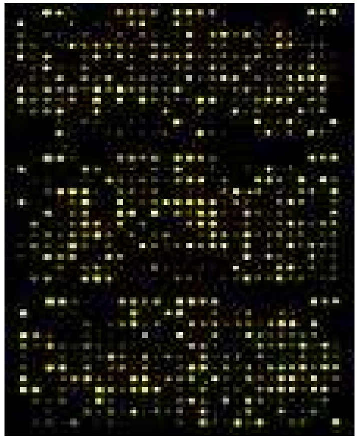
* Limited probe density *Only certain species

You can't find what you're not looking for!

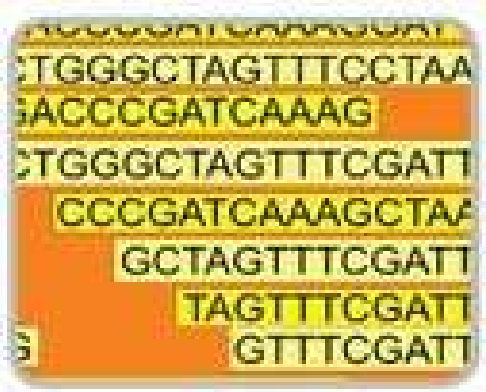
What if we could just count all the fragments, or at least an "unbiased" sample?

- ***Array hybridisation gives relative abundance measures***

- ***Sequencing promises precise quantification - with increased sensitivity, range, and for any genome***



array intensities = analogue signal



sequence reads = digital signal

So why microarrays instead of sequencing?

- ***High-end Sanger sequencing instruments (like those used in Human Genome Project) can now sequence ≈2000 fragments (1-2Mb) per day plus (substantial) sample prep time***
- ***To get decent RNA-seq results, need ≈30M fragments (1Gb) of data per sample for human!***
- ***Until recently, it has been cost-prohibitive to sequence for most genome-wide functional assays***
- ***Even now, microarrays are better for many applications (cost, access/time, maturity of analysis)***

How Next-gen Sequencing Works

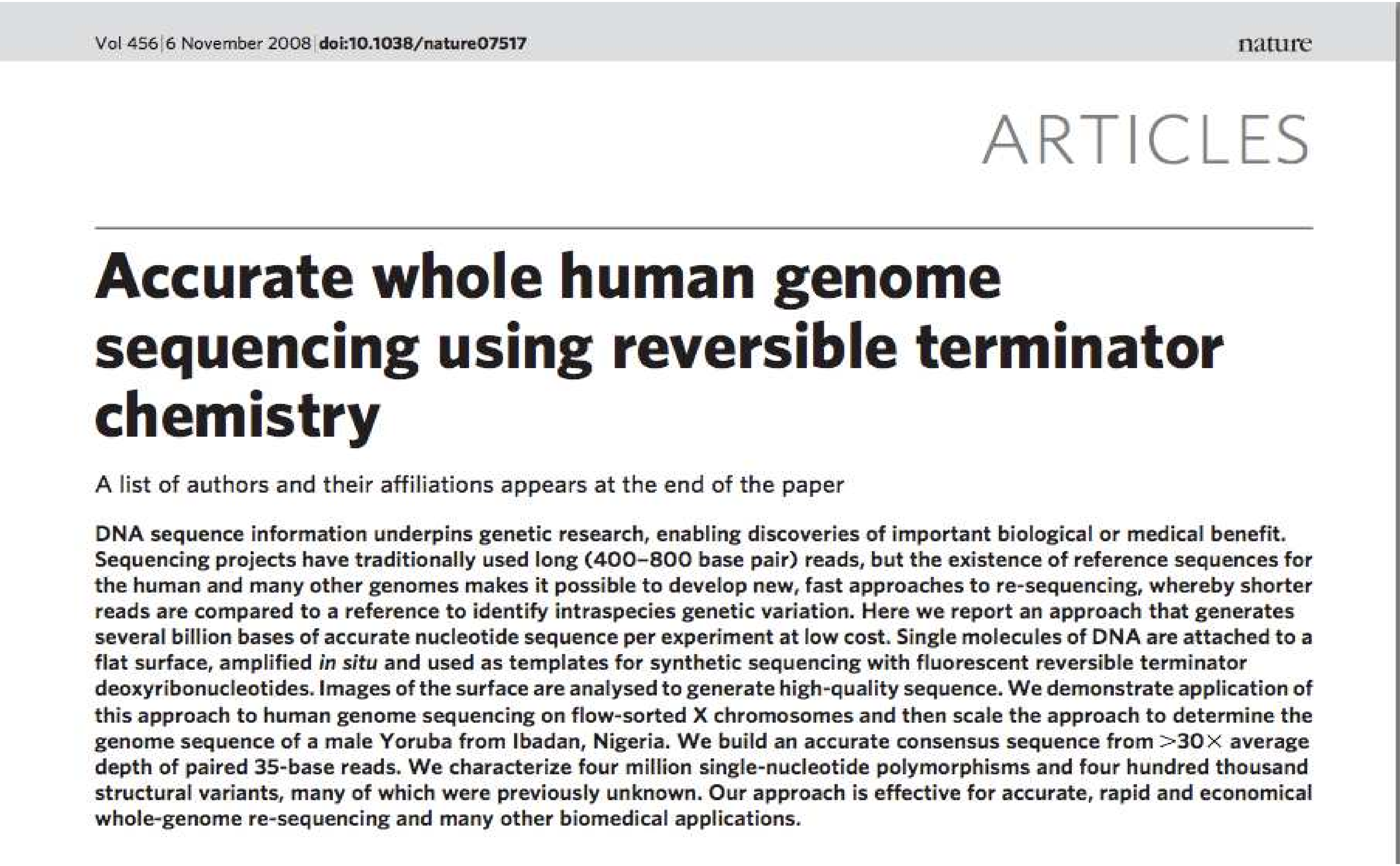
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

“Next generation” sequencing

- Use DNA polymerase (or ligase) to incorporate one base at a time; detect which base (e.g. fluorescent tags, ion detection)
- Key players:
  - Pyrosequencing (Roche 454) - RIP
  - Reversible terminators (Illumina Solexa)
  - Sequencing-by-ligation (Ion Torrent2)
  - More coming (nanopores etc.)

Solexa sequencing



Vol 456 | 6 November 2008 | doi:10.1038/nature07517 nature

ARTICLES

Accurate whole human genome sequencing using reversible terminator chemistry

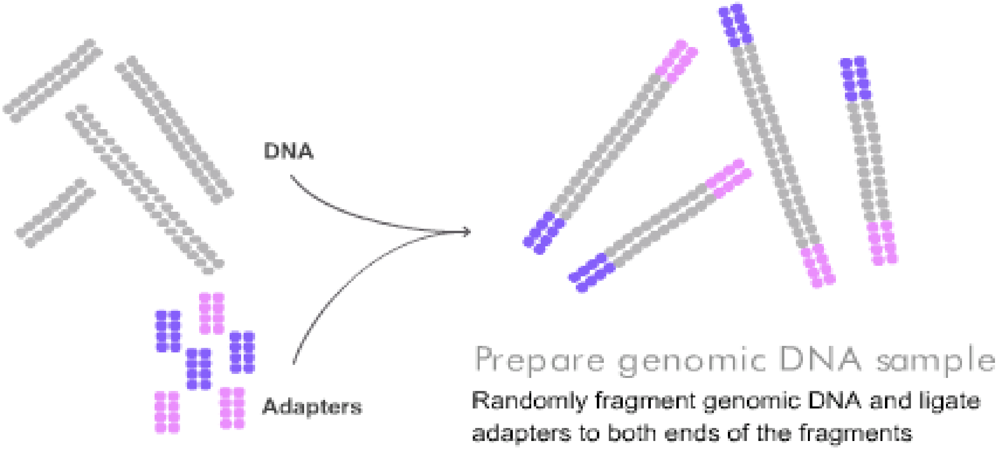
A list of authors and their affiliations appears at the end of the paper

DNA sequence information underpins genetic research, enabling discoveries of important biological or medical benefit. Sequencing projects have traditionally used long (400–800 base pair) reads, but the existence of reference sequences for the human and many other genomes makes it possible to develop new, fast approaches to re-sequencing, whereby shorter reads are compared to a reference to identify intraspecies genetic variation. Here we report an approach that generates several billion bases of accurate nucleotide sequence per experiment at low cost. Single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surface are analysed to generate high-quality sequence. We demonstrate application of this approach to human genome sequencing on flow-sorted X chromosomes and then scale the approach to determine the genome sequence of a male Yoruba from Ibadan, Nigeria. We build an accurate consensus sequence from >30× average depth of paired 35-base reads. We characterize four million single-nucleotide polymorphisms and four hundred thousand structural variants, many of which were previously unknown. Our approach is effective for accurate, rapid and economical whole-genome re-sequencing and many other biomedical applications.

Library Preparation

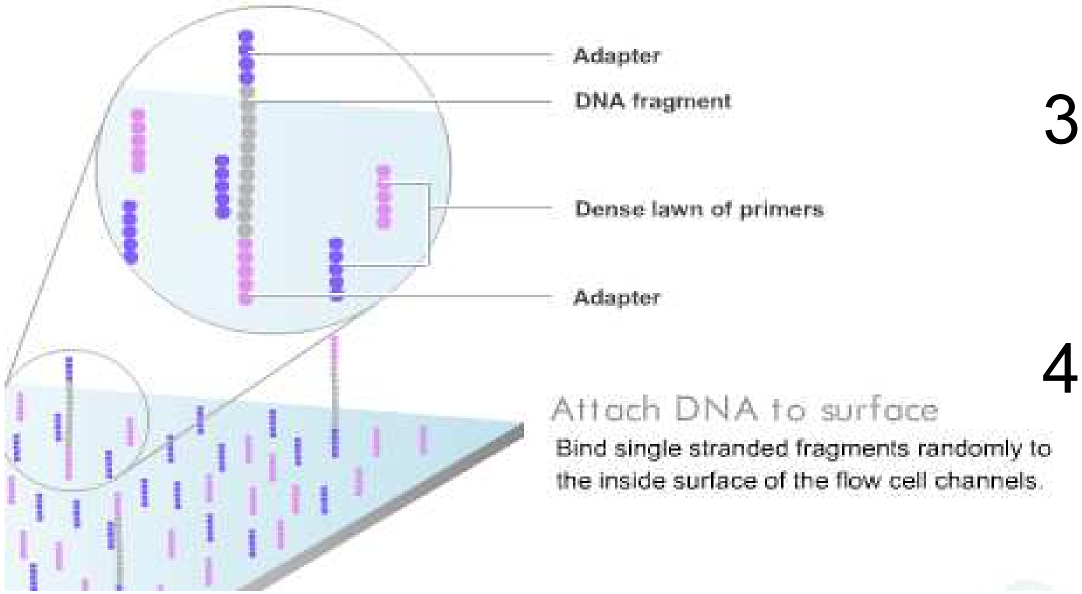
1. Fragment DNA (or cDNA) sample

2. Ligate different adaptors onto each end



Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments



Attach DNA to surface

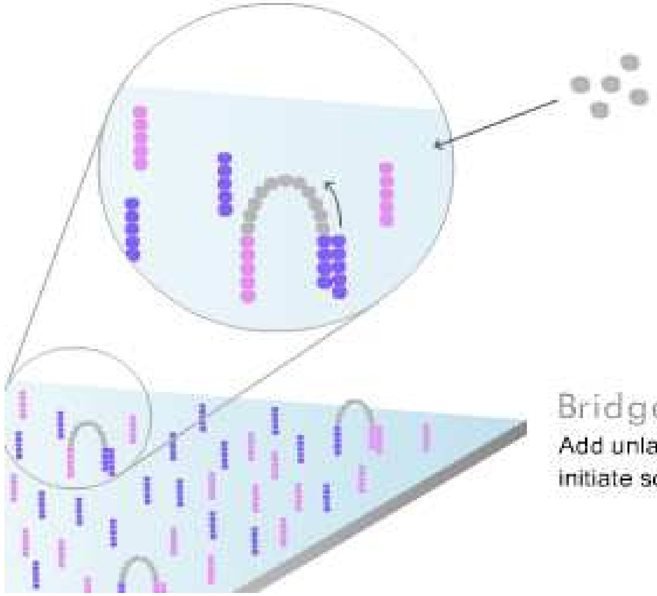
Bind single stranded fragments randomly to the inside surface of the flow cell channels.

3. Denature into single strands

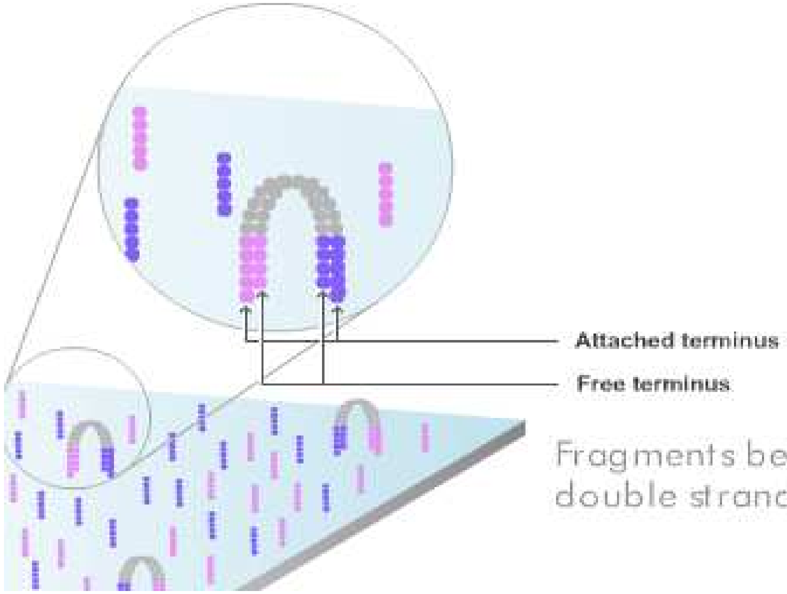
4. Hybridise to flowcell covered with complementary primers

Cluster Generation
(Bridge Amplification)

5. Strands hybridise to primers to form "bridges"



Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



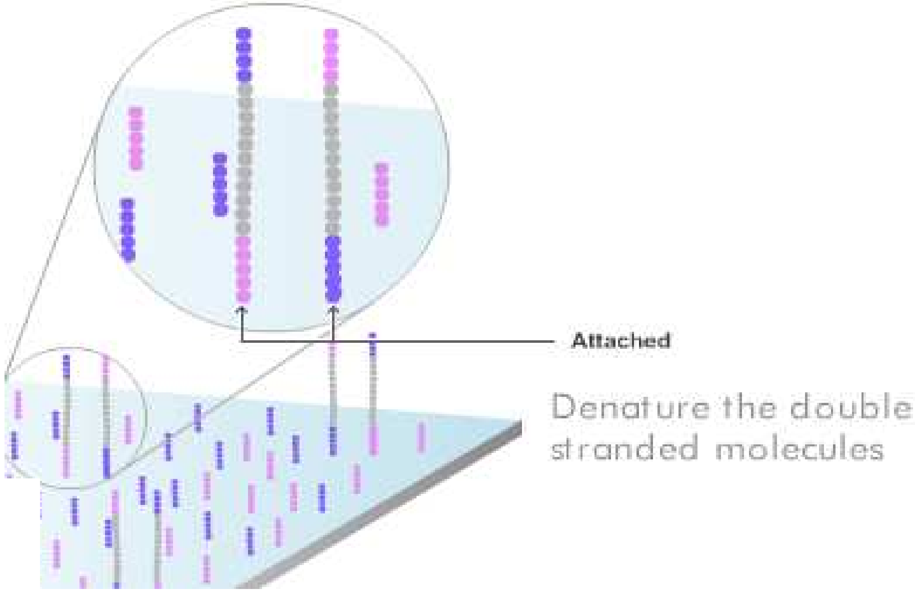
Fragments become double stranded

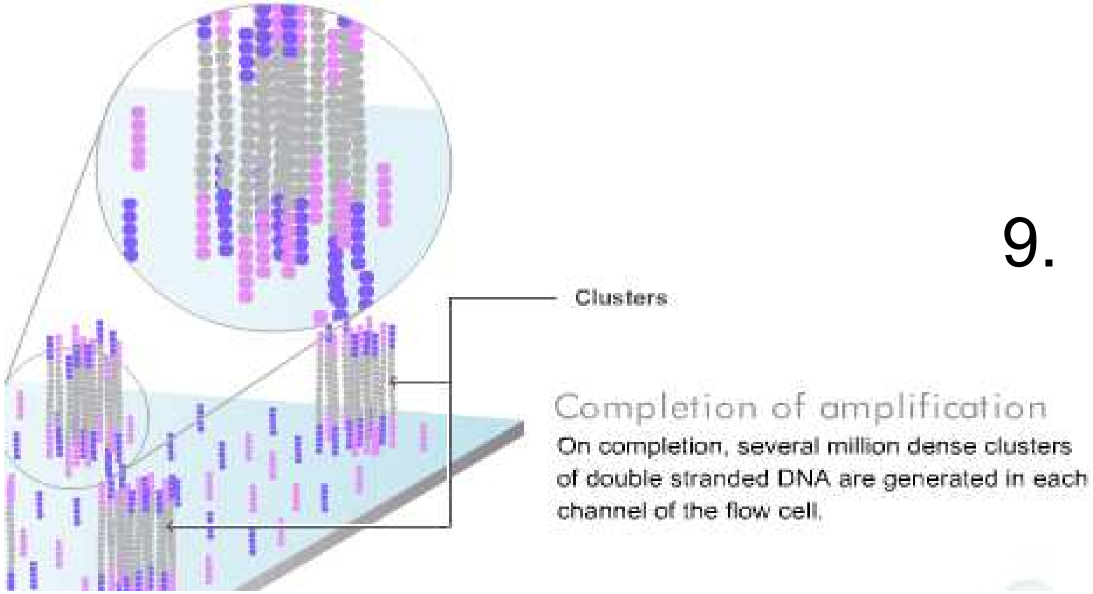
6. Extend from primer to grow second strand

7. Free one terminus of each strand

Cluster Generation (cont)

8. Denature the double strand, forming two strands, each bound on one end





Completion of amplification

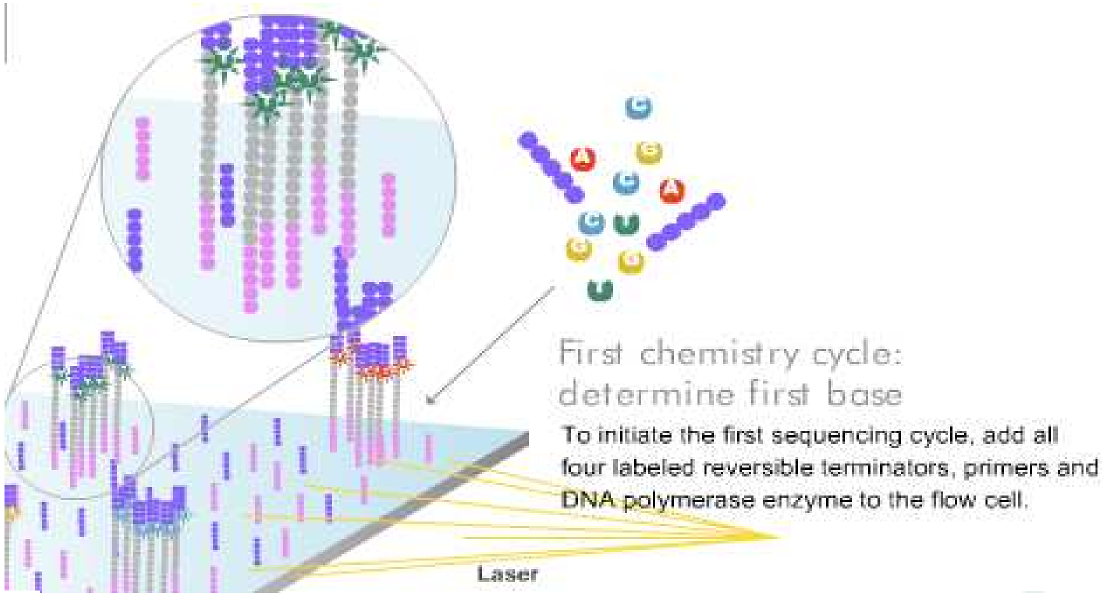
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.

9. Repeat the anneal, extend, denature process until approx.1000 copies of each original molecule per cluster

Sequencing-by-synthesis

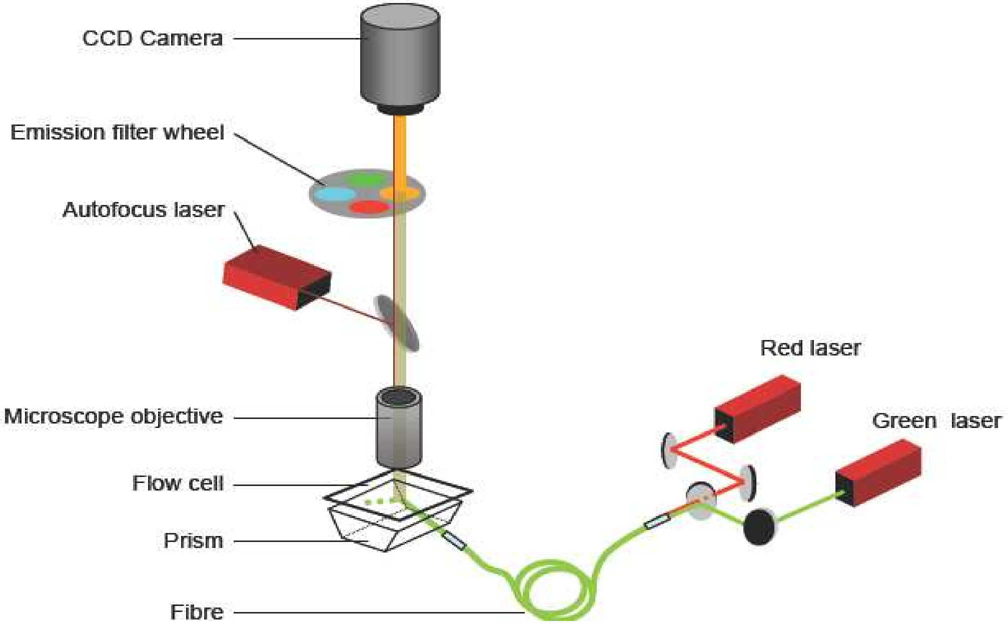
10. Extend each strand by one fluorophore-labelled nucleotide followed by blocked terminus

11. Wash off unincorporated agents



First chemistry cycle: determine first base

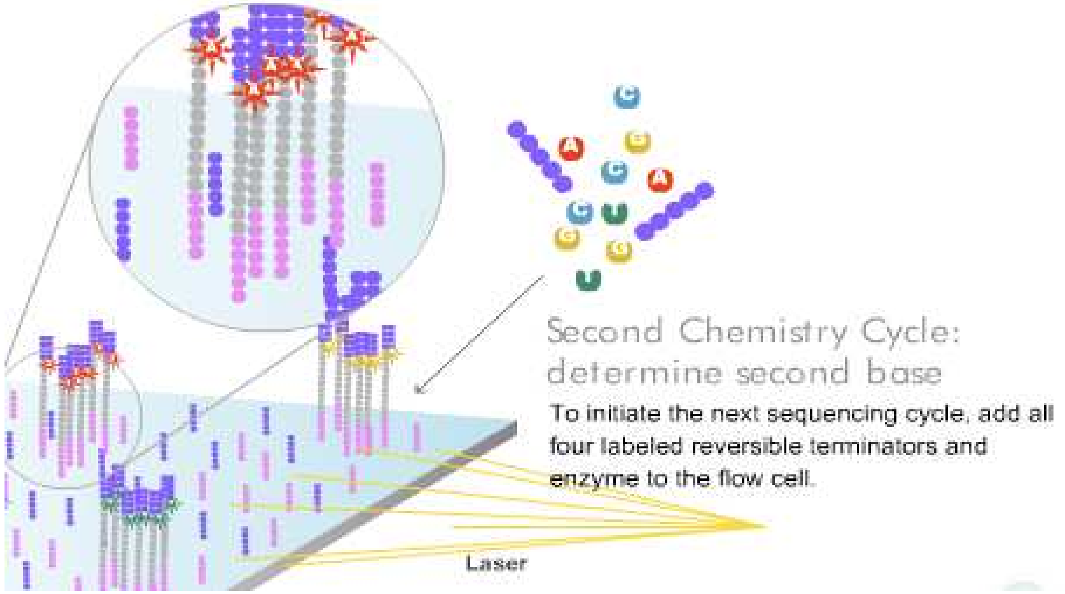
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.



12. Excite clusters with laser to detect which base was incorporated
13. Remove blocked terminus and flourophore

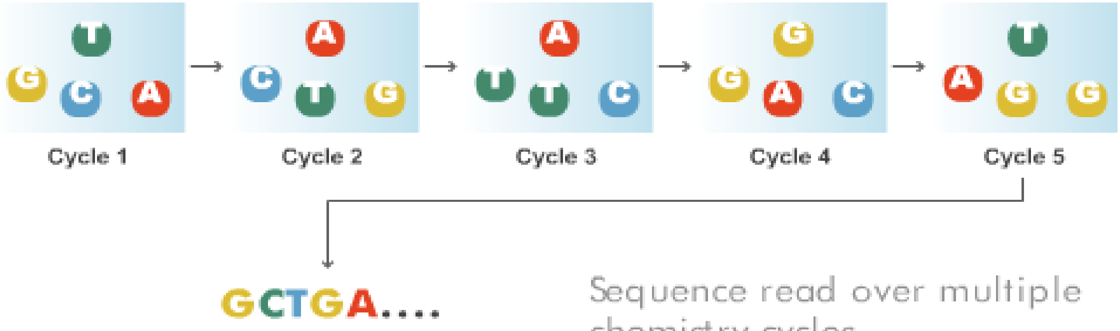
Completing a sequencing run

14. Repeat n cycles, where in is length of sequence read (limited by phasing etc.)



Second Chemistry Cycle: determine second base

To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.



Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

15. Read each cluster’s sequence by determining strongest signal for each cycle

Illumina sequencing movie

(4.5 mins)

- https://youtu.be/fCd6B5HRaZ8

Illumina HiSeq

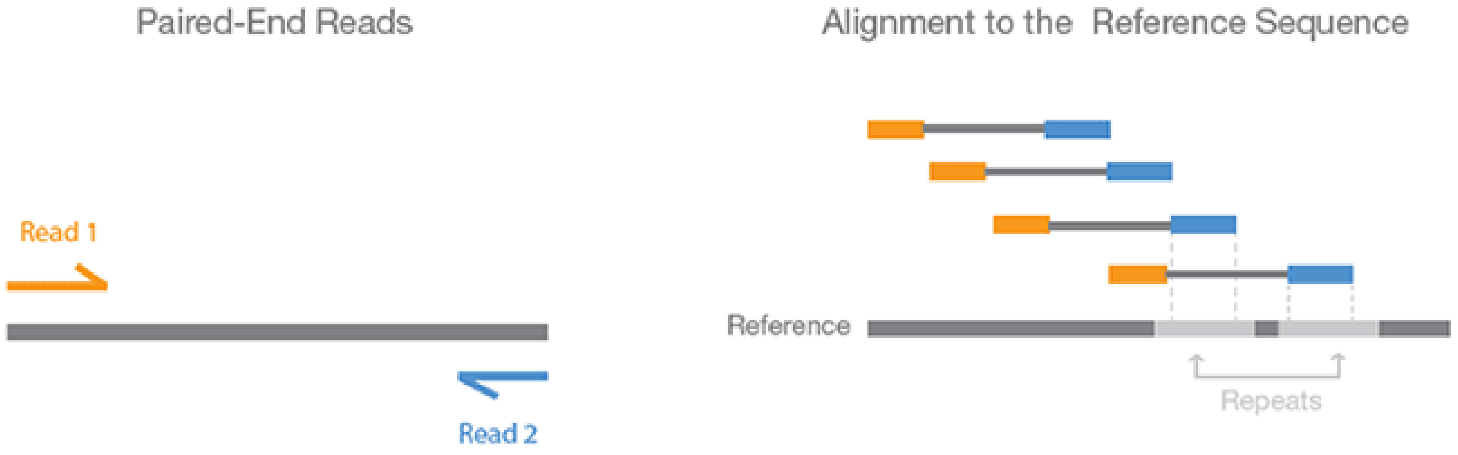
- *Flow cell divided into eight lanes, 350M reads/lane (2.8 billion reads total)*
- *Lanes divided into three rows of eight = 24 image "tiles" on each side*
- *4 images (AGCT) per cycle*
- *4 bases x 24 tiles x 2 sides x 8 lanes x 100 cycles = 153,600 images*
- *At 12MB/image, this is almost 2TB per run*
- *Most of the sequencing run time is spent imaging!*
- *2 Flow cells, never write images to disk (real-time image processing)*
- *"Rapid" mode*

Up to eight samples can be loaded onto the flow cell for simultaneous analysis on the Illumina Genome Analyzer.

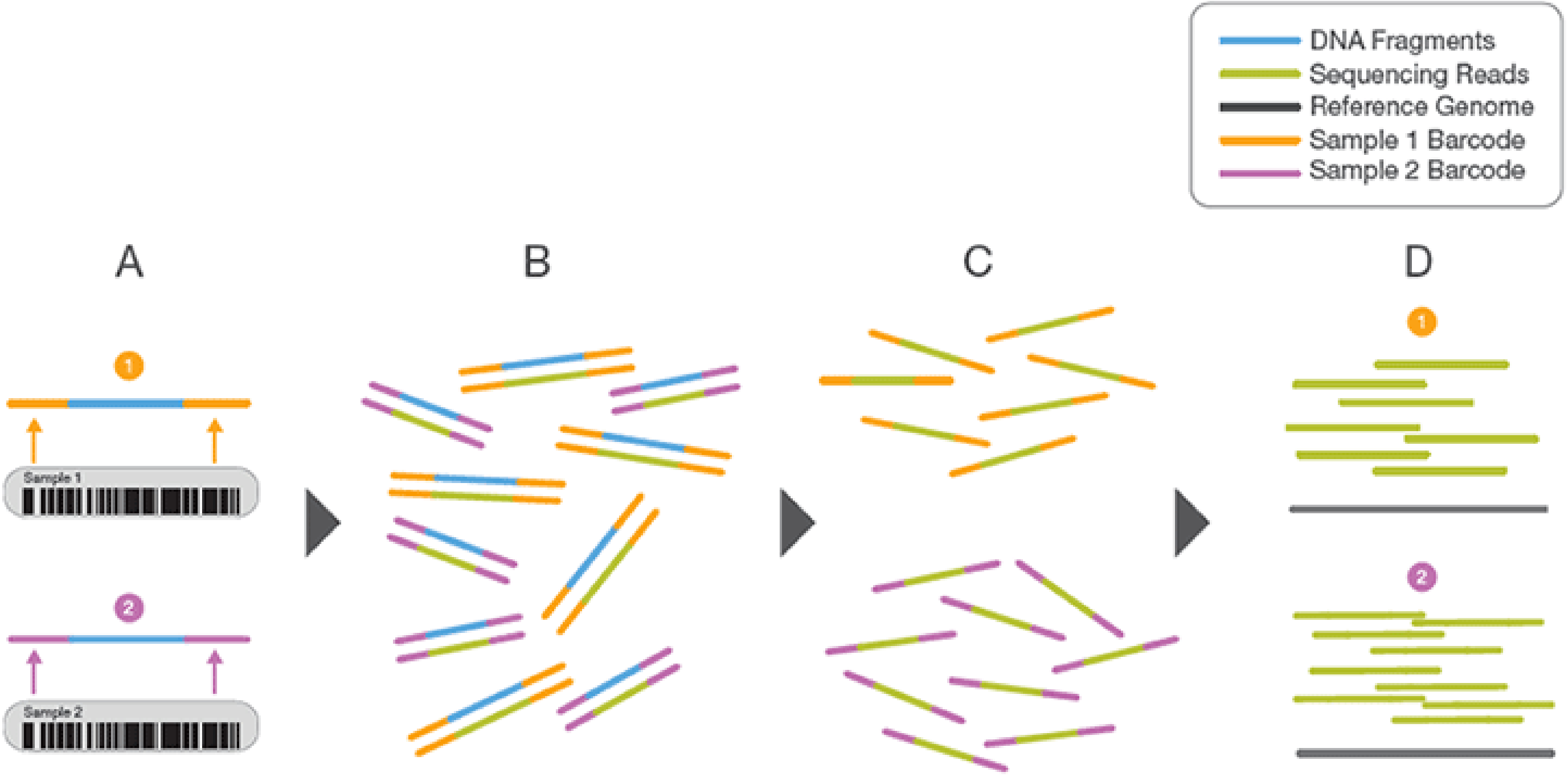
Sequencing Modes

Paired-end sequencing

- Pairs of short reads from each end of a longer fragment can deliver some of the benefits of longer reads
- Each “end” aligned separately
- Can disambiguate non-uniquely aligned reads
- Can help detect transcript isoforms
- Can detect duplications, inversions, chromosomal rearrangements
- Calculation of distribution of insert sizes



Multiplexing/barcoding



A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.

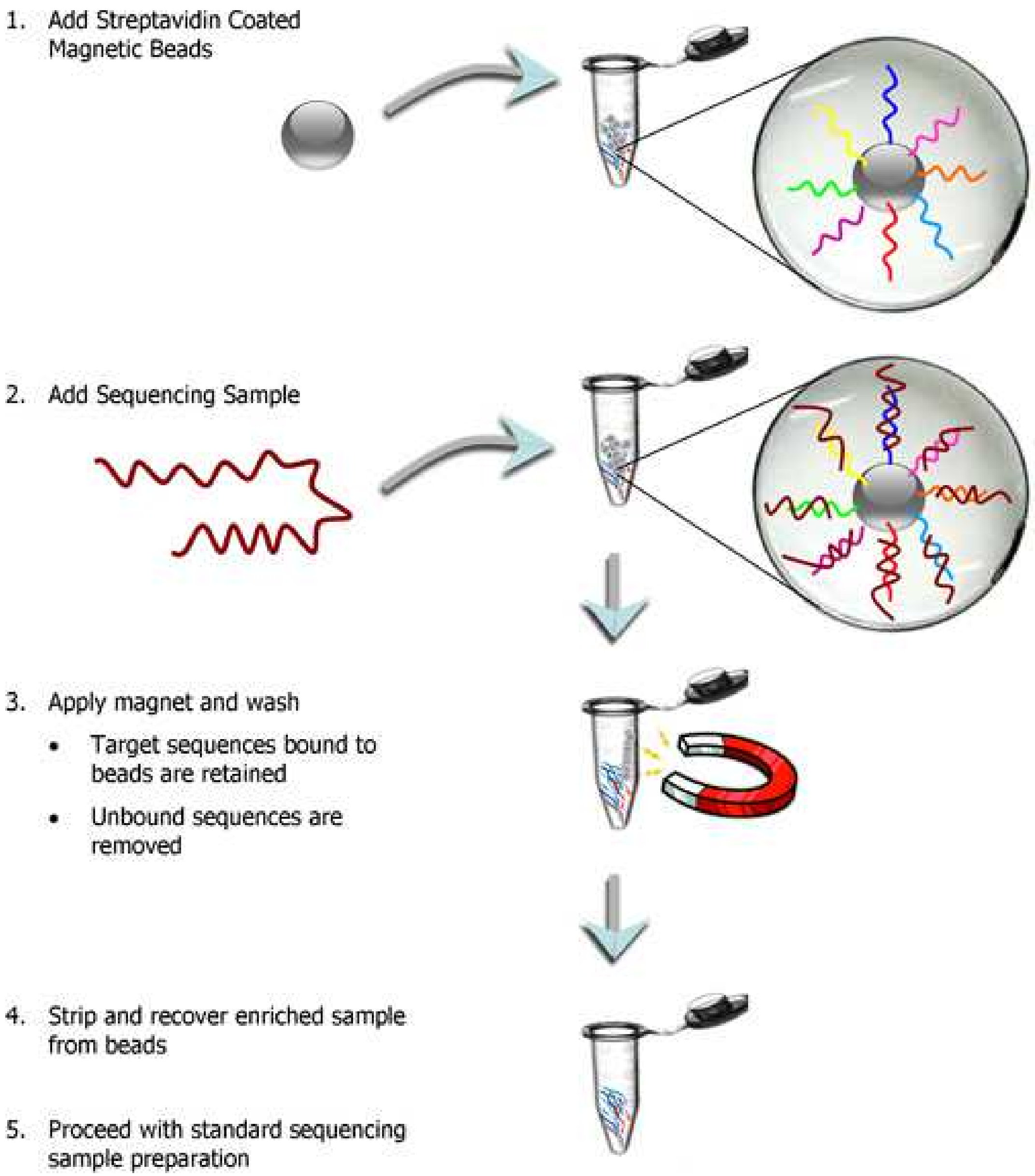
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.

C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.

D. Each set of reads is aligned to the reference sequence.

“Capture” sequencing

- Hybridisation probes “capture” targeted regions of DNA
- Like a microarray, or in solution
- Examples:
  - Exomes
  - Diagnostic panels



1. Add Streptavidin Coated Magnetic Beads

2. Add Sequencing Sample

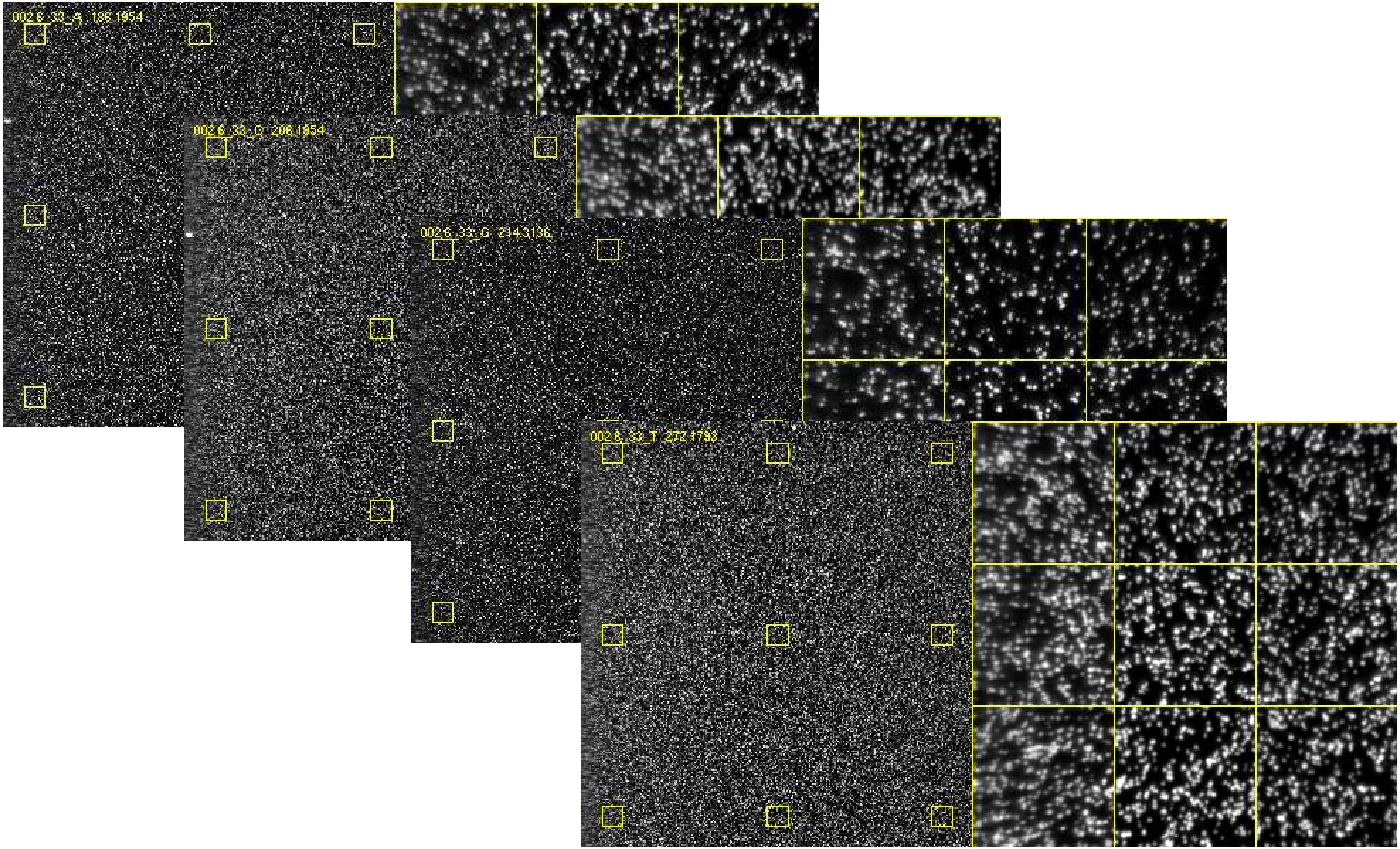
3. Apply magnet and wash
   - Target sequences bound to beads are retained
   - Unbound sequences are removed

4. Strip and recover enriched sample from beads

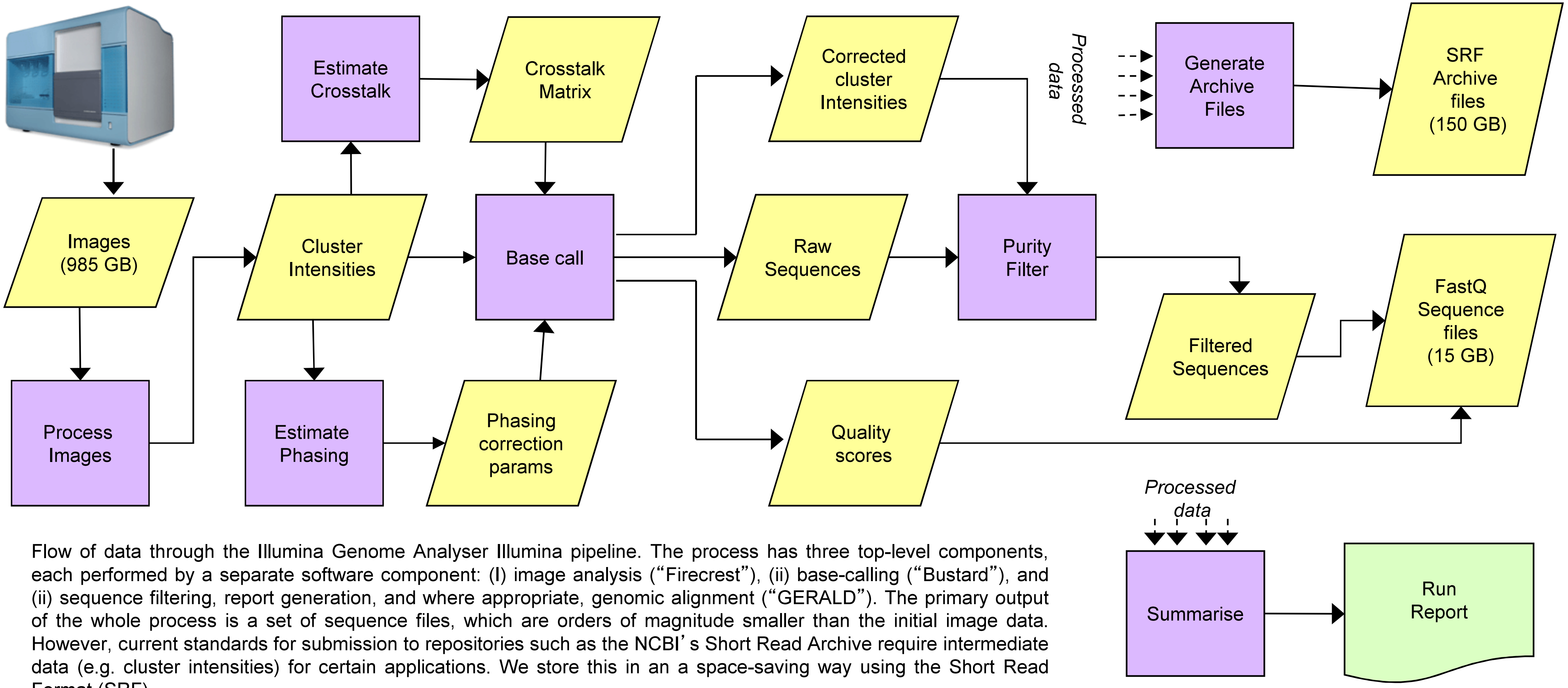
5. Proceed with standard sequencing sample preparation

Processing Sequencing Data

Solexa Images



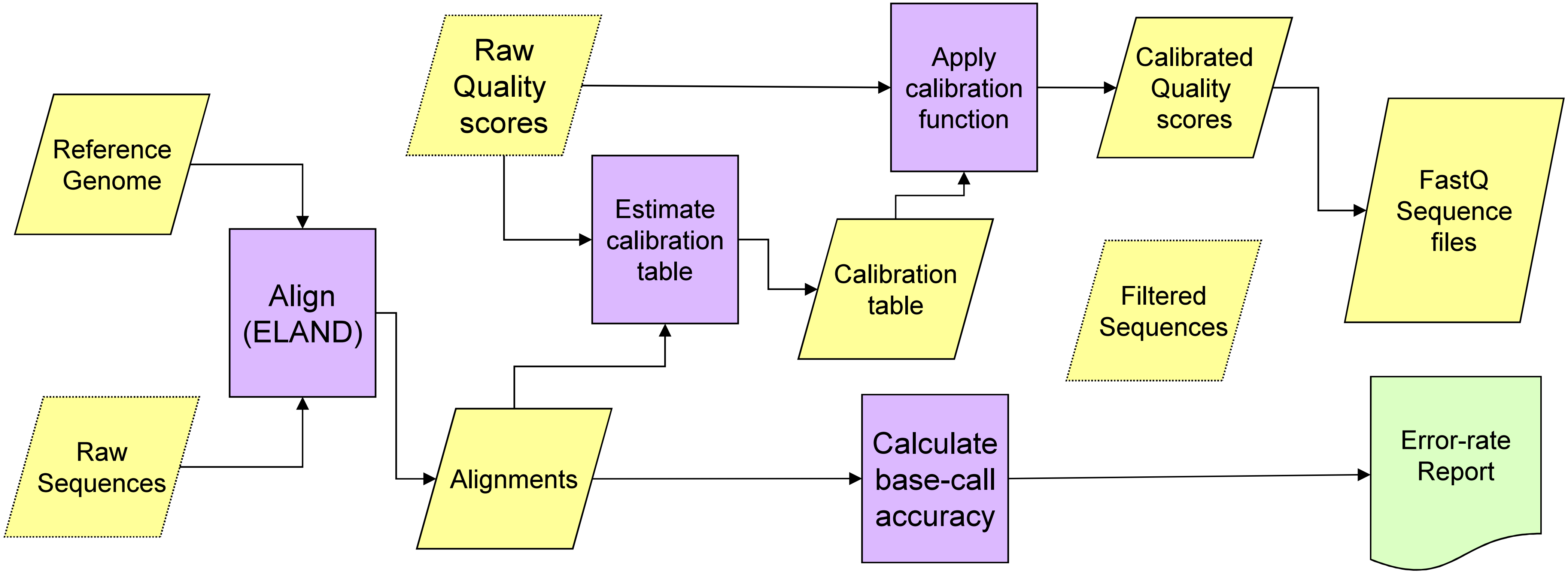
Solexa data processing pipeline



Flow of data through the Illumina Genome Analyser Illumina pipeline. The process has three top-level components, each performed by a separate software component: (I) image analysis (“Firecrest”), (ii) base-calling (“Bustard”), and (ii) sequence filtering, report generation, and where appropriate, genomic alignment (“GERALD”). The primary output of the whole process is a set of sequence files, which are orders of magnitude smaller than the initial image data. However, current standards for submission to repositories such as the NCBI’s Short Read Archive require intermediate data (e.g. cluster intensities) for certain applications. We store this in an a space-saving way using the Short Read Format (SRF).

Thanks to Kevin Howe, CRUK-CRI Bioinformatics Core

Solexa processing pipeline (cont.)



When a reference genome is available, sequence can be aligned to it using the ELAND package (part of GERALD). This allows for (a) calculations of error-rates (inferring the "correct" base calls from the reference); and (b) re-calibration of the per-base quality scores, making them more accurate estimates of the probability that a given base-call is incorrect. From version 1.0, the Illumina pipeline supports cross-calibration (i.e. estimation of the calibration table from control/training data), making calibration possible for datasets where alignment to a reference is not possible/appropriate.

Thanks to Kevin Howe, CRUK-CRI Bioinformatics Core

Short-read sequencing relies on alignment to a reference genome

- **Short reads difficult to assemble**
- **Known genome sequence serves as reference for alignment. Up to 2,500,000,000 separate alignments per run!**
- **Various fast alignment tools available: e.g. ELAND, MAQ, BWA, Bowtie, etc.**
- **Alignment issues:**
  - ***Filtered vs. unfiltered: mostly using filtered now***
  - ***Unique vs. non-unique: how unique is unique?***
  - ***Duplicates (amplification bias)***
  - ***Mismatches and Indels***
  - ***Adapters and index sequence***

Each application (DNA, mRNA, sRNA, PE, etc.) has its own alignment challenges!

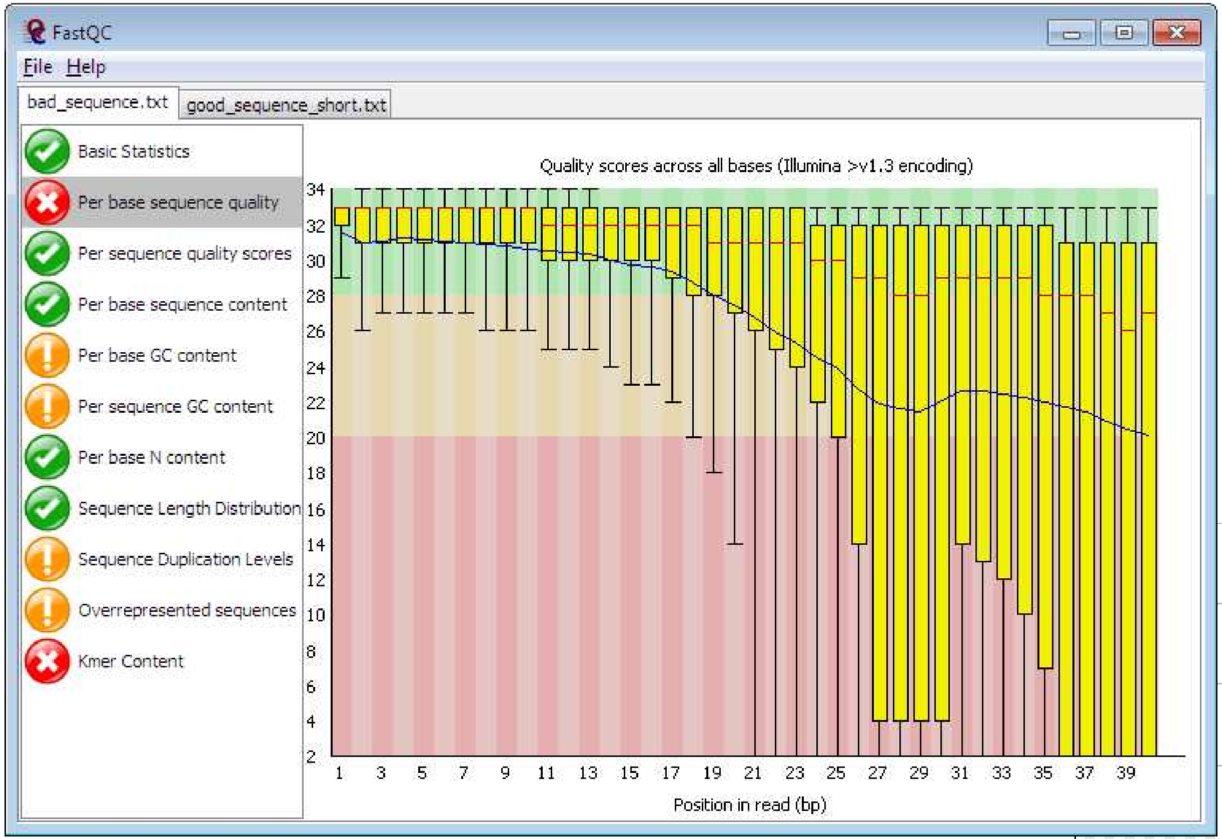
Sequencing Quality Assessment

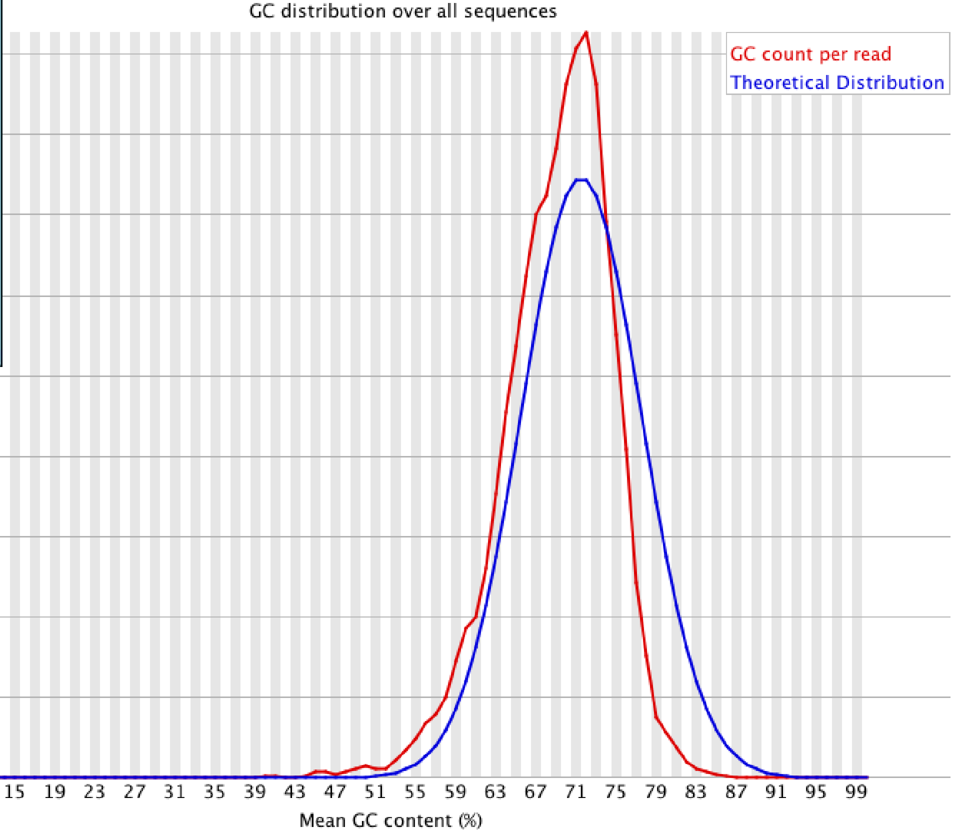
UNIVERSITY OF
CAMBRIDGE

CANCER
RESEARCH
UK

CAMBRIDGE
INSTITUTE

Quality Control: FastQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Multi-Genome Alignment Screen

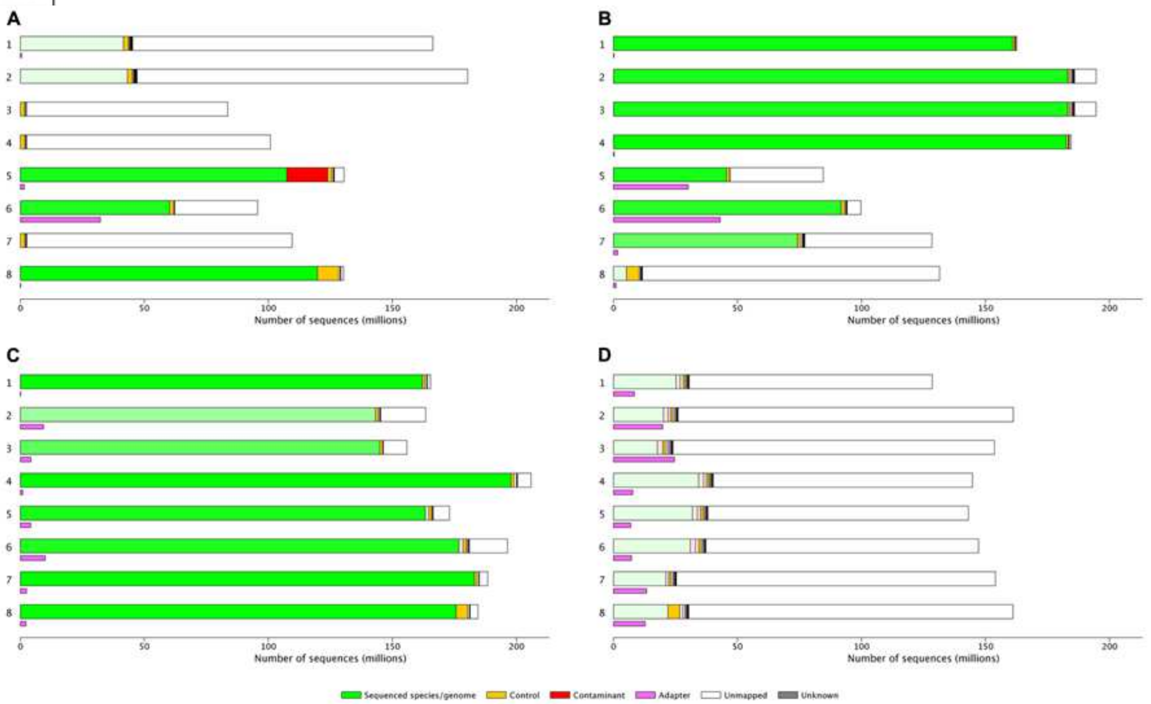
Lane 2

Sequences: 30,191,967
Yield (Gbases): 1.09

Sample ID	User ID	Institute	Sample type	Experiment type	Genome
SLX-4321	EL03	CRI	DNA	ChIP-Seq	Homo sapiens (human)
14-08-2011_6300143_1-Ctrl	14-08-2011_6300143_1	CRI	DNA	N/A	Phi X 174

Sampled sequences: 100,000

Reference ID	Reference Genome	Aligned reads	Aligned %	Error rate	Assigned reads	Assigned %
Hsa.GRCh37	Homo sapiens (human)	72257	72.3%	0.47%	72257	72.3%
Ptr.CHIMP2	Pan troglodytes (chimpanzee)	68222	68.2%	1.42%	524	0.5%
phix174	Phi X 174	402	0.4%	0.24%	402	0.4%
Ggo.gorGor3	Gorilla gorilla	65669	65.7%	1.68%	125	0.1%
Ppy.WU_202	Pongo pygmaeus (orangutan)	57954	58.0%	2.45%	46	0.0%
Xtr.JGI4_1	Xenopus tropicalis (frog)	2774	2.8%	4.91%	17	
Mml.MMUL1	Macaca mulatta (macaque)	40773	40.8%	3.29%	13	
Pha.BCM_v1	Papio hamadryas (baboon)	41420	41.4%	3.31%	7	
Cja.calJac1	Callithrix jacchus (marmoset)	22636	22.6%	3.77%	3	
Cpo.CavPor3	Cavia porcellus (guinea pig)	2754	2.8%	5.96%	2	
Cfa.BROADD2	Canis familiaris (dog)	3222	3.2%	5.67%	2	
Ocu.OryCun2	Oryctolagus cuniculus (rabbit)	2687	2.7%	5.87%	1	
Other	13 reference genomes				181	
Unmapped		26420	26.4%			
Adapter		23491	23.5%			



Hadfield and Eldridge 2014, Front. Genet. 20

Sequencing Trends

Sequencing Trends

- Smaller, cheaper “benchtop” sequencers
- Bigger, expensive, multi-genome sequencers (HiSeq X 10)
- Sequencing as a service
- Long reads (PacBio, Oxford Nanopore)
- “Long range” sequencing (10x)
- Single cell sequencing
- Cloud-based analysis (BaseSpace “apps”)

Sequencing Summary

- Massively parallel short-read sequencing ideal for many functional assays
- Quantity of data (read length, # of reads, run speed) growing faster than computer resources (CPU and especially storage)
- Many issues to be considered when preparing and assessing data for downstream analysis
- Precision of data big leap over array hybridization, but:
- Microarrays still preferable for many assays (SNP, CNV,…)