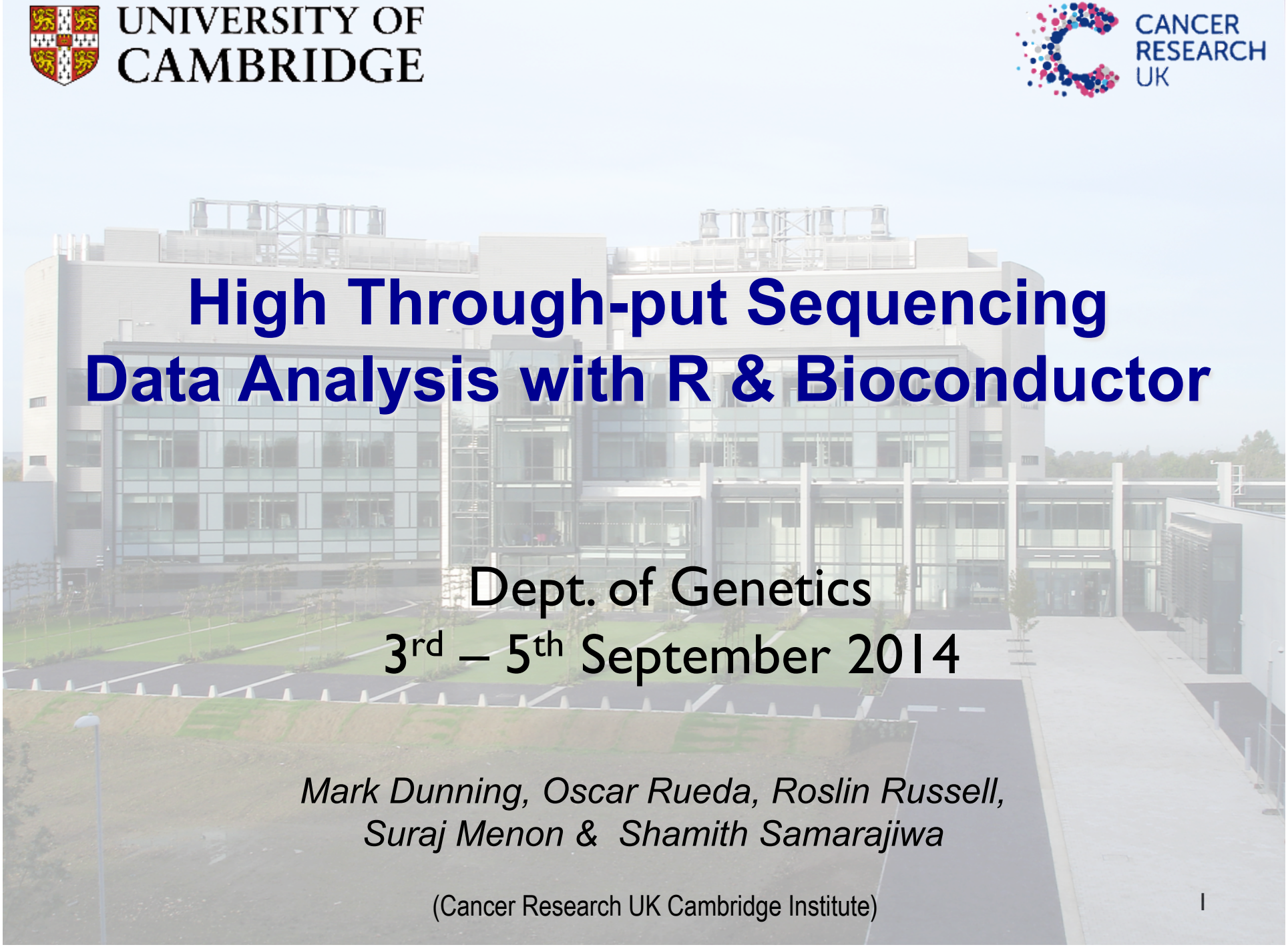


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

High Through-put Sequencing Data Analysis with R & Bioconductor

Dept. of Genetics

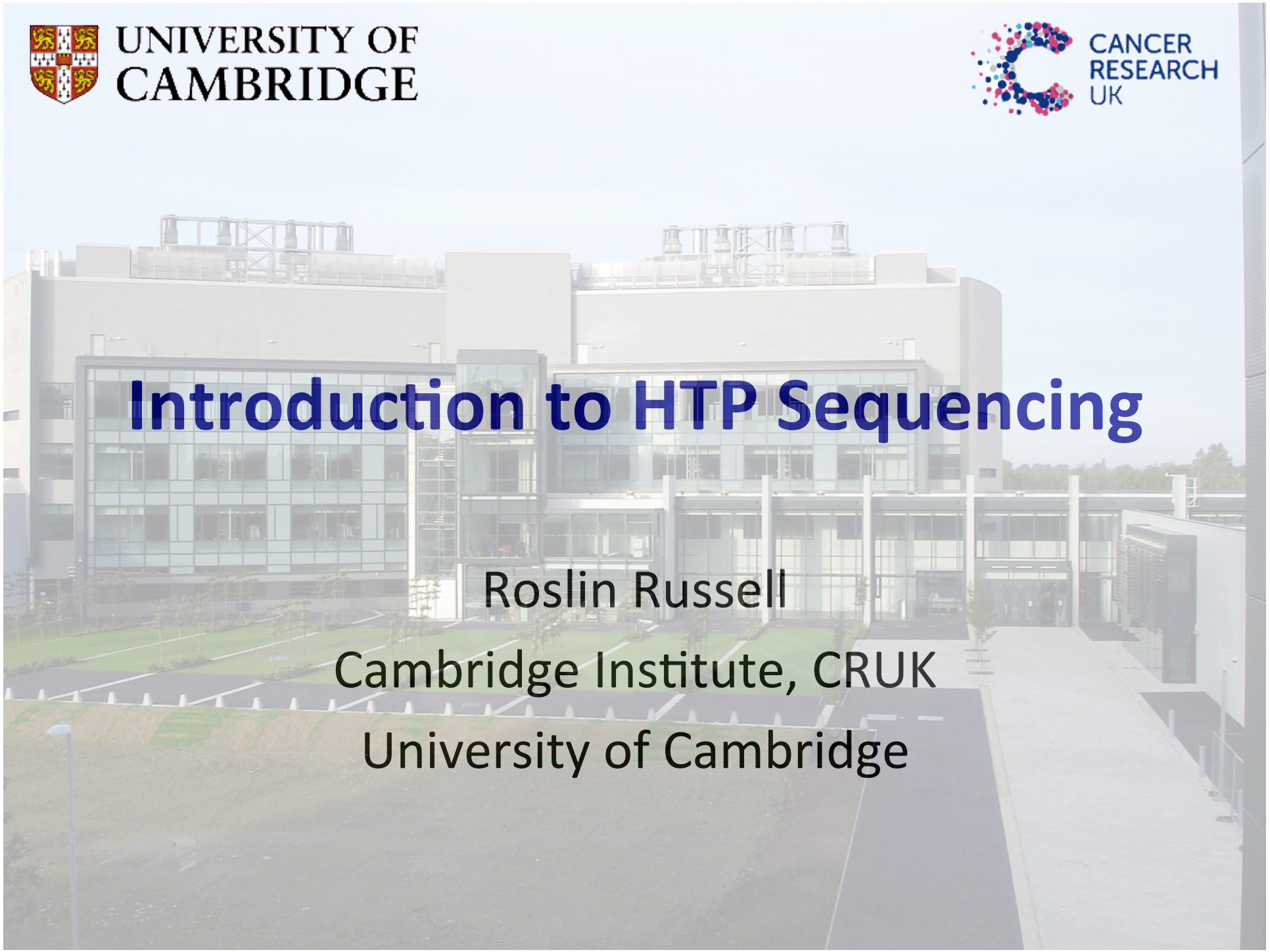
3rd – 5th September 2014

Mark Dunning, Oscar Rueda, Roslin Russell, Suraj Menon & Shamith Samarajiwa

(Cancer Research UK Cambridge Institute)

Course Outline

Day	Lecture (L) / Practical (P)	Speaker
Mon 09:30am – 10:30am	Introduction to HTP Sequencing	RR
Mon 10:30am – 12:00am	Introduction to Strings and Ranges in R (L/P)	MD
Mon 12:00pm – 01:00pm	LUNCH	
Mon 01:00pm – 02:30pm	Introduction to Strings and Ranges in R (L/P)	MD
Mon 02:30pm – 03:30pm	DNA Copy-number Analysis	OR
Mon 03:30pm – 05:00pm	Linear Models, Binomial Distribution, dispersion etc (L)	OR
Tue 09:30am – 10:30am	RNA-seq Data Analysis (L)	RR
Tue 10:30am – 12:00pm	RNA-seq Data Analysis (P)	RR/OR
Tue12:00pm – 01:00pm	LUNCH	RR/OR
Tue 01:00pm – 01:30pm	RNA-seq Data Analysis (P)	RR/OR
Tue 01:30pm – 02:30pm	Genomic Annotation and Visualisation (L)	MD/SS
Tue 02:30pm – 05:00pm	Genomic Annotation and Visualisation (P)	MD/SS
Wed 09:30am – 10:30am	CHIP-seq Data Analysis (L)	SM/SS
Wed 10:30am – 12:00pm	CHIP-seq Data Analysis (P)	SM/SS
Wed 12:00pm – 01:00pm	LUNCH	
Wed 01:00pm – 02:00pm	CHIP-seq Downstream Analysis (L)	SM/SS
Wed 02:30pm – 03:00pm	CHIP-seq Downstream Analysis (L)	SM/SS
Wed 03:00pm – 04:00pm	Further CHiP-seq (Diffbind/RCade) (L/P)	SM/SS
Wed 04:00pm – 05:00pm	Genomic Variants (L)	MD



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Introduction to HTP Sequencing

Roslin Russell

Cambridge Institute, CRUK

University of Cambridge

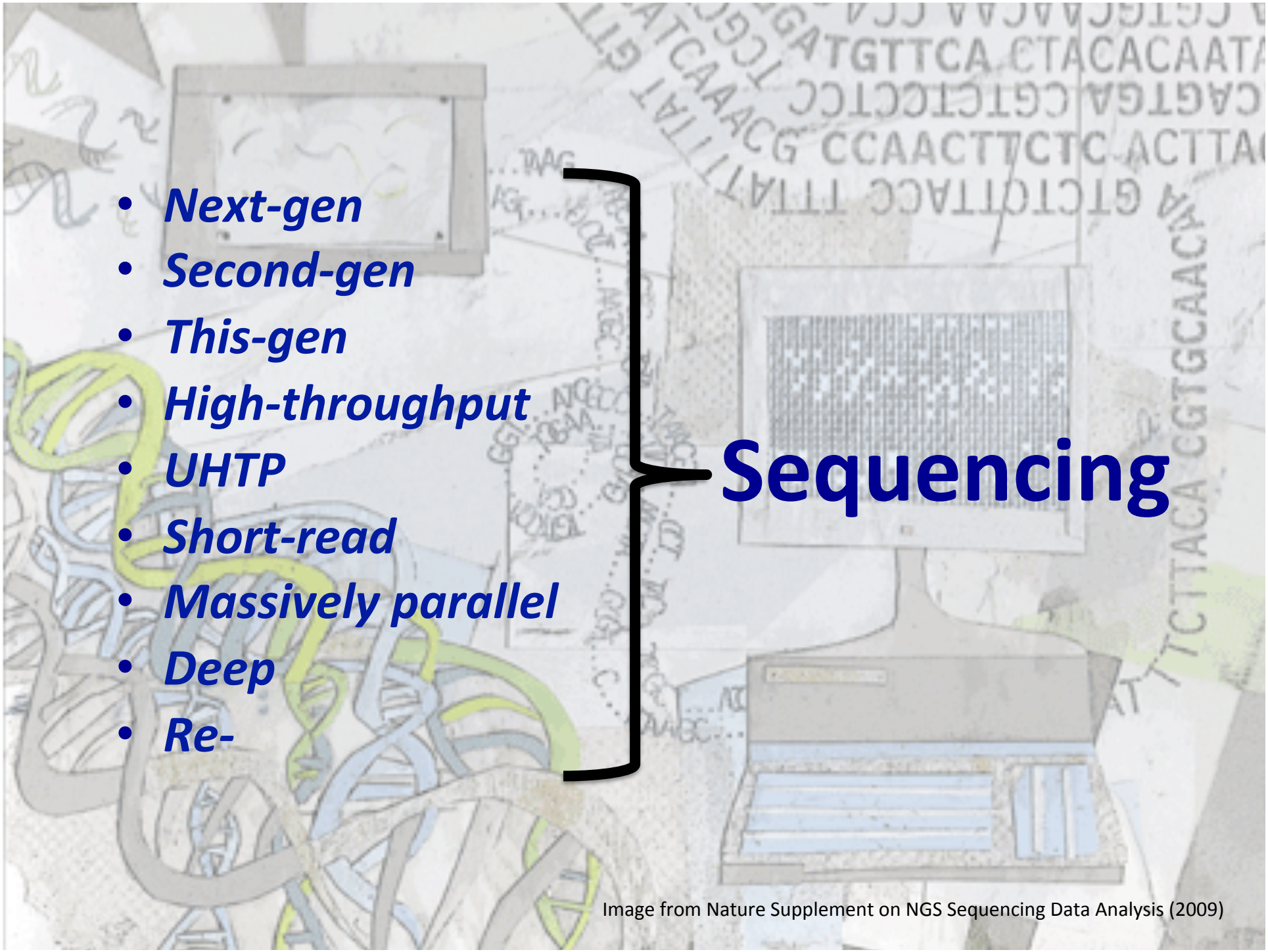
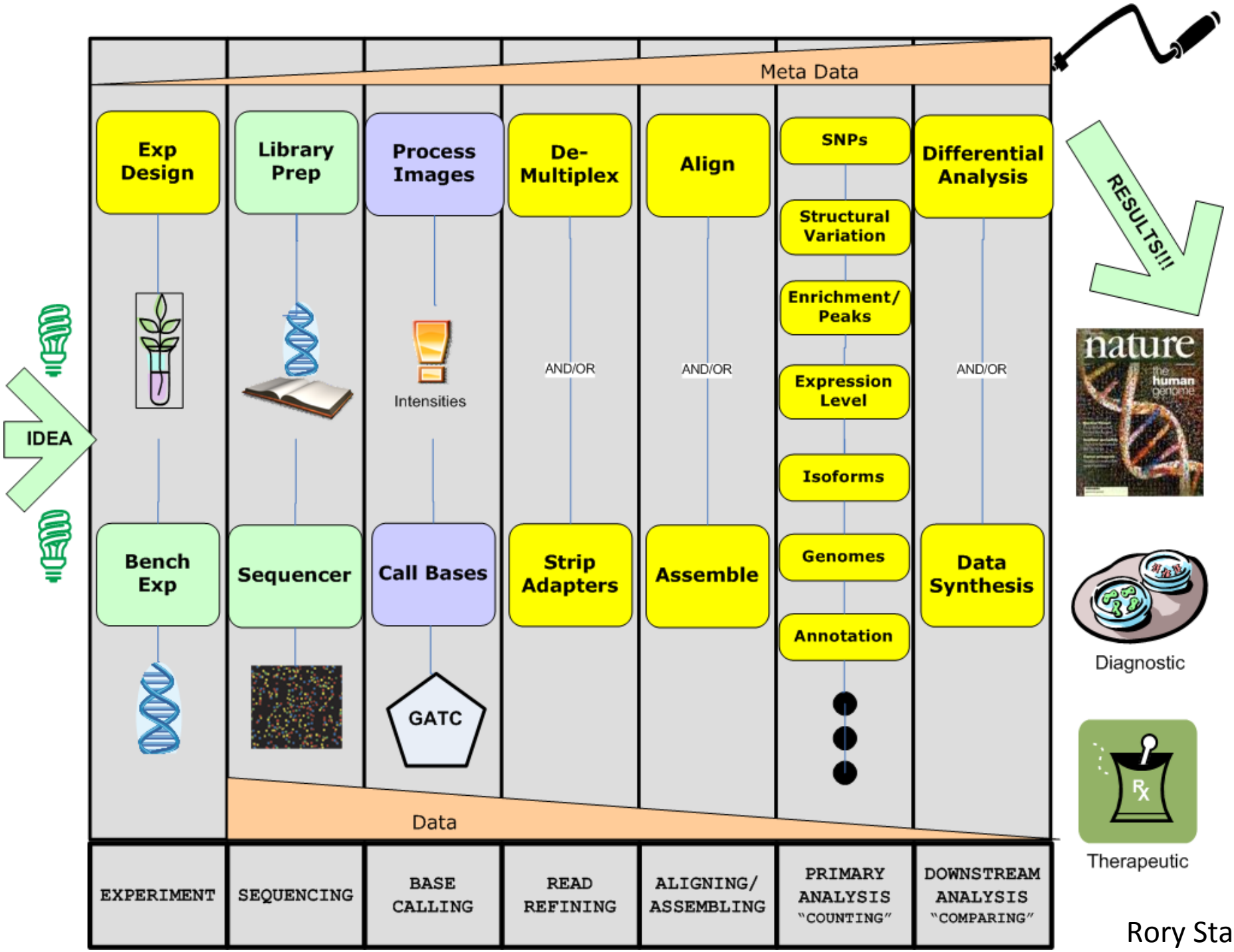


Image from Nature Supplement on NGS Sequencing Data Analysis (2009)


Meta Data
Exp Design
Library Prep
Process Images
De-Multiplex
Align
SNPs
Structural Variation
Enrichment/ Peaks
Expression Level
Isoforms
Genomes
Annotation
Differential Analysis
RESULTS!!!
IDEA
Intensities
AND/OR
AND/OR
AND/OR
Bench Exp
Sequencer
Call Bases
Strip Adapters
Assemble
Data Synthesis
GATC
Data
EXPERIMENT
SEQUENCING
BASE CALLING
READ REFINING
ALIGNING/ ASSEMBLING
PRIMARY ANALYSIS "COUNTING"
DOWNSTREAM ANALYSIS "COMPARING"
nature
the human genome
Diagnostic
Rx
Therapeutic

Rory Stark

Introduction Outline

The Technologies

- **Brief Overview**
  - Illumina Machines
  - Illumina Sequencing
- **Application & Scope**
  - Illumina Applications
  - Experimental Design, Multiplexing & Sample Preparation
- **Moore's Law**
  - Cost, CPU & Storage issues

The Bioinformatics

- **Bioinformatic Workflow**
  - Illumina pipeline
  - Alignment
  - Data formats & FASTQC
  - Integrated Pipelines
- **R & Bioconductor**
  - Background
  - HTP-sequencing Packages
  - R Basics: a refresher
  - R Programming: a refresher

PART 1:
The Technologies

PART 1:
The Technologies

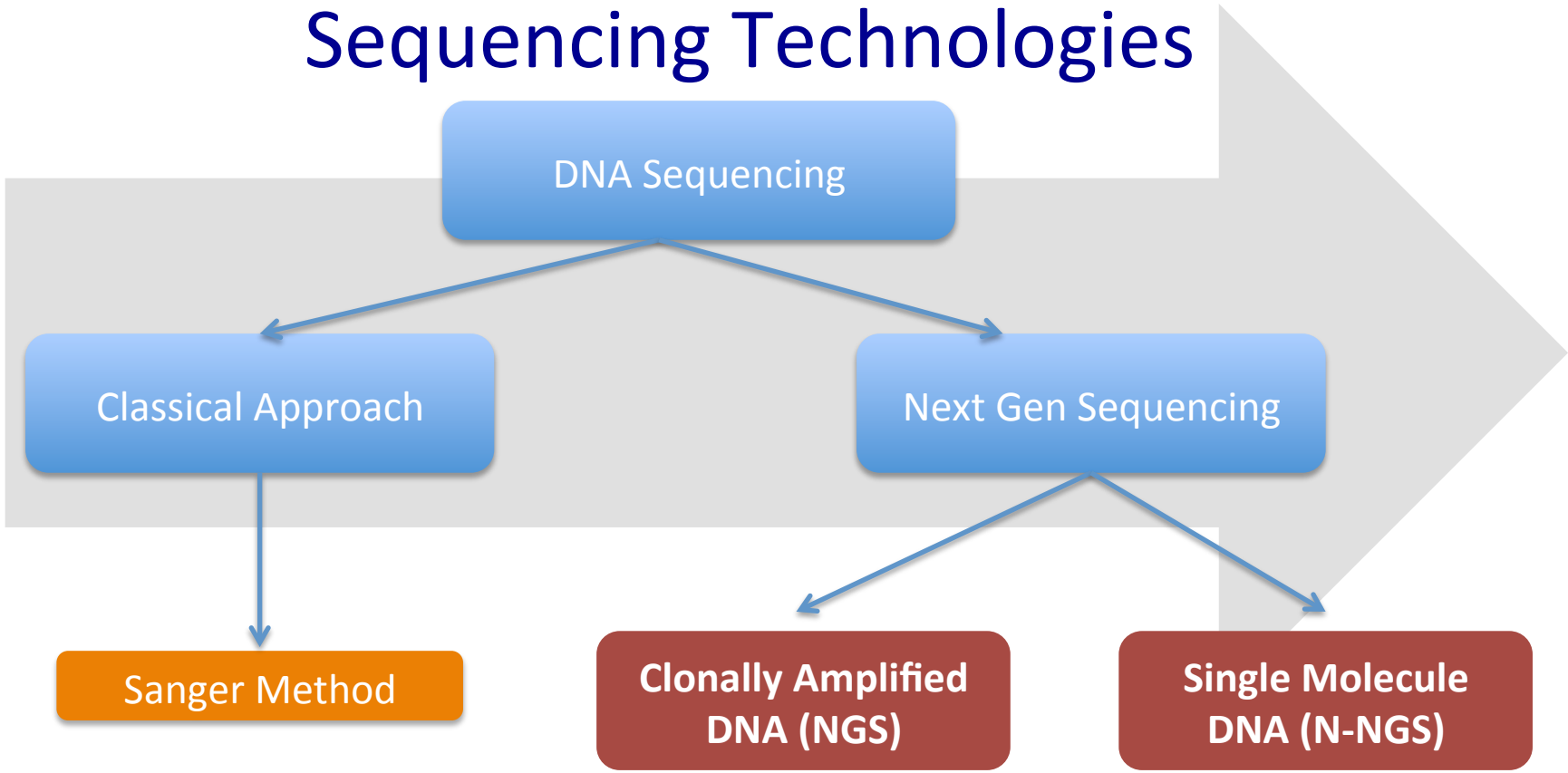
NGS Technologies
(Clonally Amplified DNAs)

Sample Preparation & Multiplexing

Application & Scope

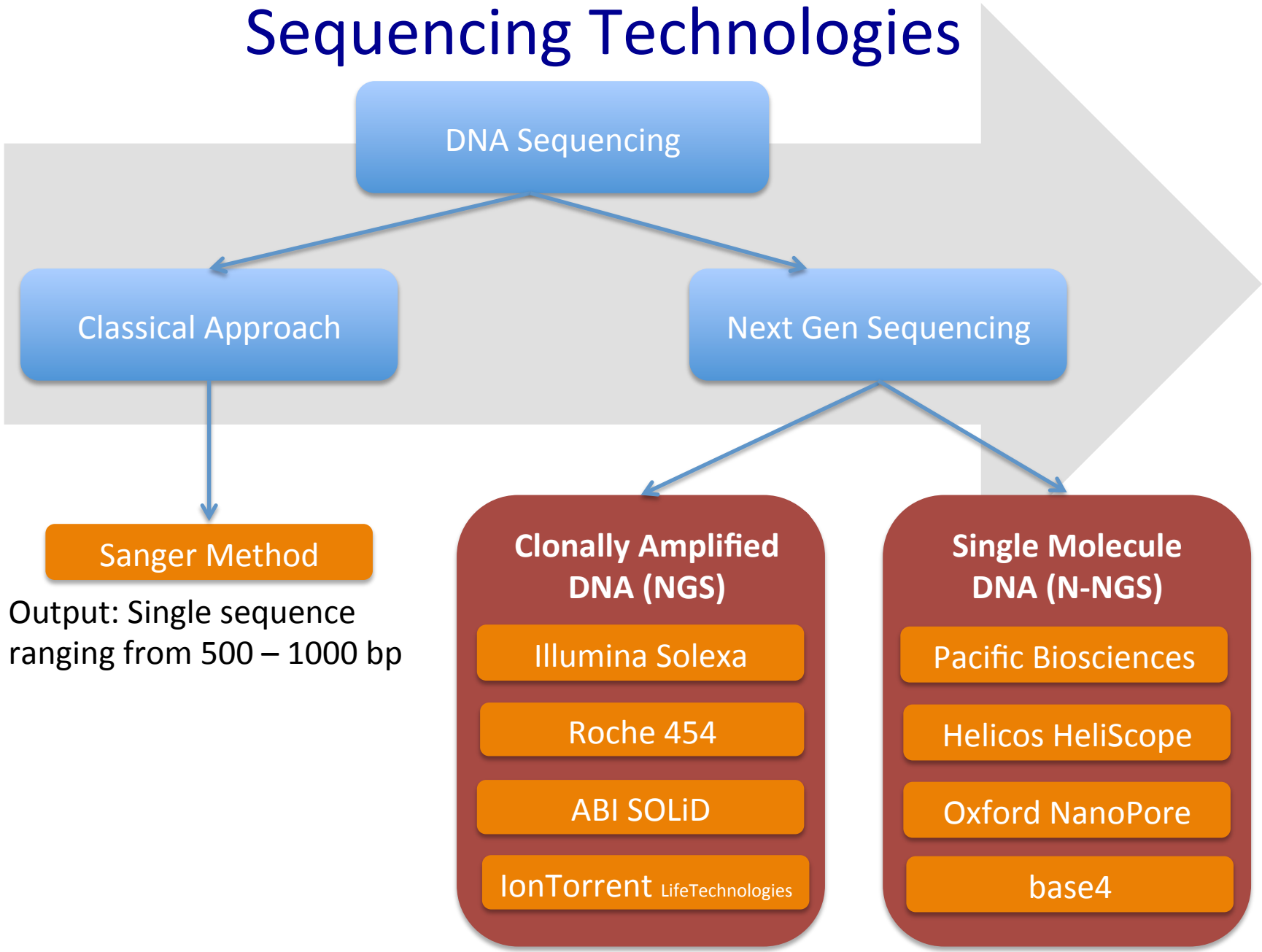
Clinical Application

Sequencing Technologies



High Throughput

Sequencing Technologies

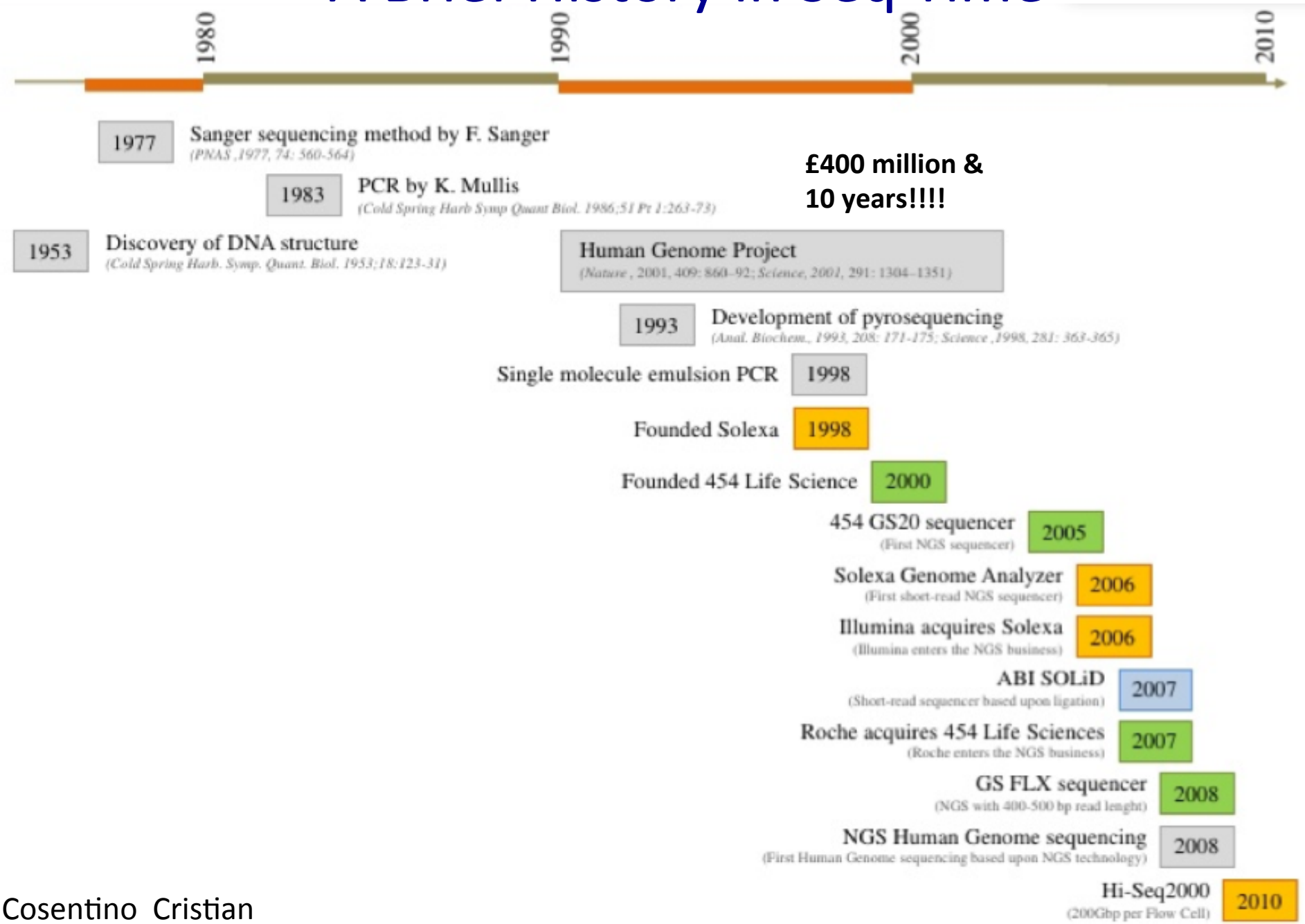


Output: Single sequence ranging from 500 – 1000 bp

Output: Gbp of sequences ranging from 25- 500 bp

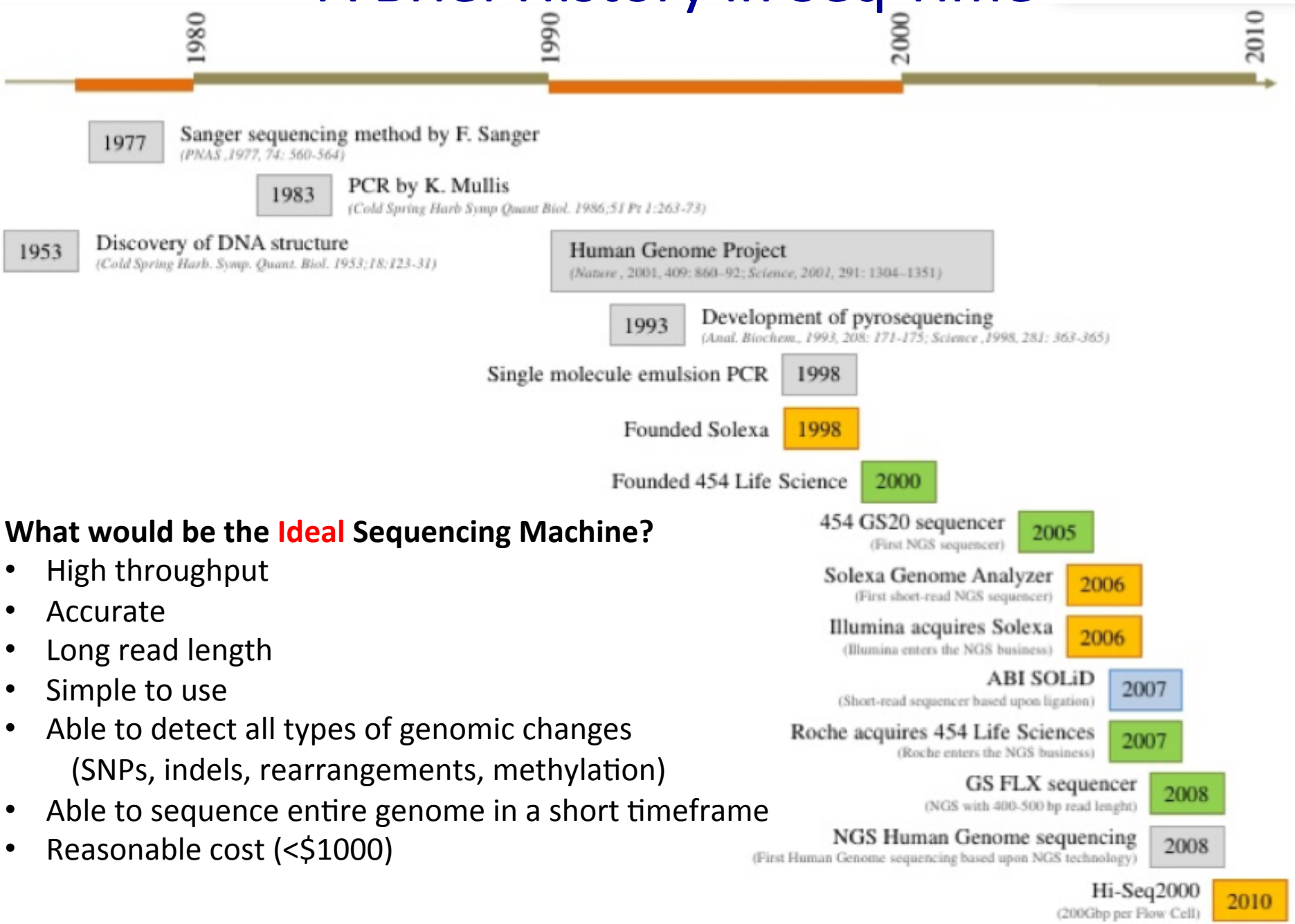
High Throughput

A Brief History in Seq Time



Cosentino Cristian

A Brief History in Seq Time



What would be the Ideal Sequencing Machine?

- High throughput
- Accurate
- Long read length
- Simple to use
- Able to detect all types of genomic changes
  (SNPs, indels, rearrangements, methylation)
- Able to sequence entire genome in a short timeframe
- Reasonable cost (<$1000)

The Trajectory of Throughput: 10 years

Mardis, Nature (2011) 470: 198-203

NGS Instruments

	Cost per base[a]	Read length (bp)[b]	Speed	Capital cost[c]
Minimum cost per base				
Complete Genomics	Low	Short	3 months	None (service)
HiSeq 2000 (Illumina)	Low	Mid	8 days	+++++++
SOLiD 5500xl (Life Technologies)	Low	Short	8 days	+++
Maximum read length				
454 GS FLX+ (Roche)	High	Long	1 day	+++++
RS (Pacific Biosciences)	High	Very long	<1 day	+++++++
Maximum speed, minimum capital cost and minimum footprint				
454 GS Junior (Roche)	High	Mid	<1 day	+
Ion Torrent PGM (Life Technologies)	Mid	Mid	<1 day	+
MiSeq (Illumina)	Mid	Long	1 day	+
Combined prioritization of speed and throughput				
Ion Torrent Proton (Life Technologies)	Low	Mid	<1 day	++
HiSeq 2500 (Illumina)	Low	Mid	2 days	++++++++

Low: < $0.10 per megabase
Mid: > $0.10 < $1 per megabase
High: > $1 per megabase

Short: <200 bp
Long: >400 bp
Very Long: >1,000 bp

Each "+": ~$100,000

Shendure & Aidan (2012) Nature Biotech

Illumina (Solexa) Sequencing

Shankar Balabsubramanium and **David Klenerman**
Founders of Solexa, Cambridge UK, 1997
Acquired by Illumina, 2006

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

"isothermal bridge amplification"

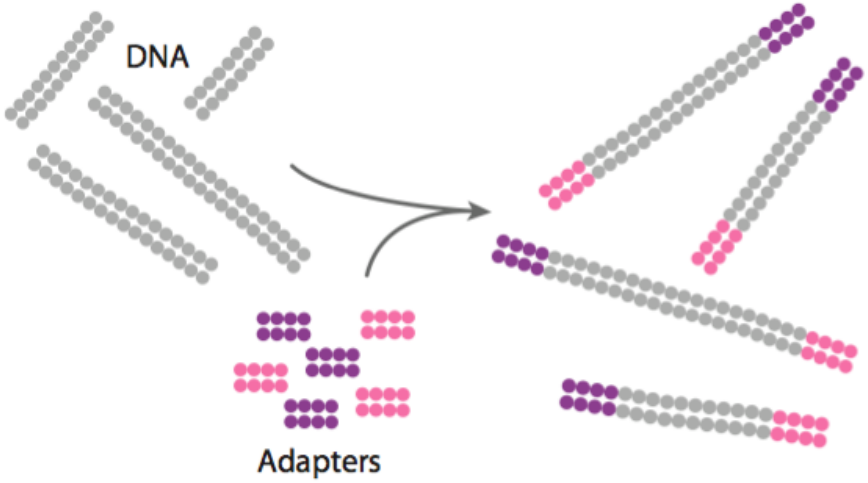
ARTICLES

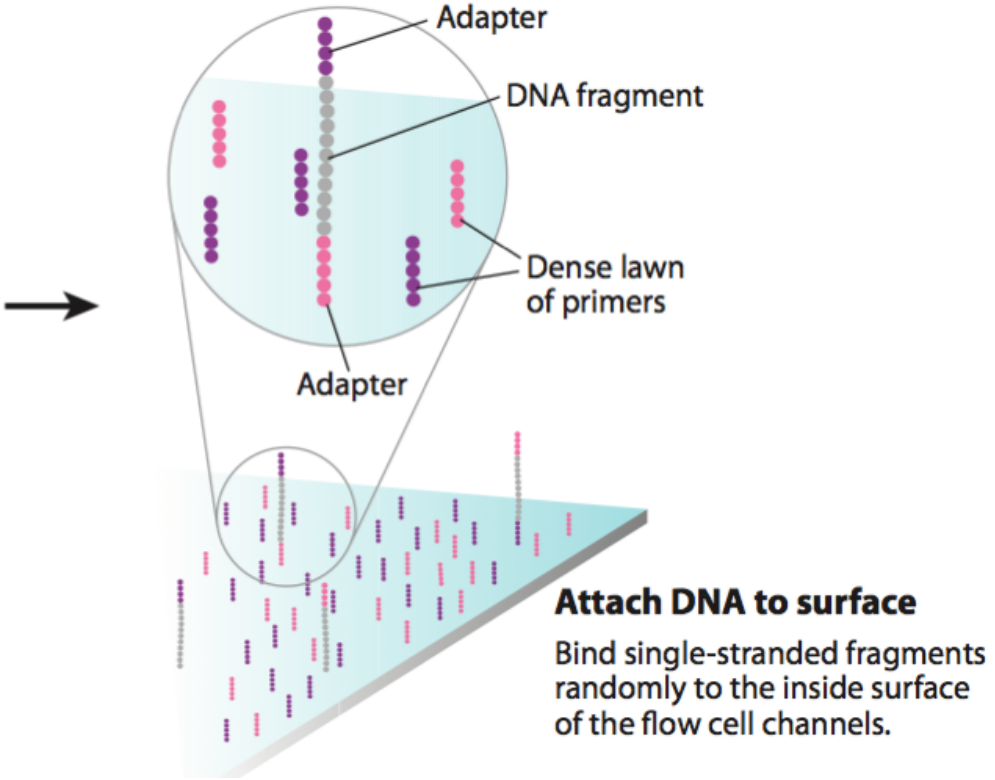
Accurate whole human genome sequencing using reversible terminator chemistry

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.

Illumina Step 1

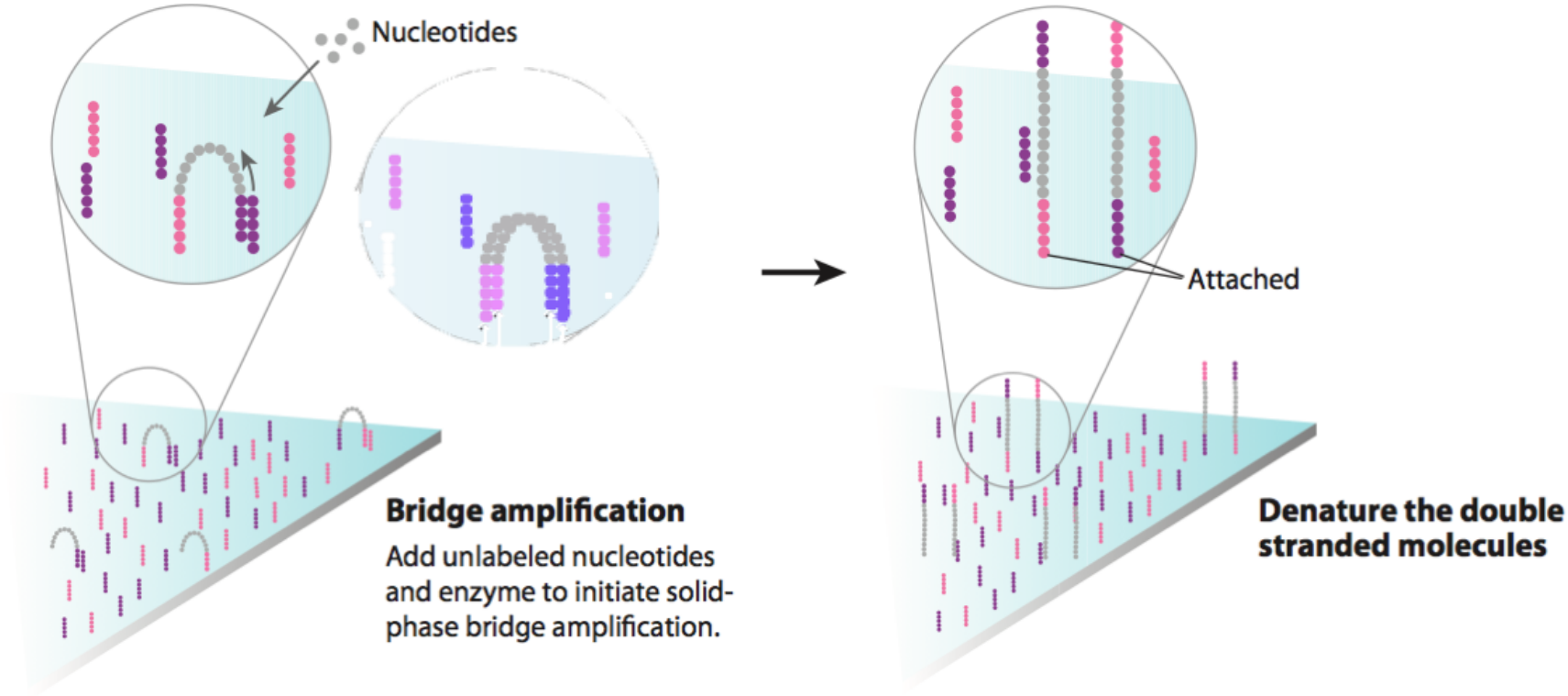




1. **Fragment** DNA (or cDNA) sample
2. **Ligate** different adaptors onto each end
3. **Denature** into single strands
4. **Bind** to flowcell covered with complementary primers

Mardis, Ann Rev Gen Hum, 2008

Illumina Step 2

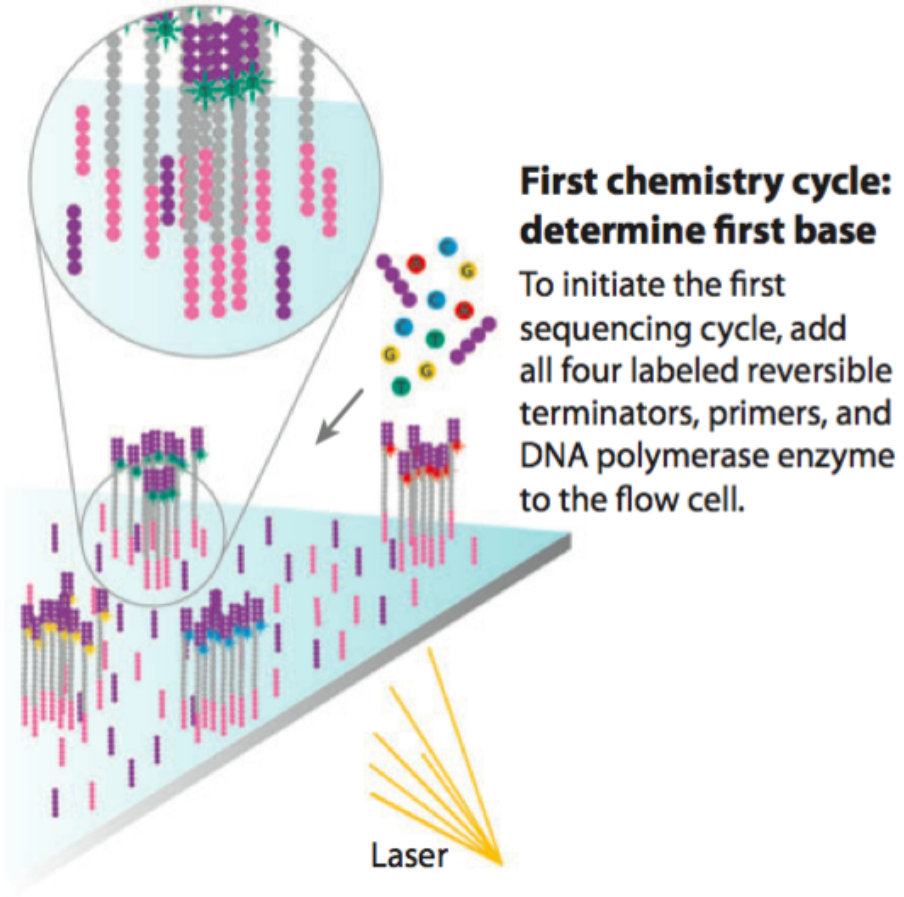


5. Strands **anneal** to primers to form "*bridges*"
6. **Extend** from primer to grow second strand
7. **Free** one terminus of each strand

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

Mardis, Ann Rev Gen Hum, 2008

Illumina Step 3

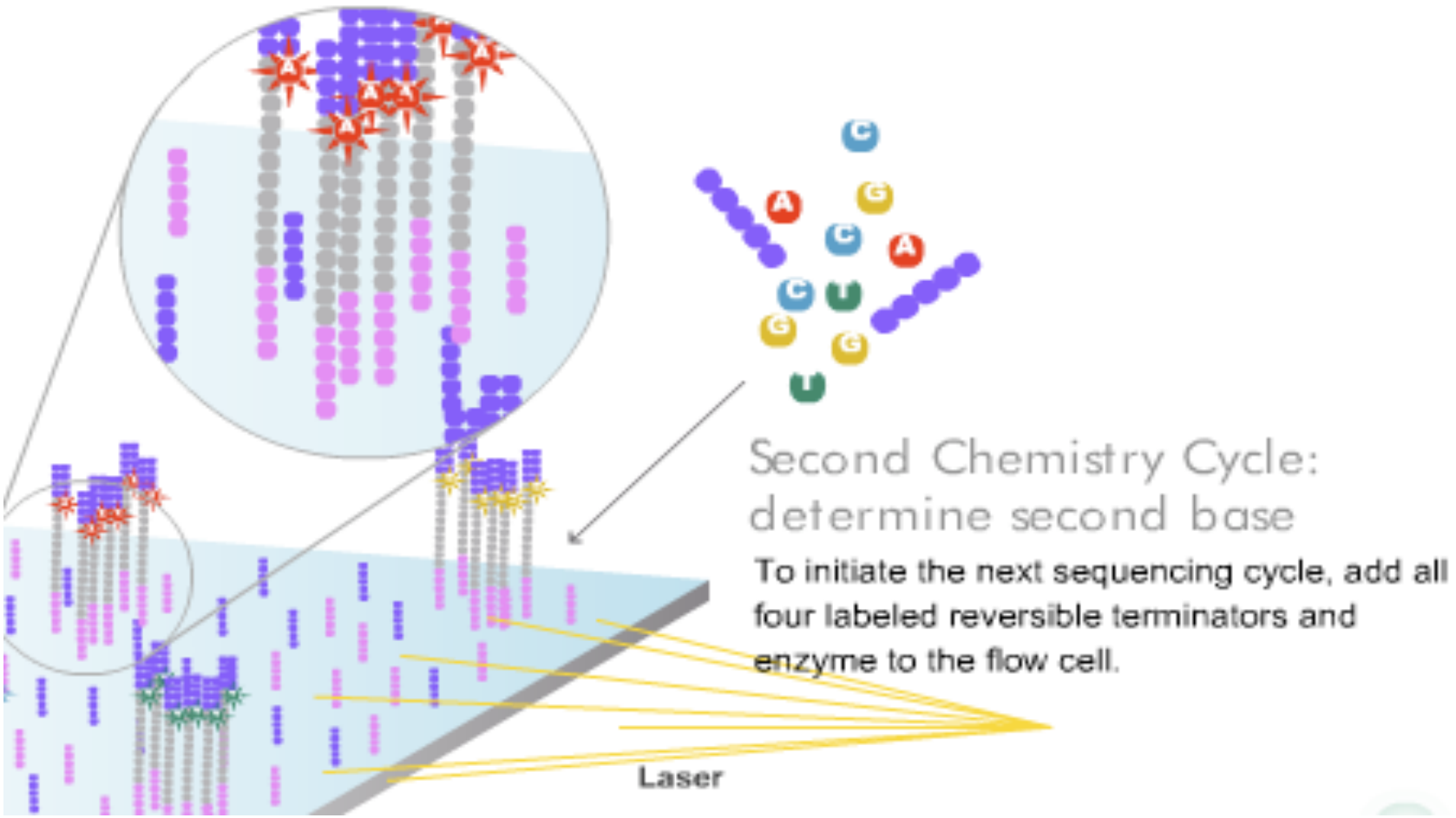


10. **Extend** each strand by one fluorophore-labelled nucleotide followed by blocked terminus
11. **Wash** off unincorporated agents

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

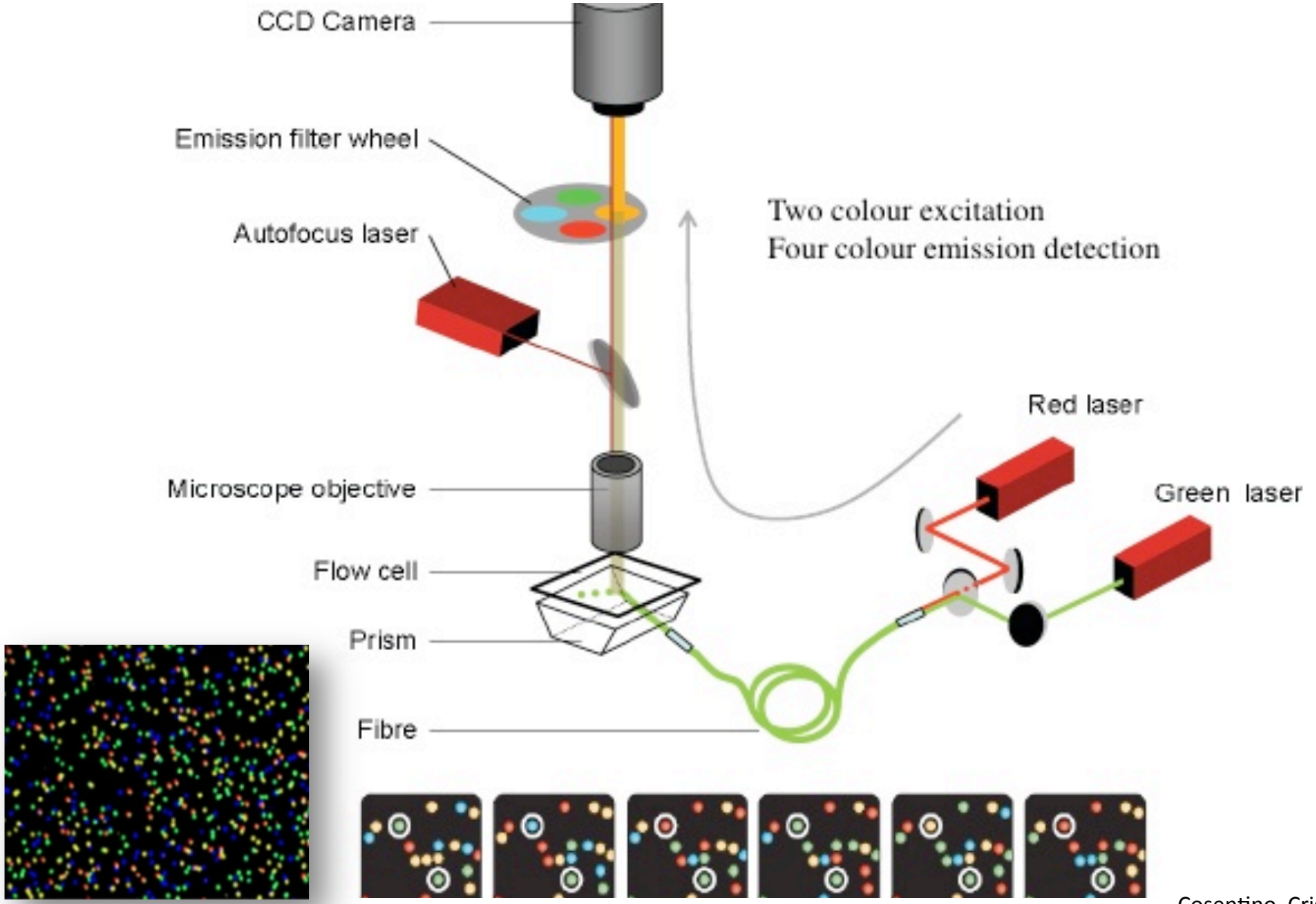
Mardis, Ann Rev Gen Hum, 2008

Illumina Cycling



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

Illumina Optical Path



Cosentino Cristian

Illumina Cycling

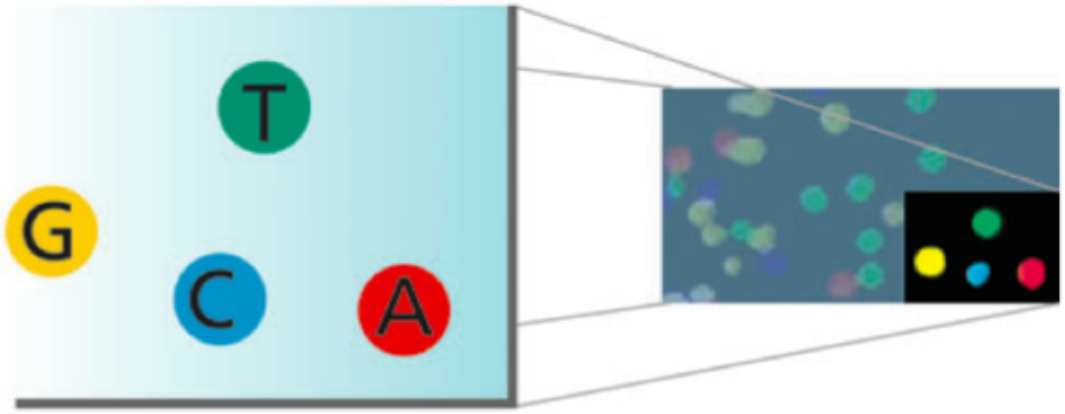
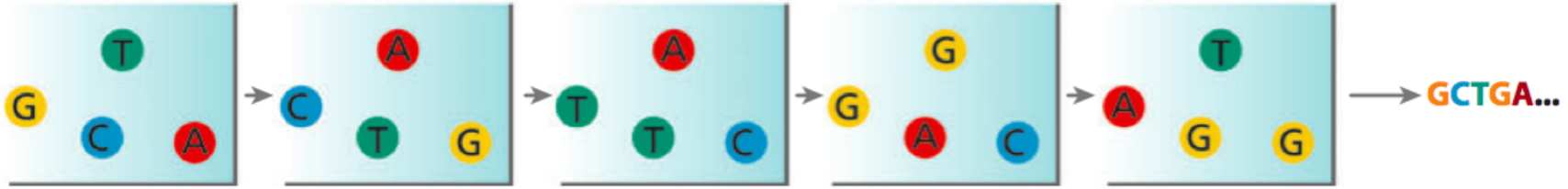


Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

The blocked 3' terminus and the fluorophore from each incorporated base are removed.



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

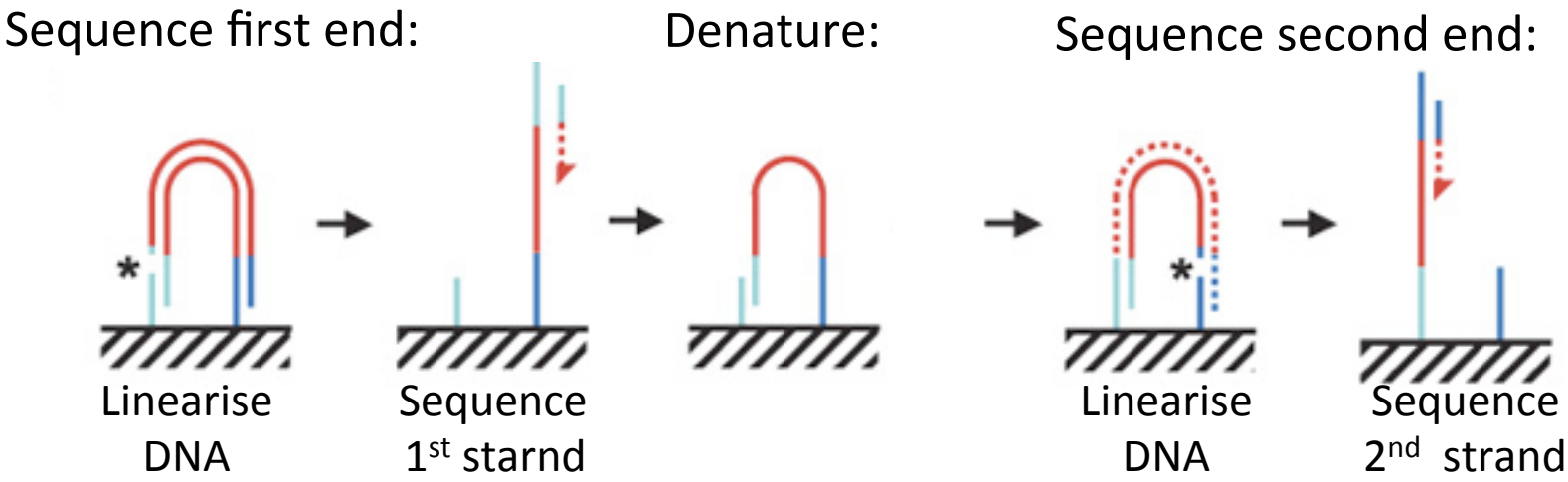
Mardis, Ann Rev Gen Hum, 2008

Paired End Sequencing

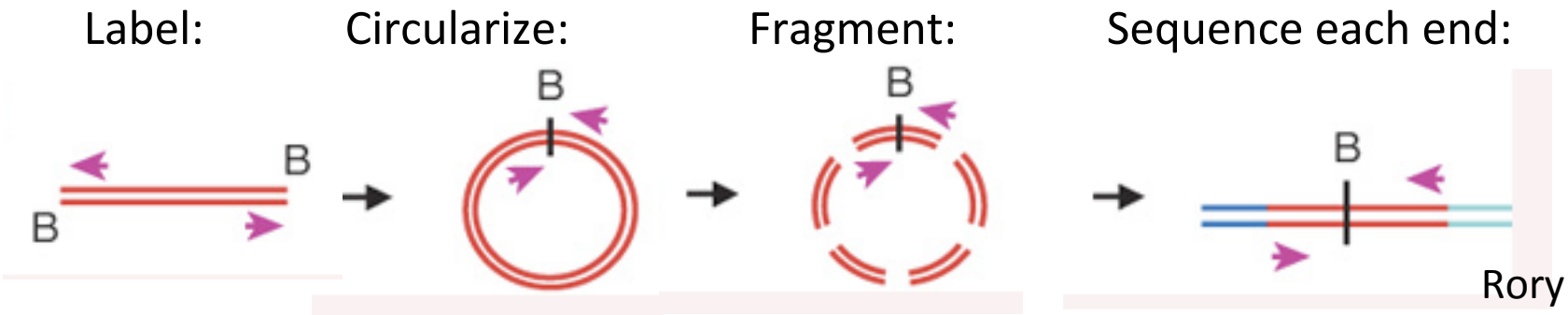
- **Pairs of short reads** from each end of a longer fragment can deliver some of the benefits of longer reads
- Can disambiguate non-uniquely aligned reads
- Can help detect transcript isoforms
- Can detect duplications, inversions, chromosomal rearrangements

Paired End Sequencing

- **Short inserts (300-500bp):**

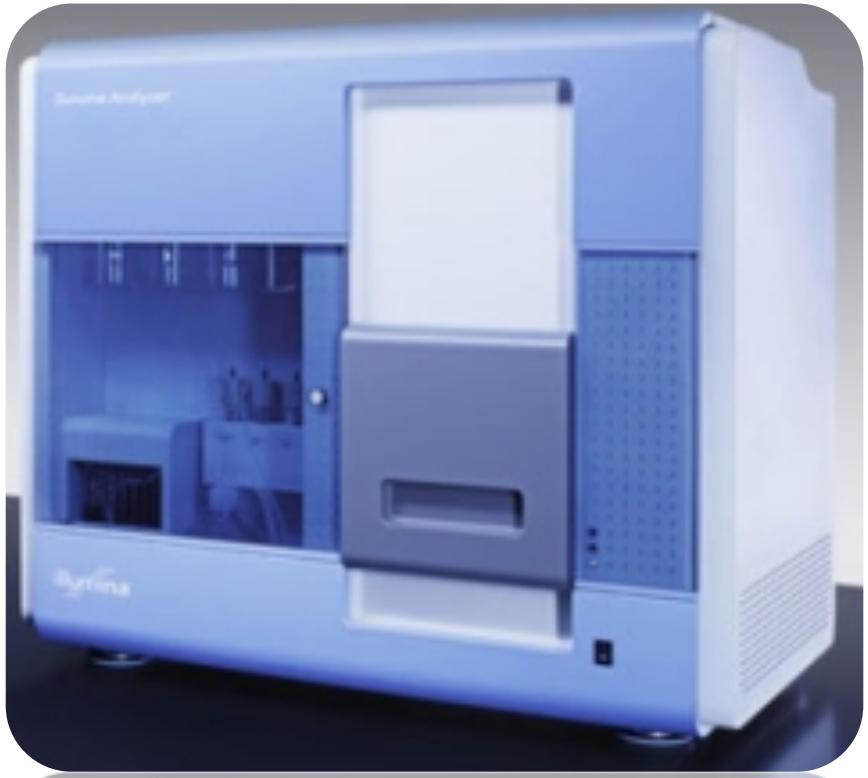


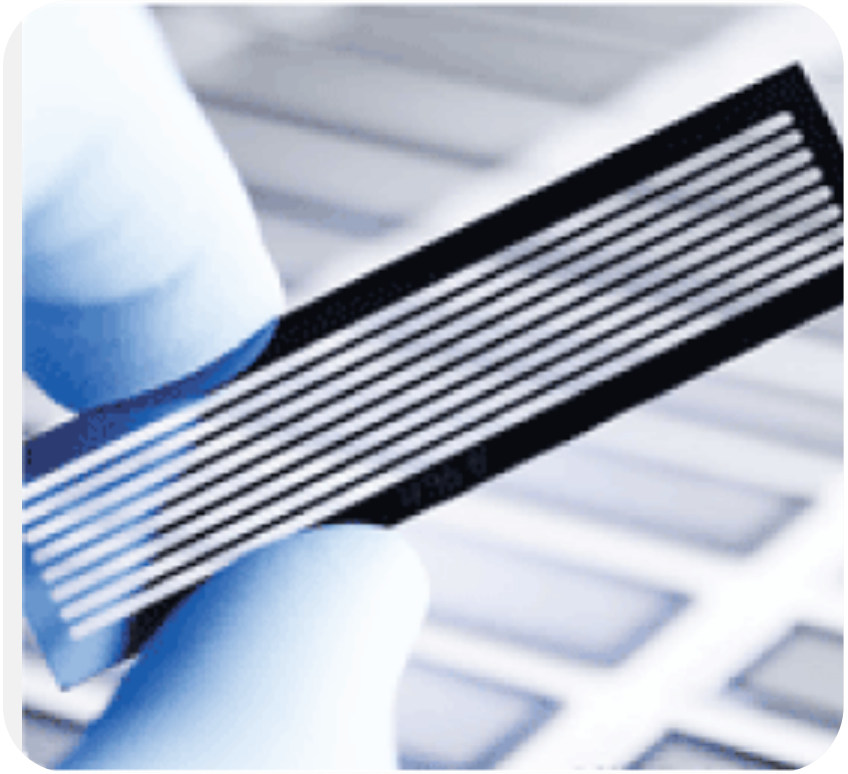
- **Long inserts (e.g. >1Kb) "Mate-pairs":**



Rory Stark

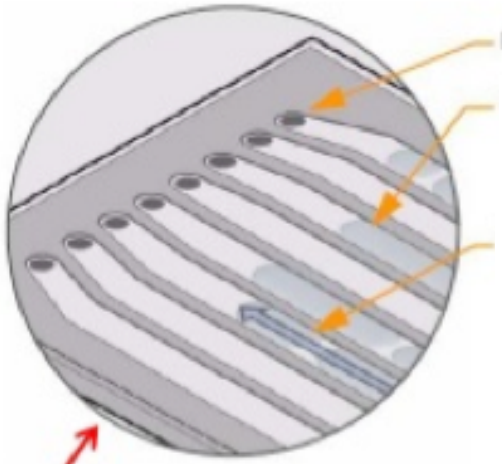
Illumina (Solexa) Genome Analyzer and Flow Cell





First GA produced 1GB run in 2006

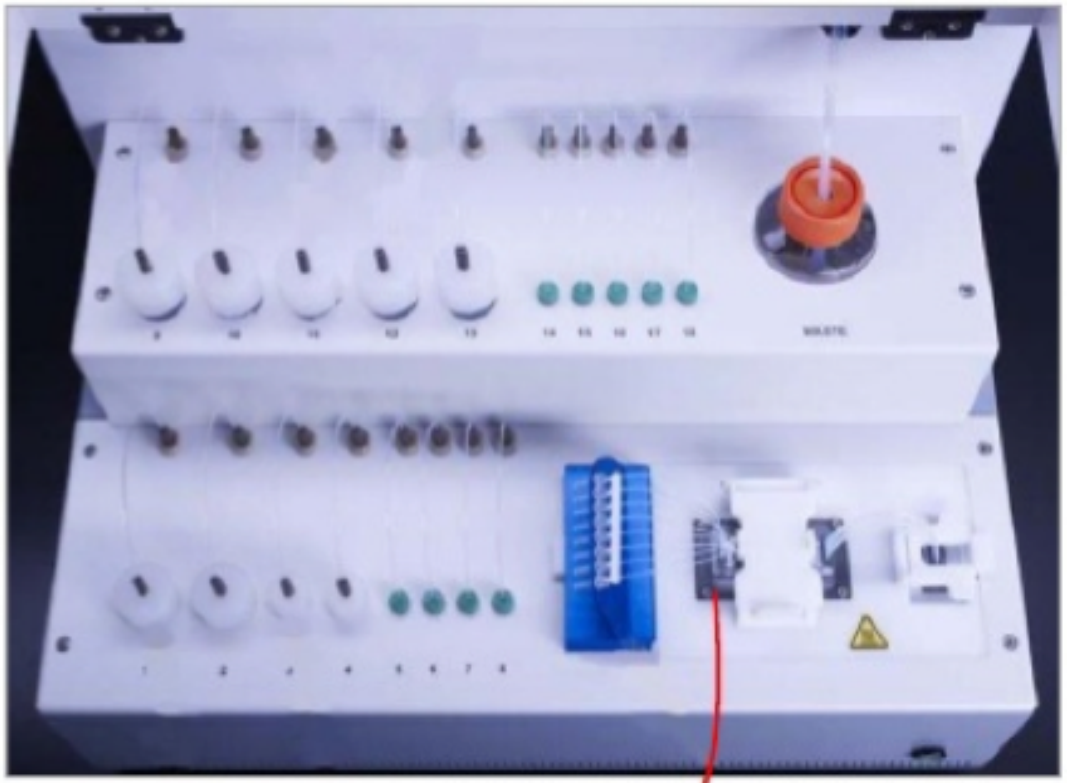
Flow Cell



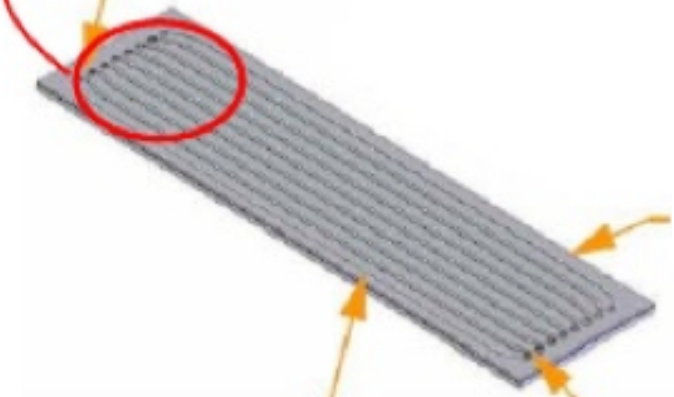
Flow Cell Port

Bubbles

Flow Direction



Outlet Manifold Ports



Lane 8

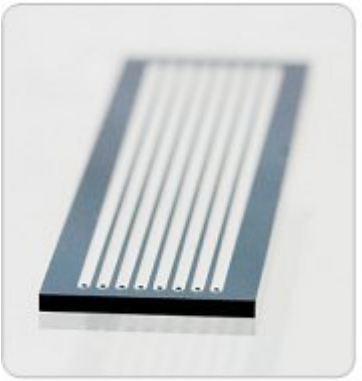
Lane 1

Inlet Manifold Ports

Cosentino Cristian

Flow Cell

- Flow cell divided into **eight lanes**
- Lanes divided into **100 image tiles**
- 4 images (AGCT) per cycle
- 4 bases x 100 tiles x 8 lanes x 45 cycles = 144,000 images
- At 7.3MB/image, this is >1TB per run
- Most of the sequencing run time is spent imaging!
- HiSeq: 2 Flow cells, dual-surface, higher res (larger) images, never write images to disk



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

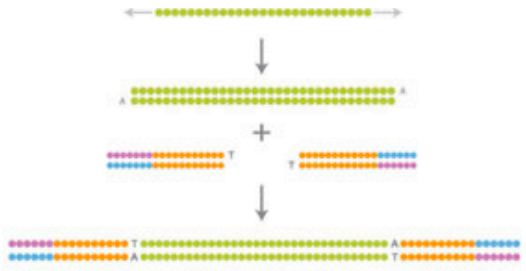
HiSeq 2000

Library Preparation
Genomic DNA; RNA-seq; small RNA; mate pair; chIP-seq
(1-14 days depending on prep type)

Cluster Generation
cBot
(5 hours)

Sequencing by Synthesis
HiSeq 2000
(2-11 days depending on read length)

Data Analysis



Library Preparation

Cluster Generation

Sequencing by Synthesis

HiSeq 2500

	RAPID RUN MODE	HIGH OUTPUT MODE
ChIP-Seq **Transcription Factor** **15M Reads** **1 x 36 bp**	40 Samples 7 Hours	200 Samples 2 Days
mRNA-Seq **>50M Reads** **2 x 50 bp**	24 Samples 16 Hours	120 Samples 5 Days
TruSeq Exome Seq **62 MB Region** **100x Coverage** **2 x 100 bp**	15 Samples 27 Hours	85 Samples 11 Days
Human Whole Genome **>30x Coverage** **2 x 100 bp**	1 Sample 27 Hours	5 Samples 11 Days
***De novo* Sequencing** **1.5 GB Genome** **100x Coverage** **2 x 150 bp**	1 Sample 40 Hours	

Source: Illumina website; August 2013

NetSeq 500 / HiSeq 2500

	NextSeq 500	HiSeq 2500
		
	Learn More »	Learn More »
Data Yield	20 - 120 Gb	10 - 1000 Gb
Maximum Read Length	2 x 150 bp	2 x 125 bp
Reads per Run	130M - 400M	300M - 4B
Run Time	12 - 30 hr	7 hr - 6 days
Price per Sample	Higher	Lower
System Price	Lower	Higher
Samples per Run (RNA-seq)	12 - 36	24 - 396
Samples per Run (ChIP-seq)	8 - 24	20 - 264
Samples per Run (Whole Genome Large)	1	1 - 10

Adapted from source: Illumina website; March 2014

Illumina MiSeq:
Personal Sequencer

Output	Read Number	Read Length
15 Gb	25 M	2x300 bp

Source: Illumina website; March 2014

PART 1:
The Technologies

NGS Technologies

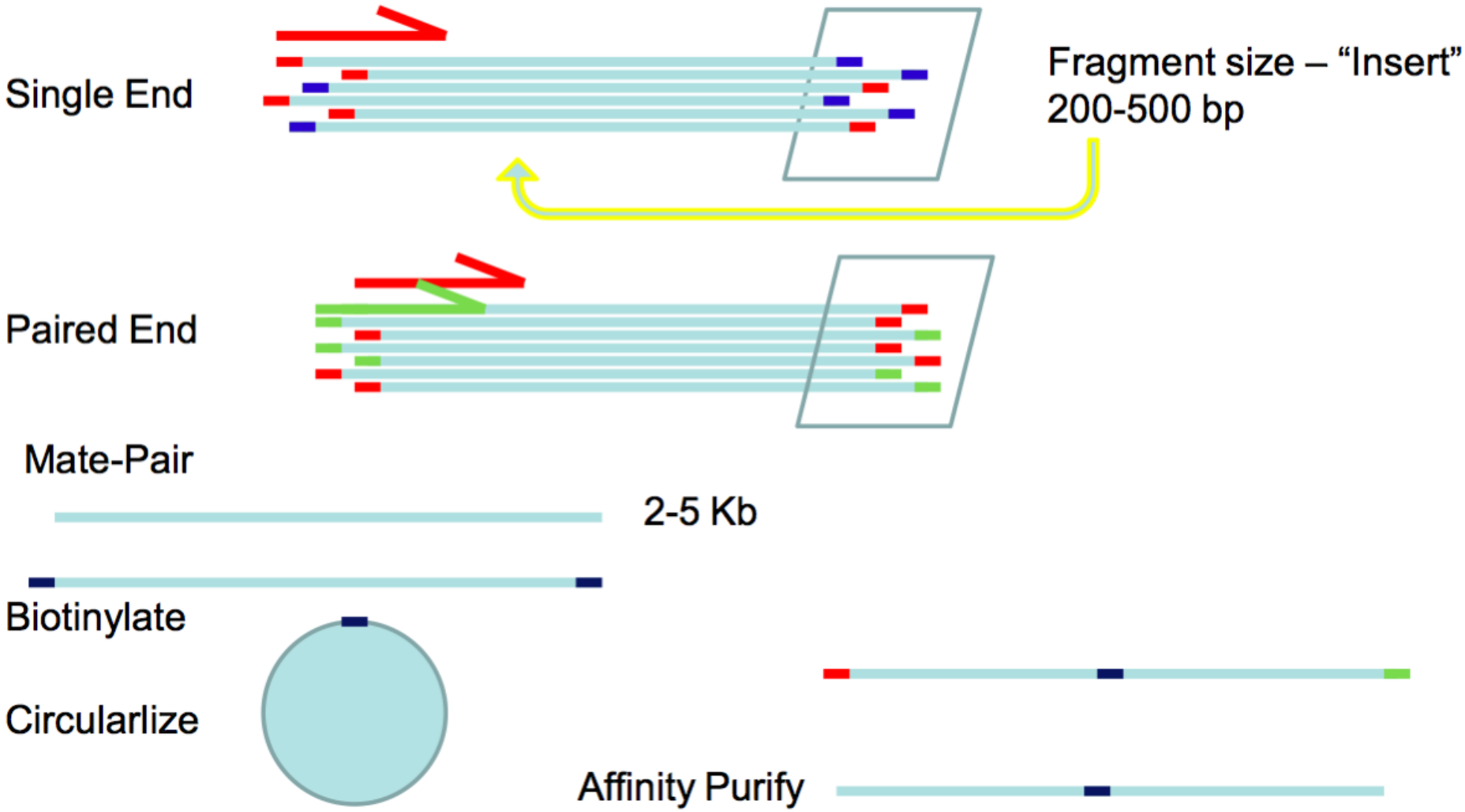
(Clonally Amplified DNAs)

- **Sample Preparation & Multiplexing**

Application & Scope

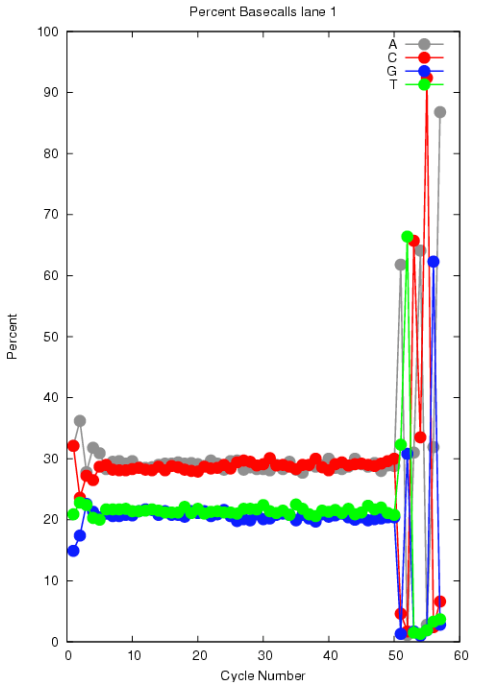
Clinical Application

Library Methods

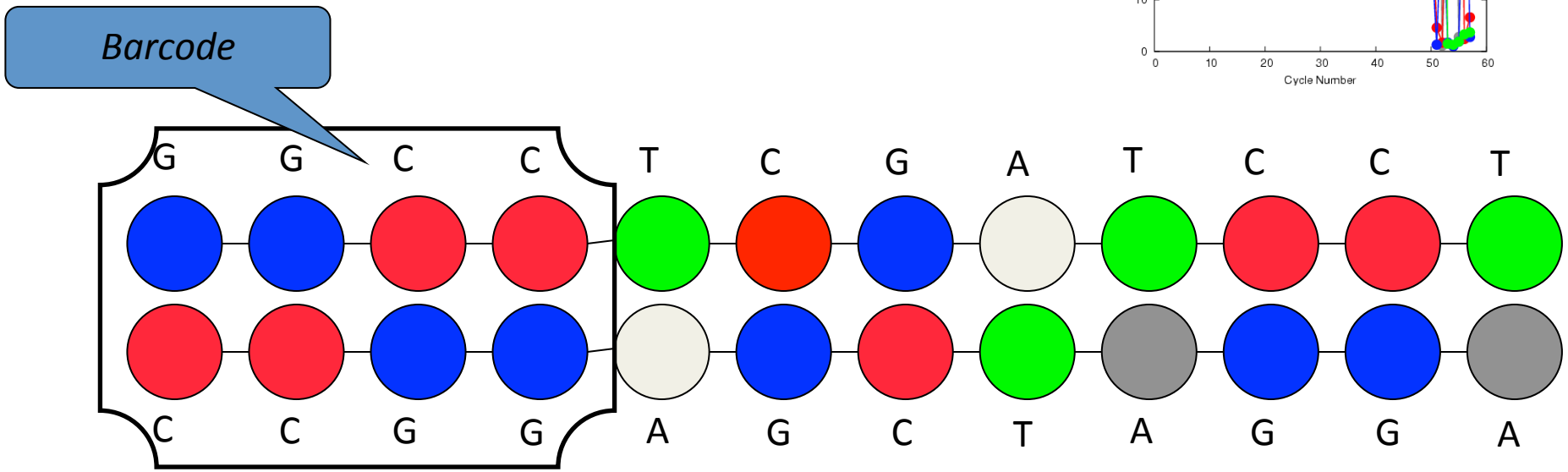


Multiplexing/barcoding

- As throughput increases, more demand for sequencing multiple samples in a single "lane."
- Barcodes are known n-mers ligated to each molecule in a specific sample.



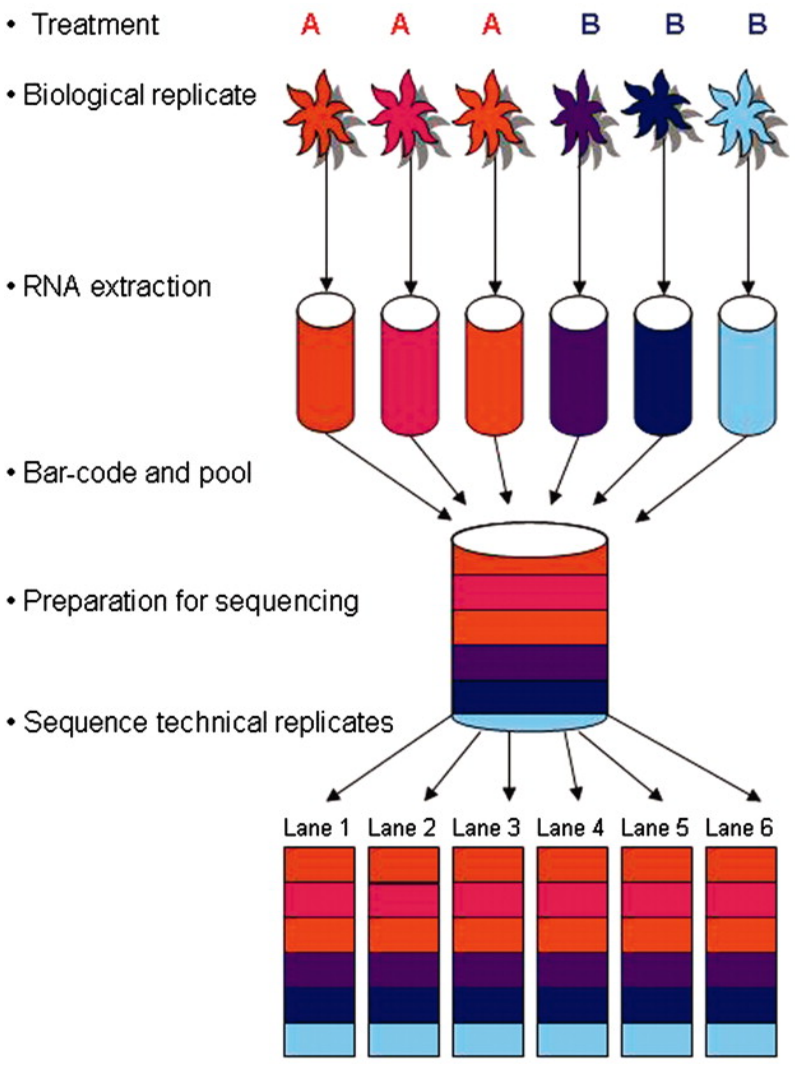
- Deconvolute samples by reading barcodes
- Also paired-end barcoding, with short second (or third) read just of barcode
- Very difficult to mix samples evenly!



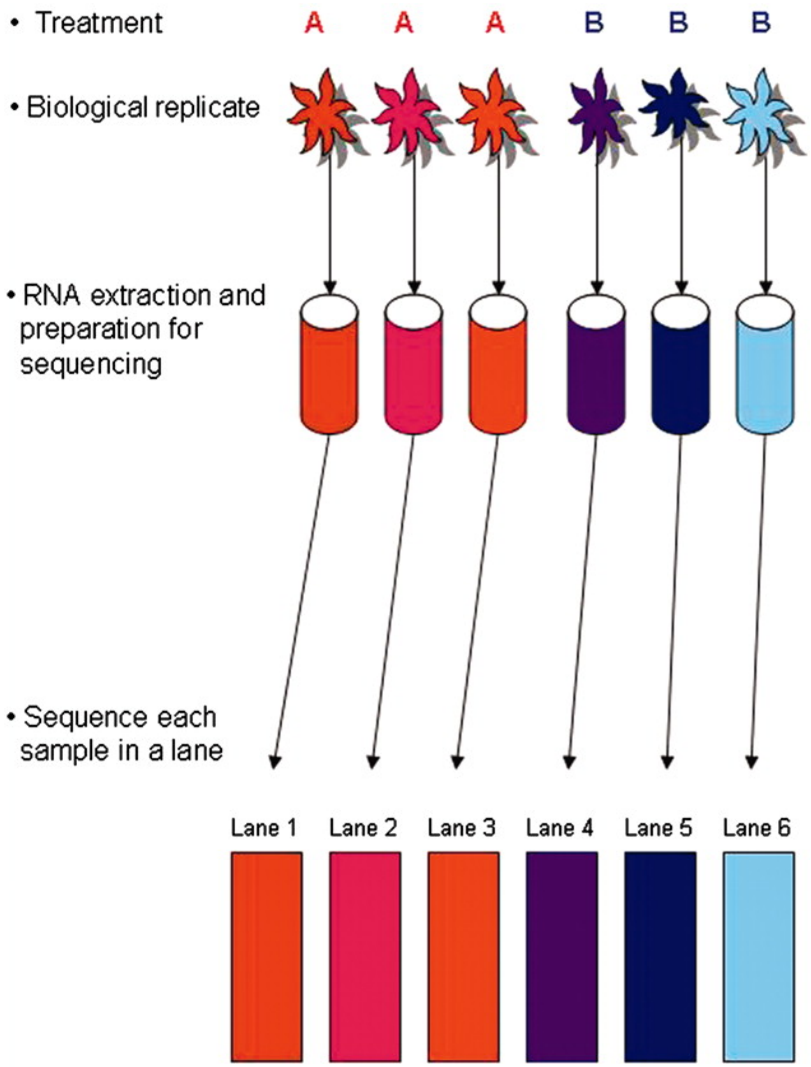
Rory Stark

Multiplexing

Balanced Blocked Design



Confounded Design

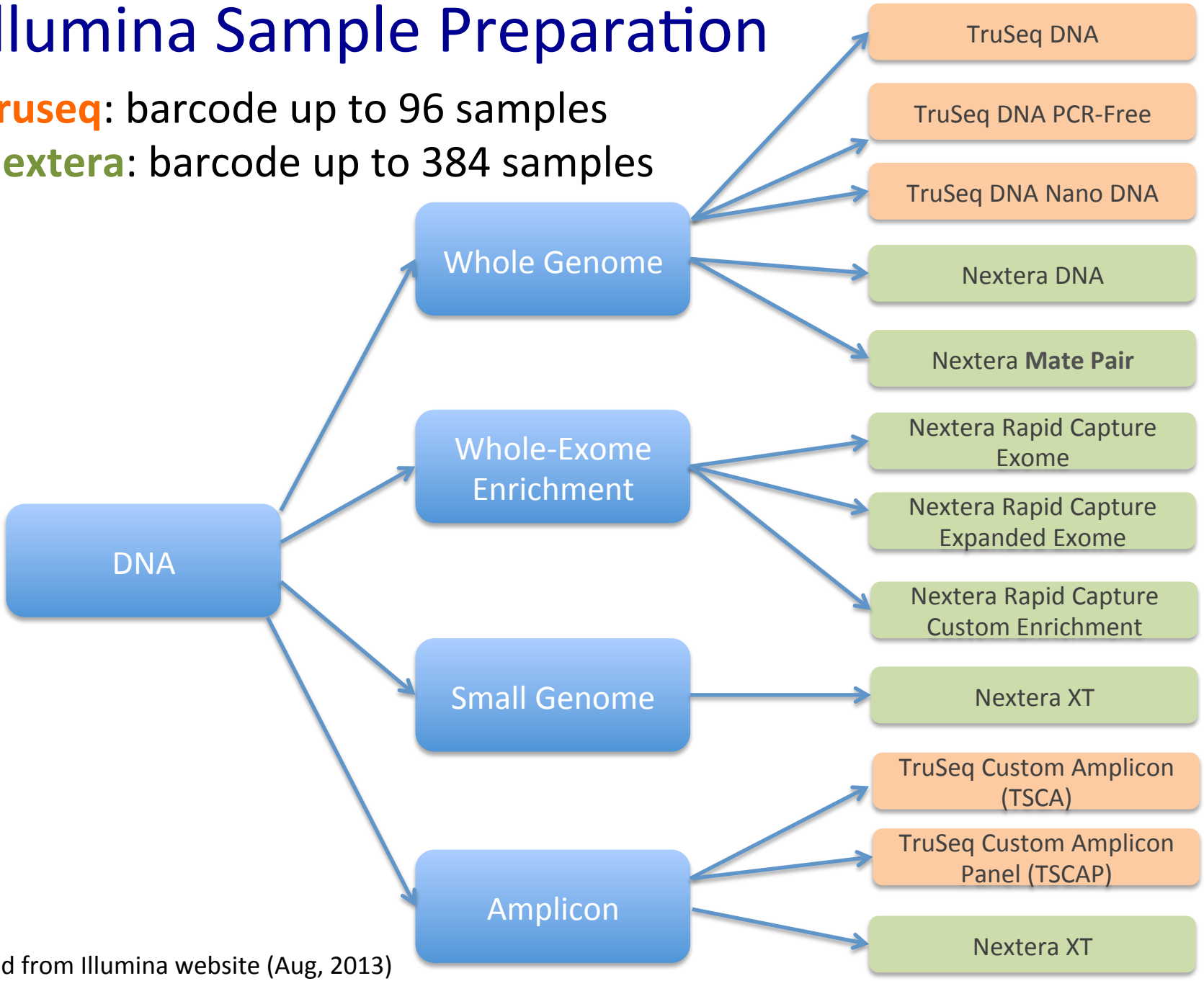


Auer P L , Doerge R W Genetics 2010;185:405-416

Illumina Sample Preparation

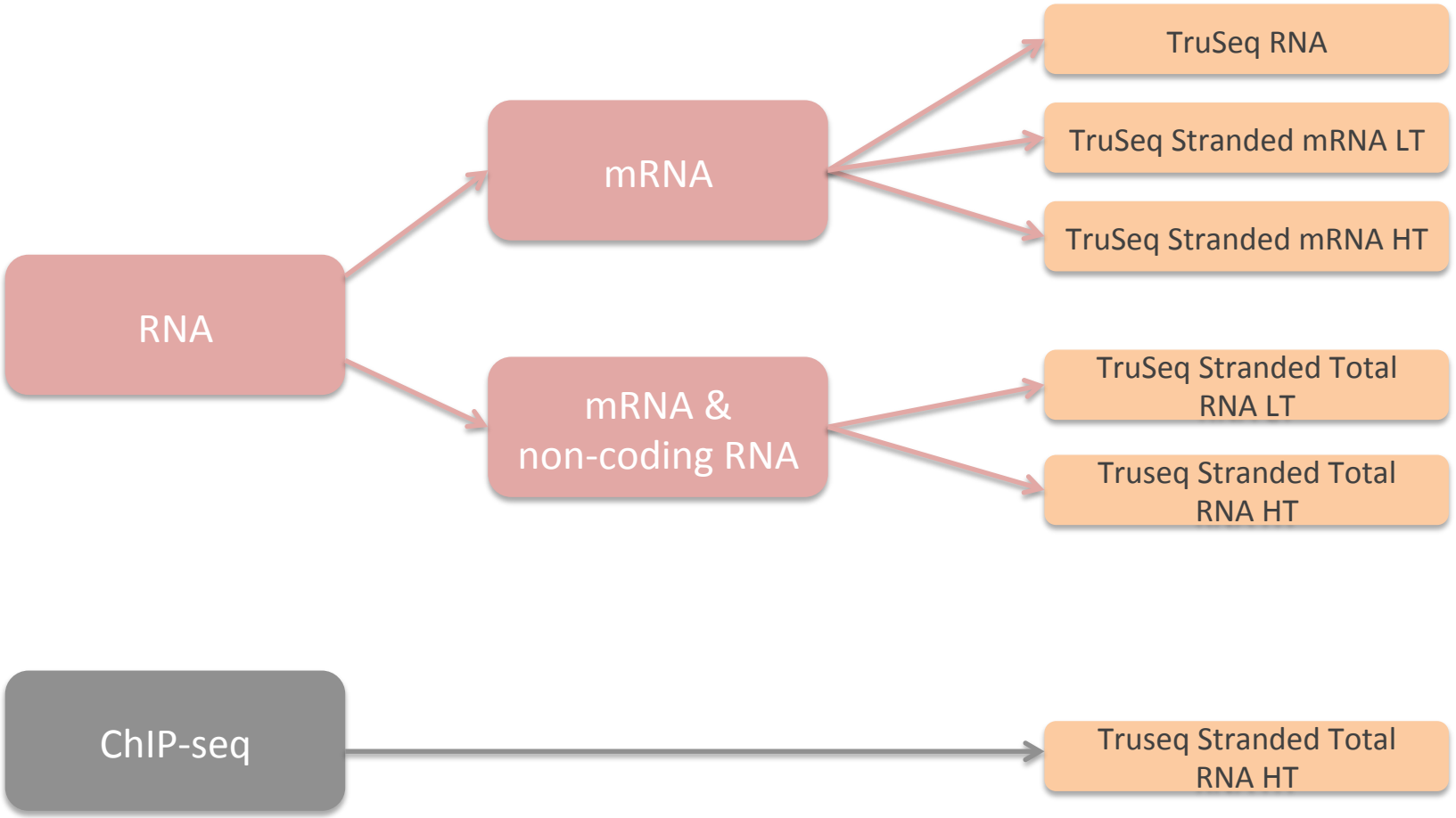
Truseq: barcode up to 96 samples

Nextera: barcode up to 384 samples



Adapted from Illumina website (Aug, 2013)

Illumina Sample Preparation

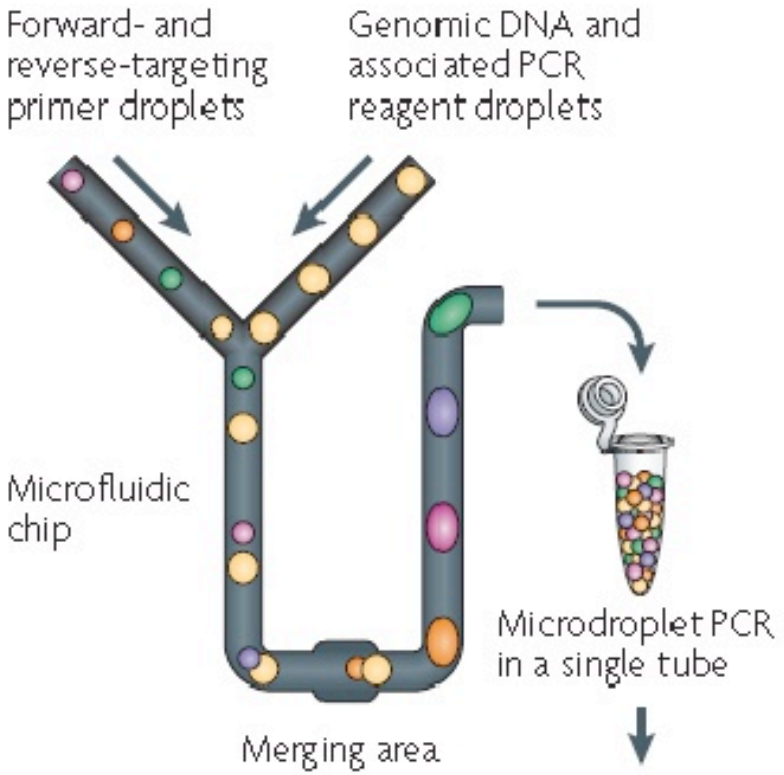


Adapted from Illumina website (Aug, 2013)

HTP Sample Preparation

RainDance

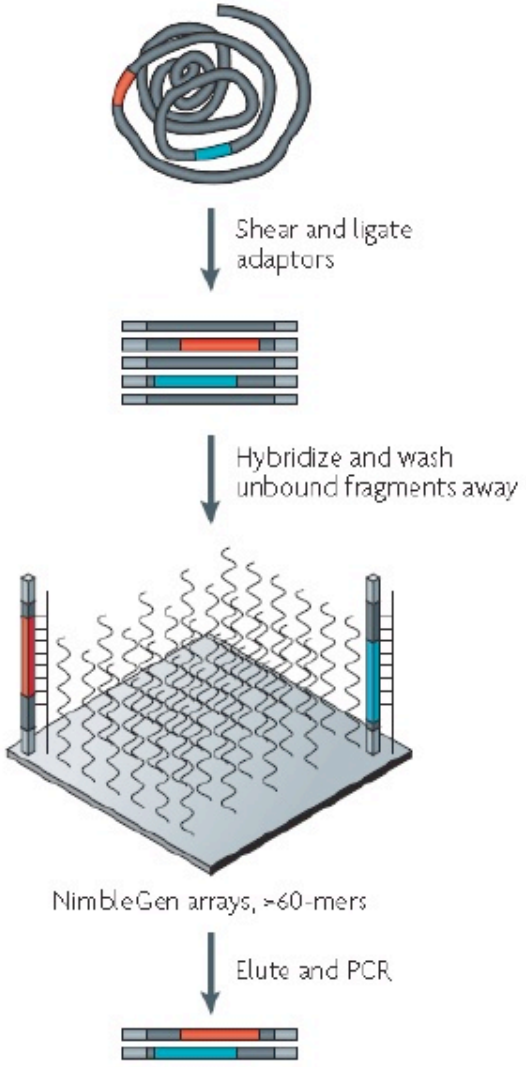
Microdroplet PCR



Reported 84% of capture efficiency

Roche Nimblegen

Salid-phase capture with custom-designed oligonucleotide microarray

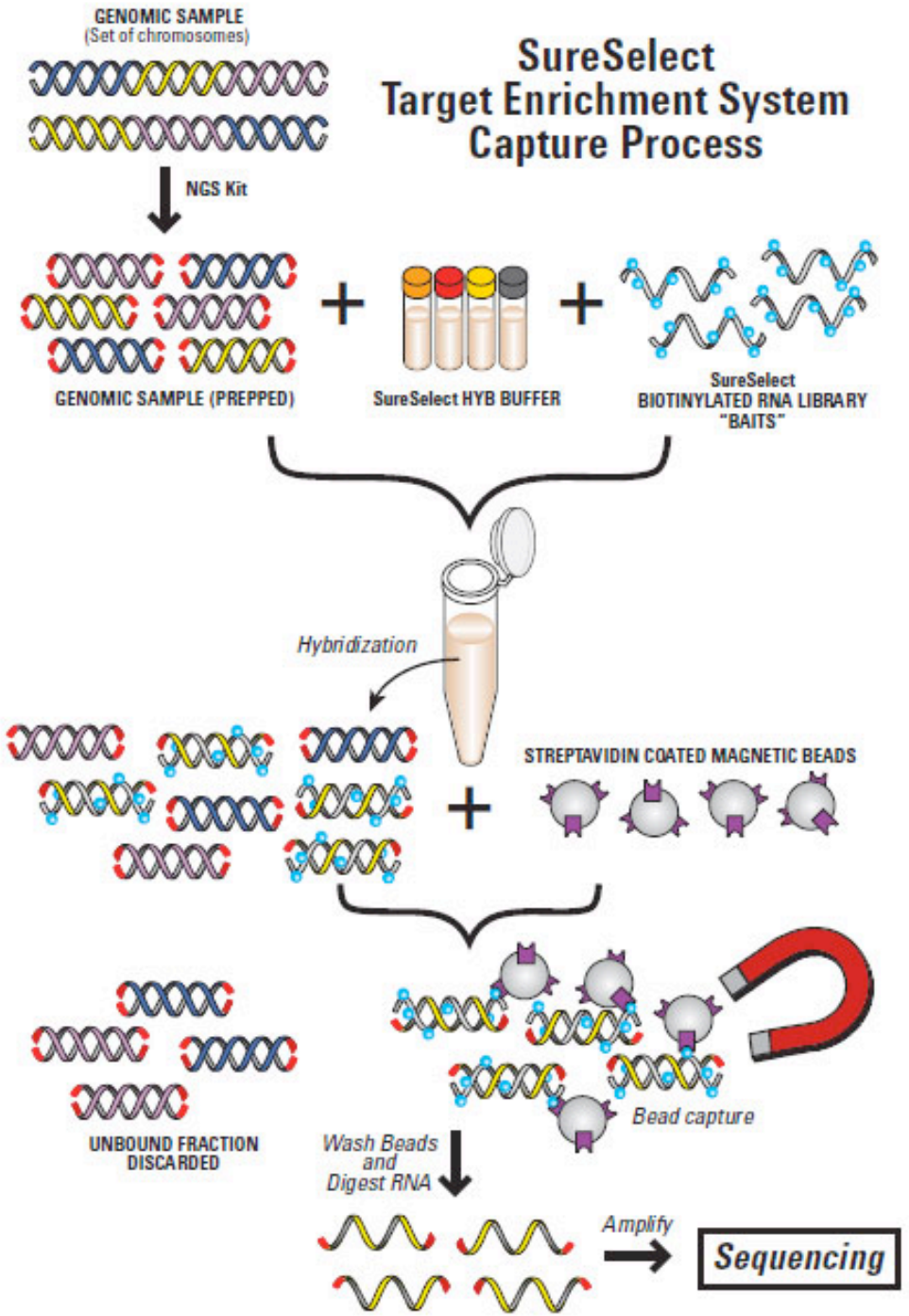


Reported 65-90% of capture efficiency

Nature Methods, 2010 7: 111-118

HTP Sample Preparation

Agilent SureSelect
Solution-phase capture with
streptavidin-coated magnetic beads



Basic NGS Parameters

- **Library size / sequencing depth:** the number of reads obtained for a given sample (e.g. 20 million)
- **Read length:** bp length of each read in the library (e.g. 150 bp)
- **Single end vs paired end:** single end sequencing only sequences one end of each DNA fragment, while paired end sequences both ends.
- **Barcoding:** how to allocate samples to lanes/runs.

PART 1:
The Technologies

NGS Technologies

(Clonally Amplified DNAs)

Sample Preparation & Multiplexing

Application & Scope

Clinical Application

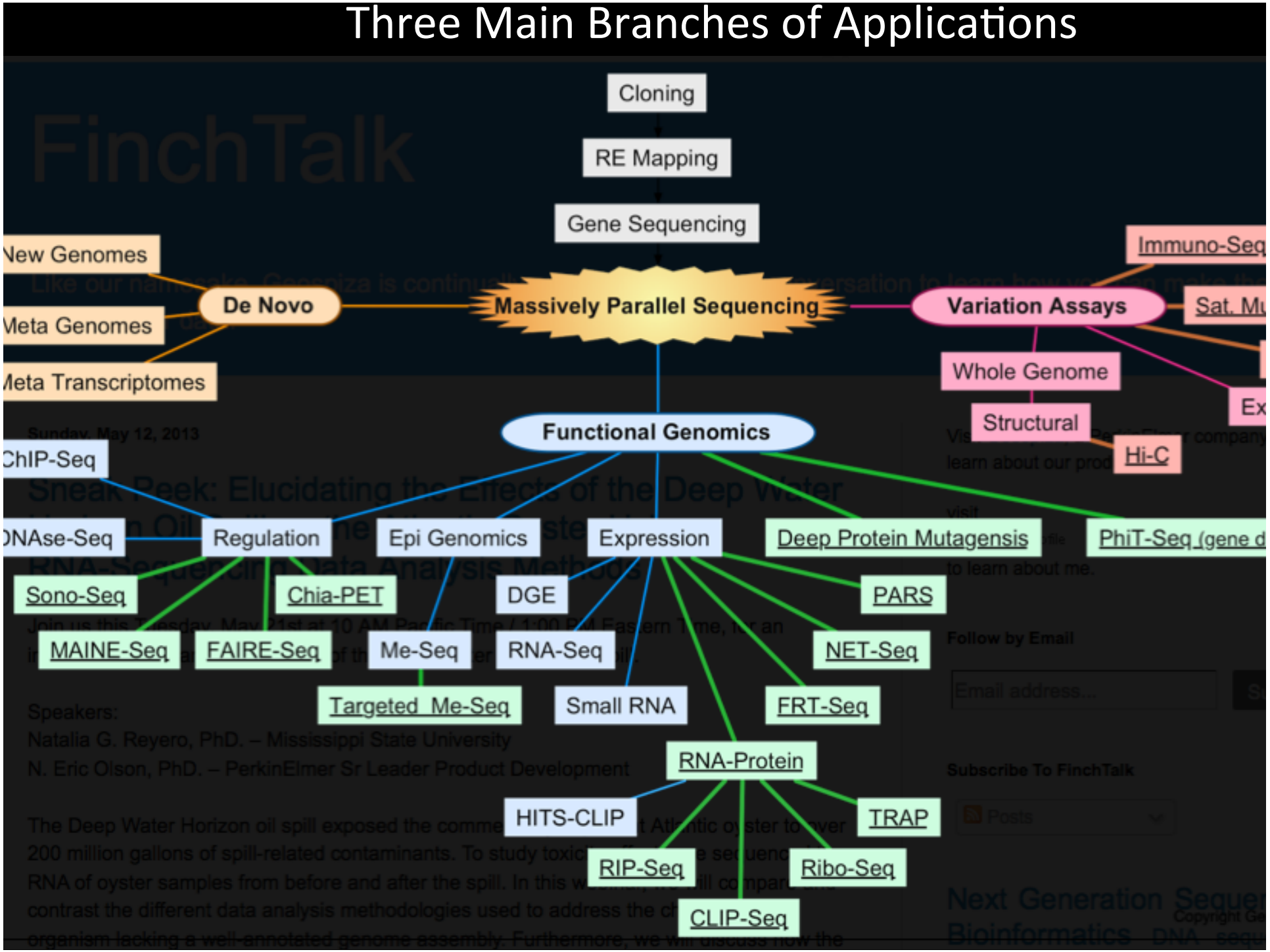
Current Main Illumina Applications

Application	Source
De novo gDNA sequencing	gDNA
Whole-genome sequencing	gDNA
Target sequencing	Target enriched DNA sequences
mRNA-seq	Total RNA
Small RNA-seq	Total RNA
CHiP-seq	Chip-DNA fragments

Method	Sequencing to determine:	reference	'Subway' route as defined in Figure 3
DNA-Seq	A genome sequence	57	Comparison, 'anatomic' (isolation by anatomic site), flow cytometery, DNA extraction, mechanical shearing, adaptor ligation, PCR and sequencing
Targeted DNA-Seq	A subset of a genome (for example, an exome)	20	Comparison, cell culture, DNA extraction, mechanical shearing, adaptor ligation, PCR, hybridization capture, PCR and sequencing
Methyl-Seq	Sites of DNA methylation, genome-wide	34	Perturbation, genetic manipulation, cell culture, DNA extraction, mechanical shearing, adaptor ligation, bisulfite conversion, PCR and sequencing
Targeted methyl-Seq	DNA methylation in a subset of the genome	129	Comparison, cell culture, DNA extraction, bisulfite conversion, molecular inversion probe capture, circularization, PCR and sequencing
DNase-Seq, Sono-Seq and FAIRE-Seq	Active regulatory chromatin (that is, nucleosome-depleted)	113	Perturbation, cell culture, nucleus extraction, DNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
MAINE-Seq	Histone-bound DNA (nucleosome positioning)	130	Comparison, cell culture, MNase I digestion, DNA extraction, adaptor ligation, PCR and sequencing
ChIP-Seq	Protein-DNA interactions (using chromatin immunoprecipitation)	131	Comparison, 'anatomic', cell culture, cross-linking, mechanical shearing, immunoprecipitation, DNA extraction, adaptor ligation, PCR and sequencing
RIP-Seq, CLIP-Seq, HITS-CLIP	Protein-RNA interactions	46	Variation, cross-linking, 'anatomic', RNase digestion, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, PCR and sequencing
RNA-Seq	RNA (that is, the transcriptome)	39	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
FRT-Seq	Amplification-free, strand-specific transcriptome sequencing	119	Comparison, 'anatomic', RNA extraction, poly(A) selection, chemical fragmentation, adaptor ligation, reverse transcription and sequencing
NET-Seq	Nascent transcription	41	Perturbation, genetic manipulation, cell culture, immunoprecipitation, RNA extraction, adaptor ligation, reverse transcription, circularization, PCR and sequencing
Hi-C	Three-dimensional genome structure	71	Comparison, cell culture, cross-linking, proximity ligation, mechanical shearing, affinity purification, adaptor ligation, PCR and sequencing
Chia-PET	Long-range interactions mediated by a protein	73	Perturbation, cell culture, cross-linking, mechanical shearing, immunoprecipitation, proximity ligation, affinity purification, adaptor ligation, PCR and sequencing
Ribo-Seq	Ribosome-protected mRNA fragments (that is, active translation)	48	Comparison, cell culture, RNase digestion, ribosome purification, RNA extraction, adaptor ligation, reverse transcription, rRNA depletion, circularization, PCR and sequencing
TRAP	Genetically targeted purification of polysomal mRNAs	132	Comparison, genetic manipulation, 'anatomic', cross-linking, affinity purification, RNA extraction, poly(A) selection, reverse transcription, second-strand synthesis, adaptor ligation, PCR and sequencing
PARS	Parallel analysis of RNA structure	42	Comparison, cell culture, RNA extraction, poly(A) selection, RNase digestion, chemical fragmentation, adaptor ligation, reverse transcription, PCR and sequencing
Synthetic saturation mutagenesis	Functional consequences of genetic variation	93	Variation, genetic manipulation, barcoding, RNA extraction, reverse transcription, PCR and sequencing
Immuno-Seq	The B-cell and T-cell repertoires	86	Perturbation, 'anatomic', DNA extraction, PCR and sequencing
Deep protein mutagenesis	Protein binding activity of synthetic peptide libraries or variants	95	Variation, genetic manipulation, phage display, *in vitro* competitive binding, DNA extraction, PCR and sequencing
PhIT-Seq	Relative fitness of cells containing disruptive insertions in diverse genes	92	Variation, genetic manipulation, cell culture, competitive growth, linear amplification, adaptor ligation, PCR and sequencing

Shendure & Aidan (2012) Nature Biotech

Three Main Branches of Applications



How Deeply Should You Sequence?

- **Depends on the application:** e.g. some parts of the genome will be more tricky to sequence, so full coverage for genome assembly may require very deep sequencing.

- **RNA-seq differential expression:** very roughly (for human and mouse) - at least 10 million reads per sample is required, and 30 million reads per sample should suit most applications. Can sequence deeper if low expressed genes are particularly important.

- **If in doubt:** you can run a pilot experiment and see how much coverage and saturation you get, and potentially sequence further after that.

Sequencing depth and coverage: key considerations in genomic analyses

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger and Chris P. Ponting

Abstract | Sequencing technologies have placed a wide range of genomic analyses within the capabilities of many laboratories. However, sequencing costs often set limits to the amount of sequences that can be generated and, consequently, the biological outcomes that can be achieved from an experimental design. In this Review, we discuss the issue of sequencing depth in the design of next-generation sequencing experiments. We review current guidelines and precedents on the issue of coverage, as well as their underlying considerations, for four major study designs, which include *de novo* genome sequencing, genome resequencing, transcriptome sequencing and genomic location analyses (for example, chromatin immunoprecipitation followed by sequencing (ChIP–seq) and chromosome conformation capture (3C)).

Sequencing Depth depends on Technology/Application

- ***De novo* genome sequencing**

- **DNA resequencing:**
  - whole-genome sequencing (WGS)
  - whole-exome sequencing (WES)
  - *SNV and indel detection*
  - *CNV detection*

- **Transcriptome sequencing**
  - *Differential expression analyses.*
  - *Analyses of alternative splicing*
  - *Transcript Discovery and gene fusions*

- **Location-based methods:**
  - DNA–protein interactions - *ChIP-seq;* ChIP-exo
  - RNA–protein interactions – CLIP; CLIP-seq; iCLIP; PAR-CLIP
  - RNA–DNA interactions - CHART; CHiRP
  - DNA–DNA interactions – 3C; 4C; 5C; ChIA–PET
  - DNA-state:
    - chromatin openess - DNaseI-seq
    - DNA-methylation - MeDIP–seq; CAP–seq

Sequencing Depth depends on Technology/Application

Table 2 | **Representative read counts for location-based approaches**

Techniques	Read counts in representative studies	Refs
DNaseI-seq and FAIRE–seq	20–50 million	79
CLIP–seq	7.5 million; 36 million	89, 90
iCLIP and PAR–CLIP	8 million; 14 million	105, 106
CHiRP and CHART	26 million	72
4C	1–2 million	92
ChIA–PET	20 million	107
5C	25 million	108
Hi-C	>100 million	94
MeDIP–seq	60 million	109
CAP–seq	>20 million	110
ChIP–seq	>10 million per sample (point source); >20 million per sample (broad source)	79

What Read Length?

- **The read length should match your sample fragment length:** you want it to be around the size of and just a bit shorter than your fragments.

- **Recommended for Illumina:**
  sRNAseq, ribosomal profiling = 50 bp read length
  mRNAseq (fragment size ~80-200bp) = 150 bp read length longer fragments = 250 bp

- **For longer reads / niche applications:** other sequencers also exist. PacBio has an impressive read length (often >10 kb)

Single End vs. Paired End

- **Single end:** a bit cheaper, less sequencing required. Suitable for most general purposes, such as differential expression analysis.

- **Paired end:** contains more length and positional information about the sequenced fragment - can tell where it starts and stops.

- **Applications for which you need paired end sequencing:**
  - splice junctions
  - rearrangements: indels and inversions

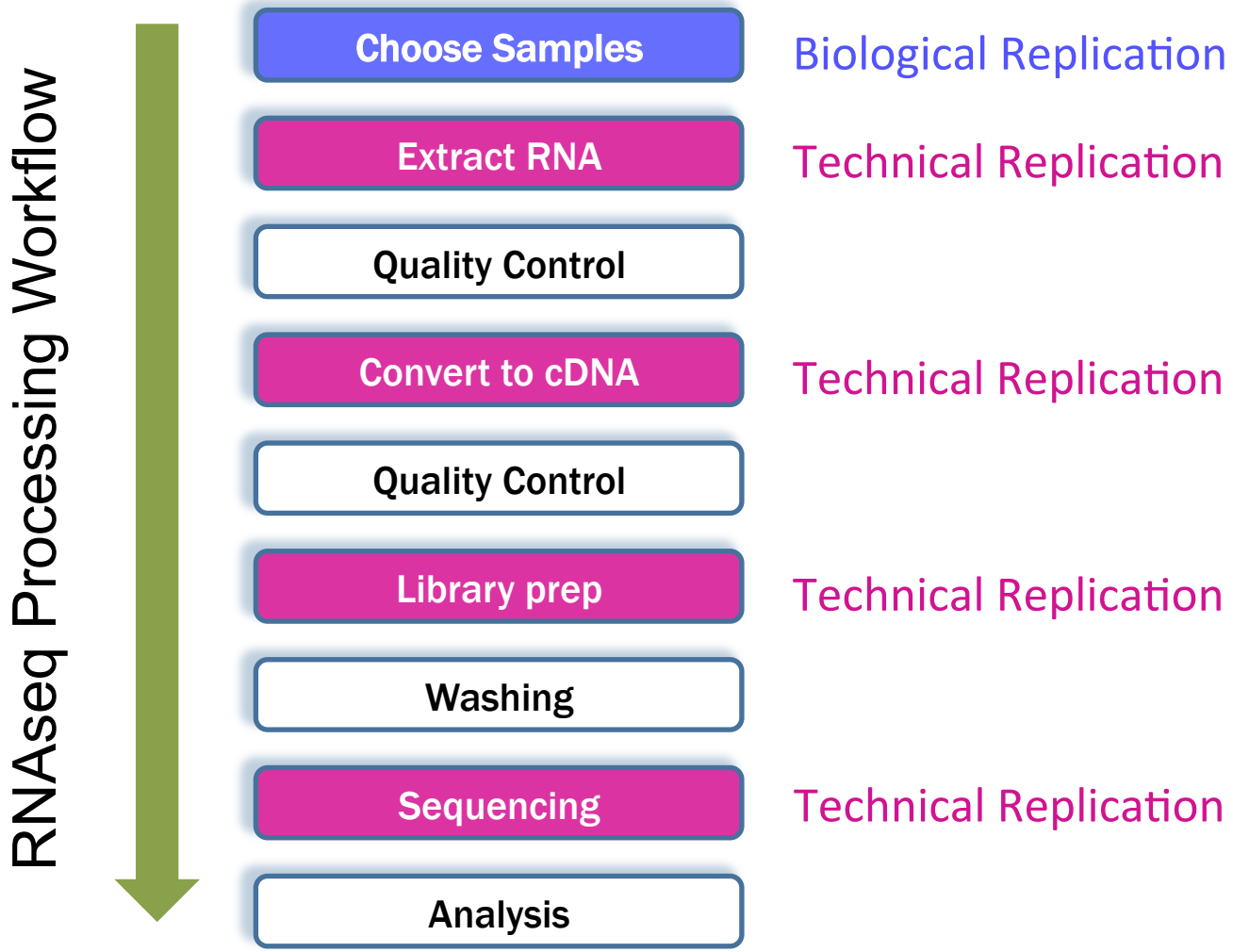
Sequencing Depth vs. Replicates

- **Replicates are good!** Even if you don't sequence any more deeply.

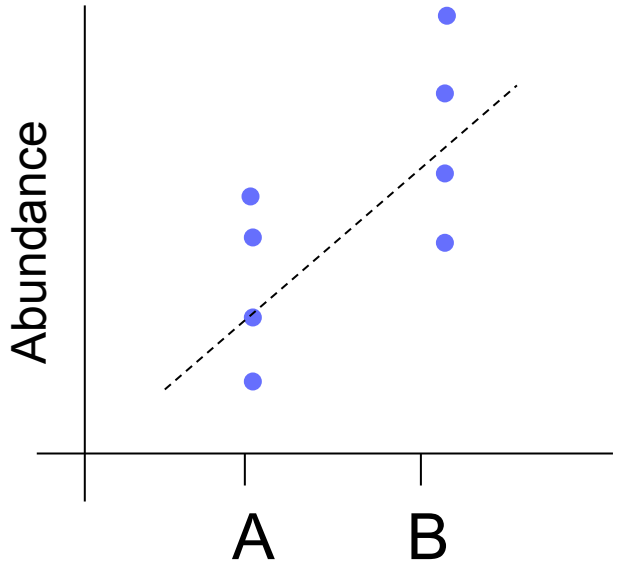
- A study looking at RNAseq (Liu et al. 2013) found that adding more replicates at 10 million reads library size was a more cost- efficient way of improving the power of the experiment, than adding more sequencing coverage. This was true for 3-7 replicates.

- So, if you don't have money for a lot of sequencing, you can still multiplex a number of biological replicates!

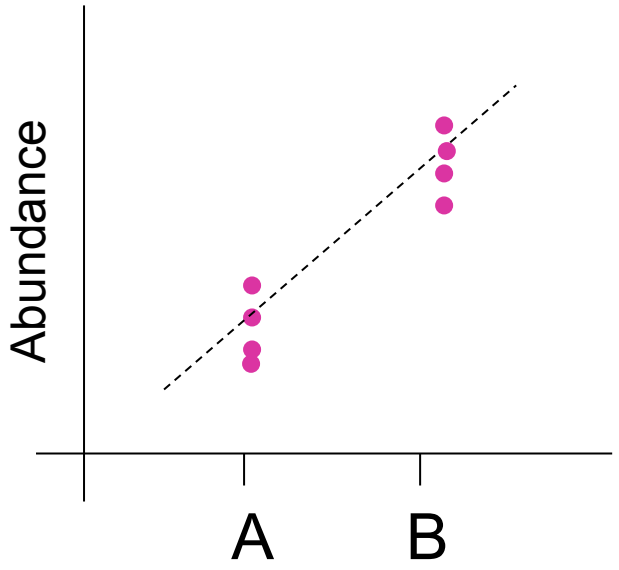
Sample Prep:
Biological or technical replicates?



Biological or technical replicates?



Biological Replicates



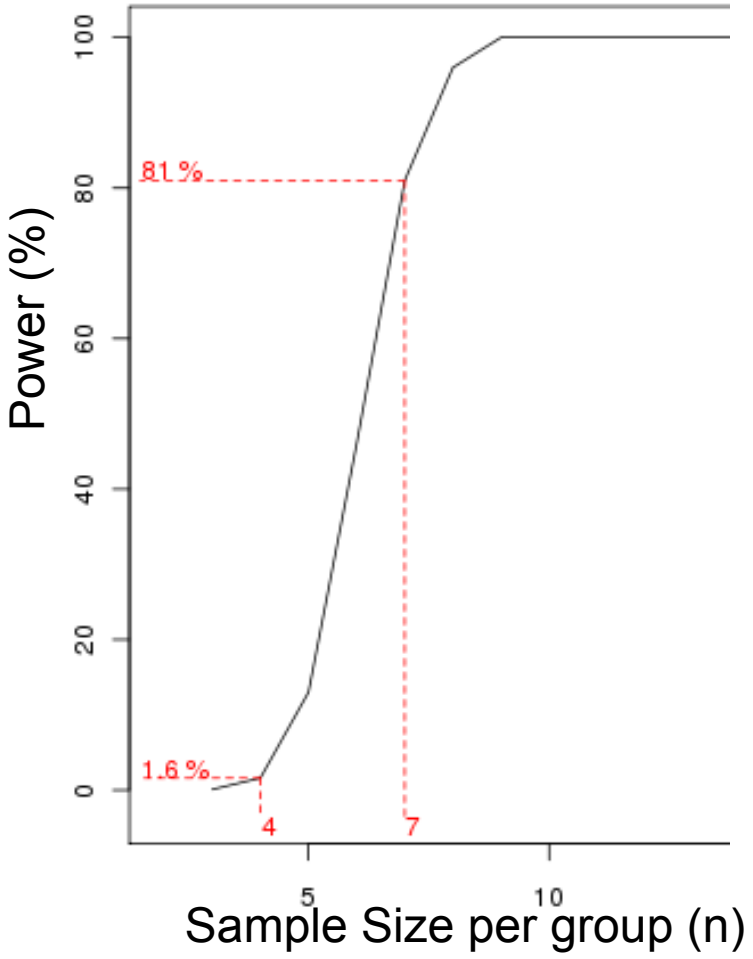
Technical Replicates

Don’t do technical replicates unless you are planning a technical rather than a biological study.

Adequately Powered

*The power of an experiment is the **probability** that it can detect an effect, if it is present*

- **Too <u>few</u> samples:**
  - may <u>lack power</u> and miss a scientifically important effect.
  - Also wastes resources and is **unethical**.
- **Too <u>many</u> samples:**
  - wastes resources e.g. animals, money, time and effort, and is **unethical**.



How Do I Tell?

- It's possible to calculate the required power of an experiment. For RNA-seq, there's even an online tool (and there are various general calculators):

Gene Expression

Scotty: A Web Tool For Designing RNA-Seq Experiments to Measure Differential Gene Expression

M.A. Busby, C. Stewart, C. Miller, K. Grzeda, G. Marth
Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill MA, USA

- Alternatively, use rules of thumb based on other people's systematic reviews of a particular experiment type.

PART 1:
The Technologies

NGS Technologies

(Clonally Amplified DNAs)

Experimental Design, Sample Preparation & Multiplexing

Application & Scope

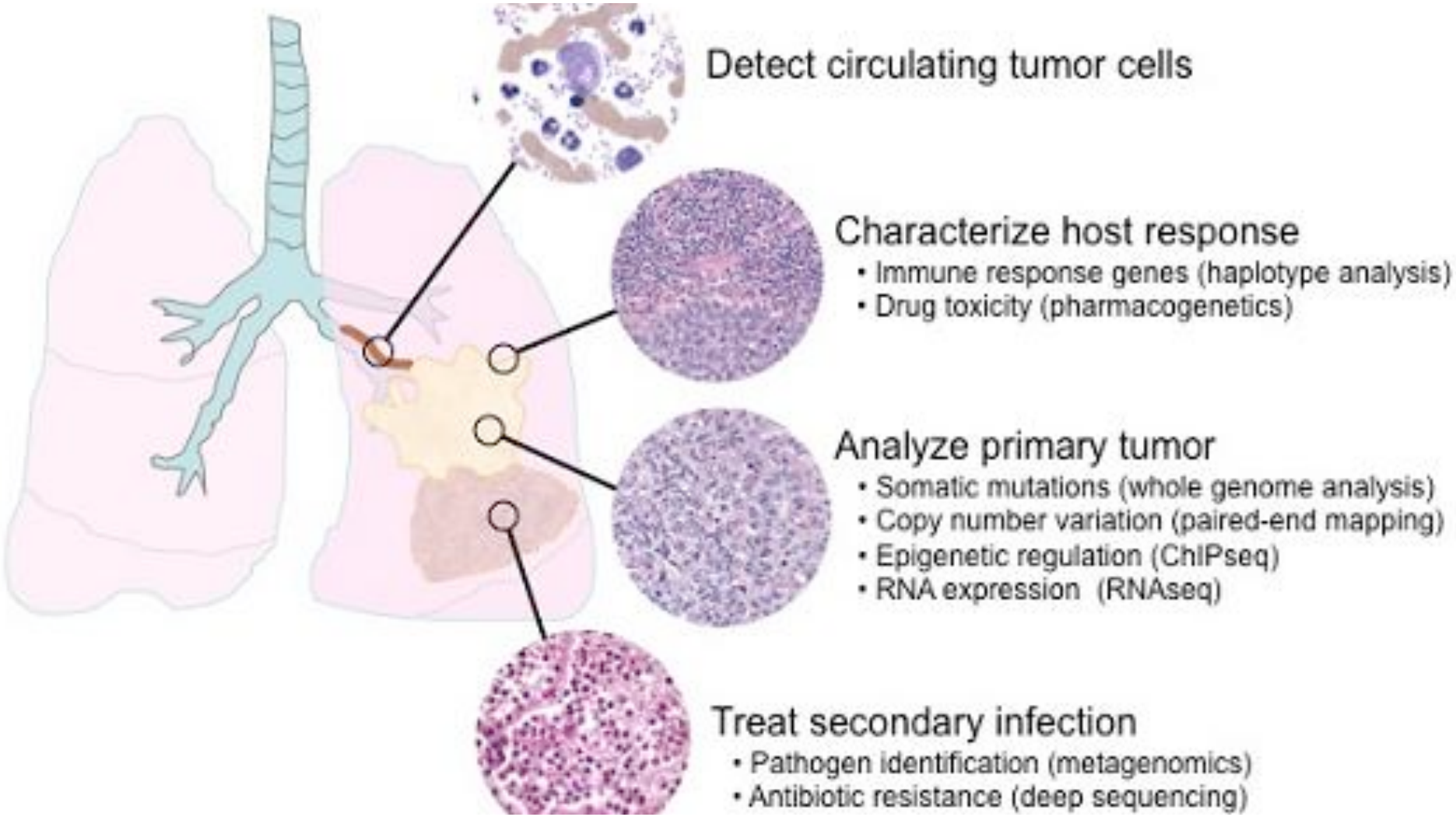
Clinical Application

How will High-throughput DNA sequencing impact clinical diagnostics?

- Improved detection of **SNPs** and **somatic mutations**.
- More accurate **haplotype analysis**.
- Detect balanced and unbalanced **copy number** variation.
- Better understand **epigenetic regulation**
- Quantitative **transcriptome analysis**
- **Deep sequencing** applications (meta-genomics and minimal residual disease detection)

Matthew Anderson

Lung Cancer diagnosis in 2020?

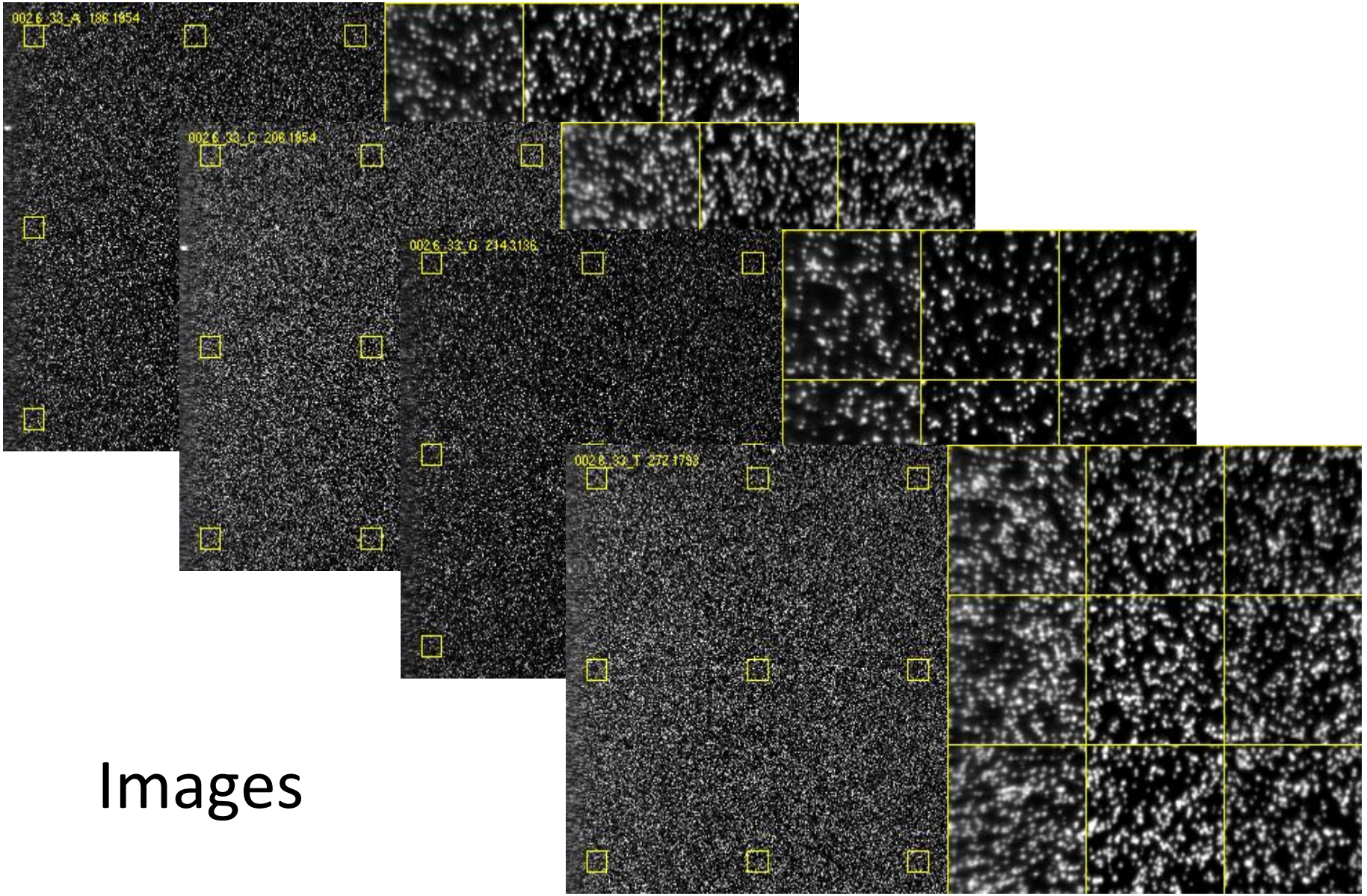


Anderson & Schrijver, Genes 2010, *1*(1), 38-69

PART 2:
The Bioinformatics

Bioinformatics Workflow

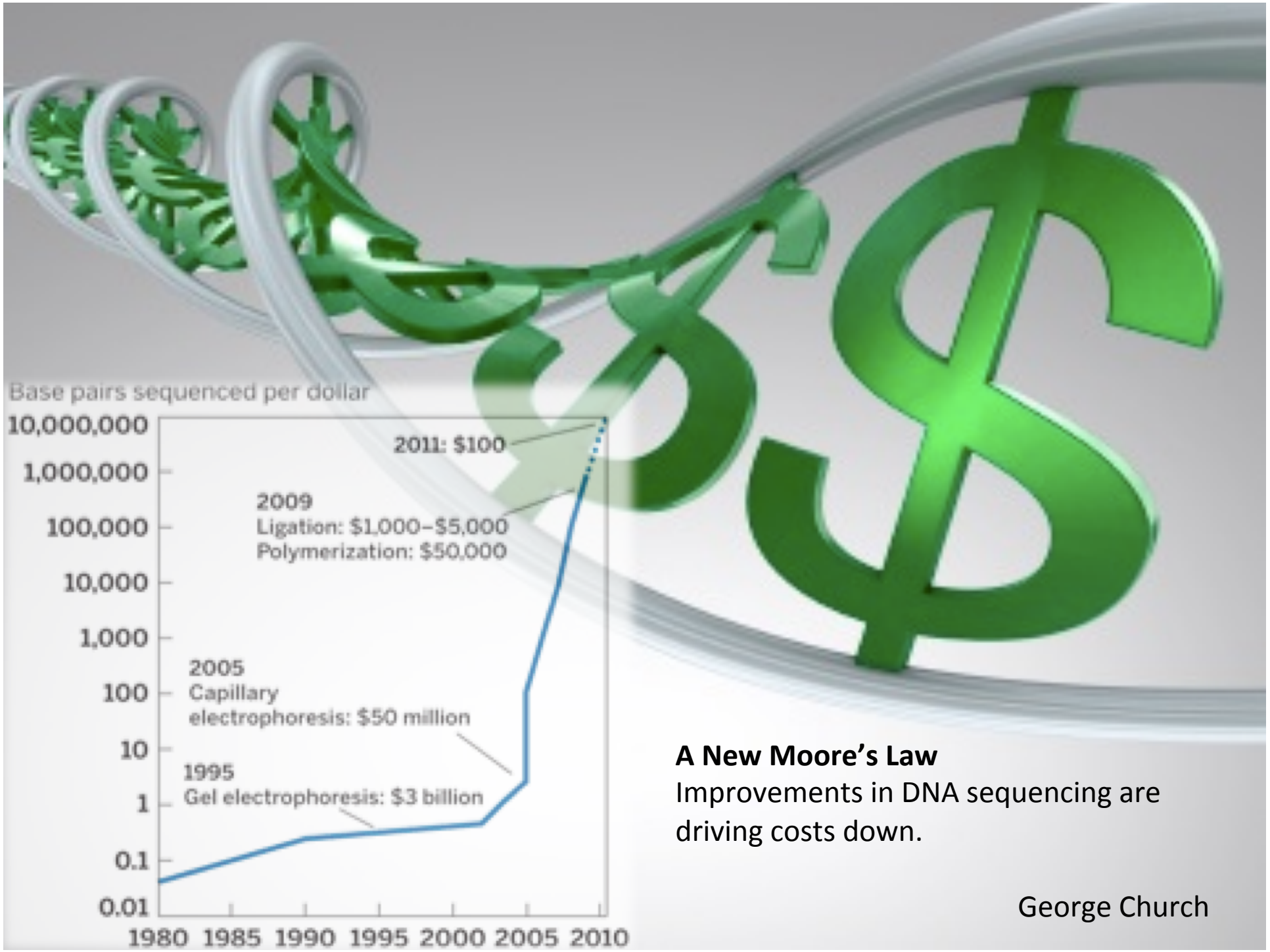
Data Output



Images

Data Output

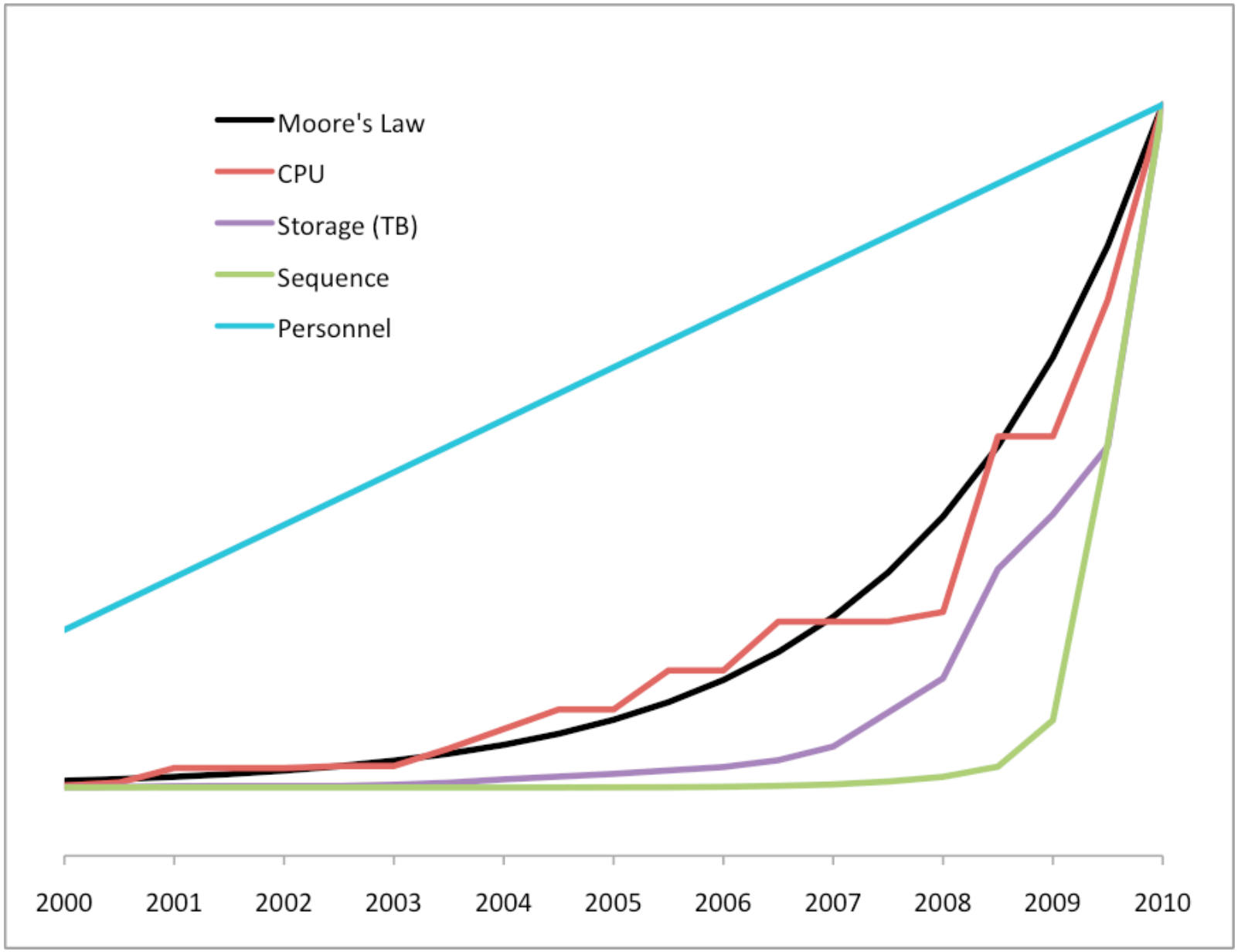
- Image data output (tiff files)
- 100 tiles per lane, 8 lanes per flow cell, 100 cycles
- 4 images (A,G,C,T) per tile per cycle = 320,000 images
- Each tiff image is ~ 7 MB = 2,240,000 MB of data (2.24 TB !)
- 4.5 TB for 100 nt Paired-end read



A New Moore's Law
Improvements in DNA sequencing are driving costs down.

George Church

Moore's Law: CPU, Storage & Sequencing



“lossy” compression

- Are we **running out of data storage space**?
- In some databases at the EBI it’s gotten very close to that. They don’t just store the genomes, but they store the **raw data**: the output of the Illumina machines.
- It has to be compressed in order to store it but we’re getting to the point where we have to use “lossy” compression, so we’re beginning to **lose information.**
- There’s been a lot of discussion in this field about **what** can we afford to lose and **how much** can we afford to lose.
- This field is sort of on the edge of deciding what we are going to throw away - we’re outrunning **Moore’s law**!!!!

What do we do with all of this data?

Illumina Throughput & Output

- Read Length 36, 75, 100 nt
  - 150 nt and more, soon
  - Single Read, Paired-End reads as well as Mate-Pair reads
- 8 lanes per flow cell, 15-20 million reads per lane
- ~ 30 Gbases per flow cell
  - PE,100nt
- 1TB raw data / run
- **Accuracy** is ~99 - 99.5%
  - Primary type of error: Substitution
  - 150 million errors per flow cell!
- **Serious computing** required for primary analysis:
  - Image analysis to base calling
  - Alignment
  - Assembly

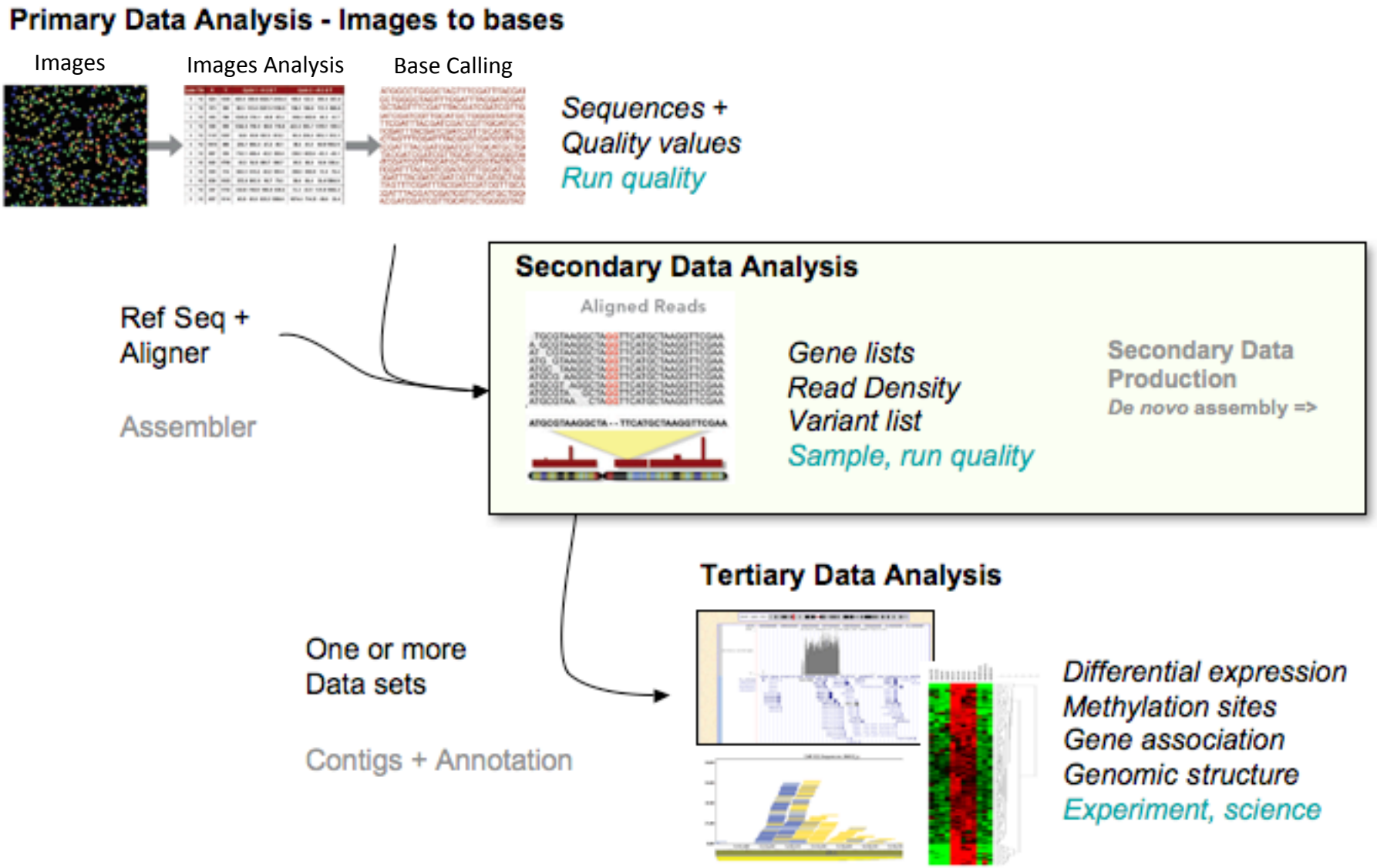
What do we do with all of this data?

- Short read lengths lead to difficulty in reassembling the genomic sequence.
- Multiple reads necessary to increase confidence of variant calls
- Analysis strategy depends on particular application:
  - *De novo* genome assembly.
  - Alignment to a reference genome.
  - Alignment to a synthetic reference genome
  - 'Counting' applications (copy number variation, RNA expression)
- Must account for error modalities of platform

Bioinformatics Workflow

- Image Extraction
- Base Calling, quality scoring
- Align reads to known sequence OR each other
- Assemble Reads
- Analysis of genes, regions
- Coverage, quantification
- Annotation

Three-phase Analysis



Alignment

- A way of arranging the sequences of DNA to identify regions of similarity.

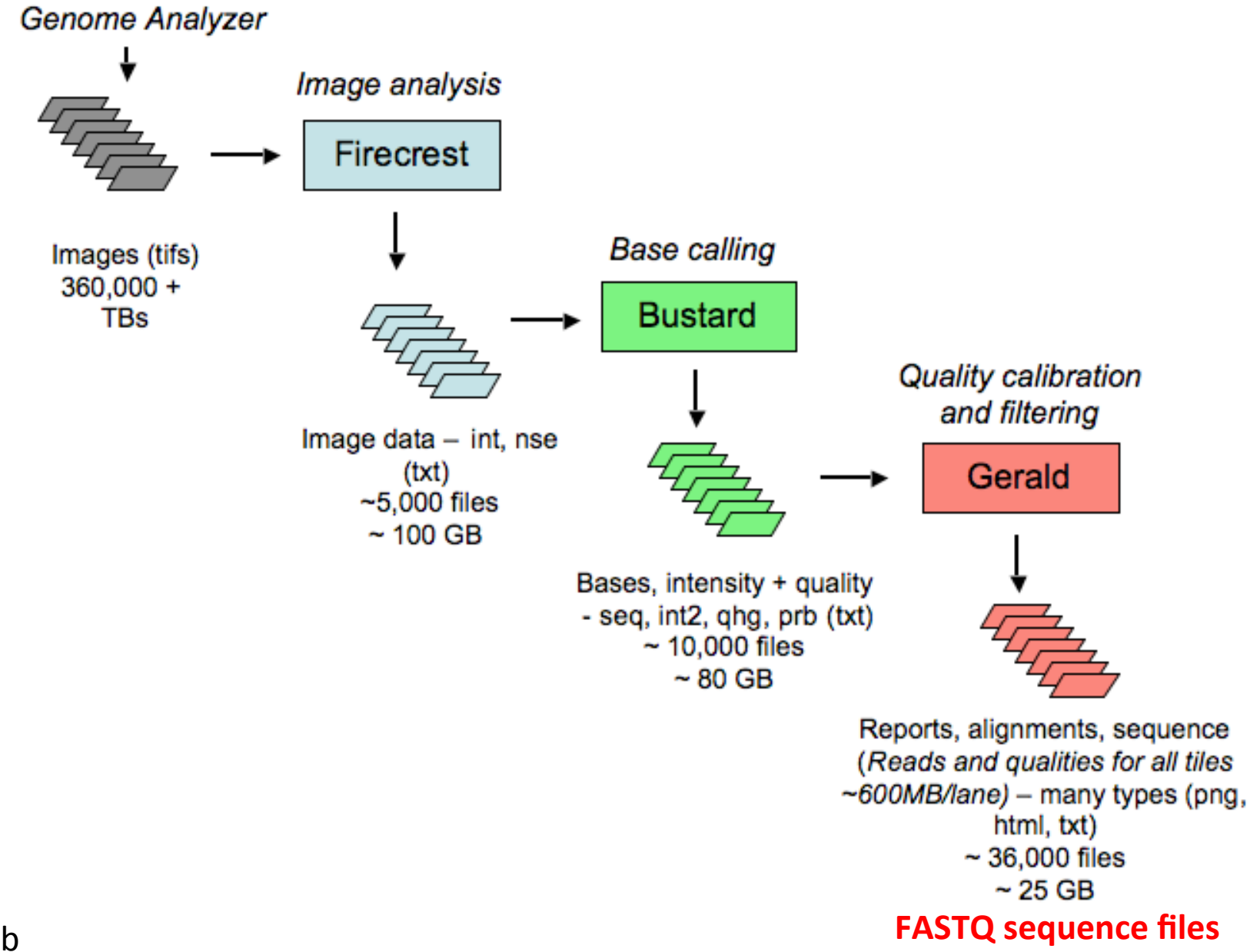
- Lengthy and highly variable sequences a challenge!

- Computational approaches to sequence alignment generally fall into two categories:
  - ***global* alignments**:

    Find the best overall alignment between sequences.
  - ***local* alignments**:

    Find short regions of highly conserved sequence.

Illumina Pipeline



Illumina Pipeline

- Firecrest: image analysis
  - Locates clusters and calculates intensity and noise
- Bustard: base calling
  - Deconvolutes signal and corrects for cross-talk, phasing
- GERALD:
  - **Local** alignment using software **ELAND** (aligns & creates files containing alignment locations and information
  - Read data aligned to genome sequence, RefSeq dataset, or mirbase sequences, etc.
  - Quality control is performed by randomly selecting 10% of the reads and aligning them to the reference using the GEM mapper.
  - Generates FastQC reports.
  - ELAND Limitations:
    - tolerates only two mismatches
    - does not allow indel errors
    - read length limitation
    - trims all reads to the same length

Illumina Pipeline

- GERALD calculates ***Chastity Score*:**

  "the ratio of the highest of the four (base type) intensities to the sum of highest two".

- Score is used to remove clusters with low signal to noise ratio
  - often caused by clusters being too close to each other so their signals bleed into one another.

- Default cut-off for filtering is >=0.6.

Other Software Application for Alignment & Assembly

Align/Assemble to a reference

* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hou workstation with 2 gigabytes of memory. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony Solexa 1G machine.
* EULER - Short read assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research).
* Exonerate - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Bi EMBL. C for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C
* MOSAIK - Reference guided aligner/assembler. Written by Michael Strömberg at Boston College.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has prelimina handle ABI SOLiD data. Written by Heng Li from the Sanger Centre.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficie library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, A Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. P required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available free for evaluatio use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* RMAP - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinforma OS required.
* SeqMap - Works like ELand, can do 3 or more bp mismatches and also INDELs. Written by Hui Jiang from the Wong lab at Stanford. Builds available for n
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Bru Stephen Rumble at the University of Toronto.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a sequence or a set of reference sequences.. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto referen Author is Ruiqiang Li at the Beijing Genomics Institute. C++ for Unix.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases usi Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq genome builds for a number of organisms. Web Portal. OS independent.

***de novo* Align/Assemble**

* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing te or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecul
* SSAKE - Version 2.0 of SSAKE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X co paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

http://seqanswers.com/

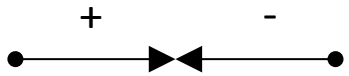
Sequence Alignment Tools

- **BFAST**: Blat like Fast Alignment Tool; Fast and accurate mapping of short reads to reference sequences.
- **Bowtie**: Ultrafast short read aligner; Burrows-Wheeler-Transformed (BWT) index.
- **BWA**: Burrows-Wheeler *Alignment* tool; Gapped global alignment.
- **ELAND**: part of Illumina SW & runs on single processor; Local Alignment.
- **SOAP**: Short Oligonucleotide Alignment Program; Efficient gapped and ungapped alignment of short oligos onto reference sequences. SOAP2 based on BWT.
- **SSAHA**: Sequence Search and Alignment by Hashing Algorithm (Smith-Watermann).
- **SOCS**: SOLiD reference based, un-gapped alignment with bisulfite capability; Rabin-Karp string search algorithm, which uses hashing.
- ***Bioconductor*** *packages:*
  - **`Rsubread`** package;
  - **`matchPDict`** in the **`Biostrings`** package - particularly useful for flexible alignment of moderately sized subsets of data.

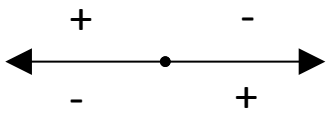
Processing Paired End Data

1. Align each end

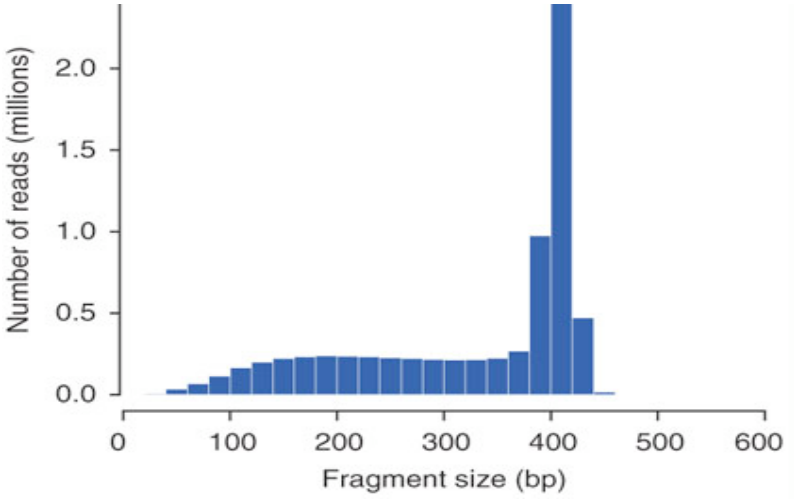
2. Match orientation
   - Short inserts:



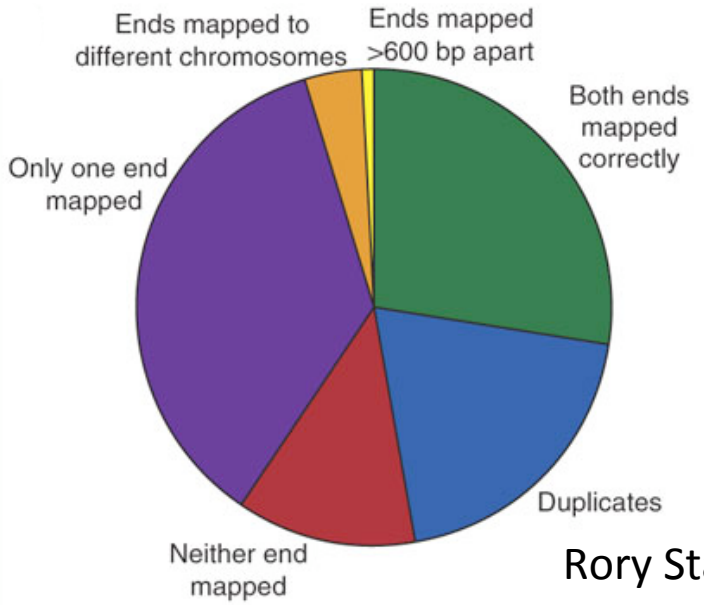
   - Long inserts:



3. Compute insert size:



4. Check for irregular alignments
   - Insert size beyond 3 sd: deletions
   - Mismatched orientations: inversions
   - Difference chromosomes: chromosomal rearrangements



Rory Stark

File Formats & FASTQC

Raw and Aligned Reads

- Raw data is a (large) set of sequences
- Typical file format is FASTQ

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
hhhhhhhhhhhhhhhdhhhhhhhhhhhdRehdh
```

Read identifier

Bases called

Base quality codes

- Alignment to genome is done by efficient indexing of seed sequences
- Aligned reads in SAM format

```
@HWI-… 163 chr19 9900 10000 16M2I25M
```

Read identifier

Where this read matched

Start and end positions

Codes for match:
16 matches, 2 extra,…

Phred Quality Scores

- Q scores are used to measure **base calling accuracy**.
- Q scores are defined as a property that is logarithmically related to the **base calling error probabilities** (P):

$$Q = -10 \log_{10} P$$

- When sequencing quality reaches Q30, virtually all of the reads will be perfect, having zero errors and ambiguities.
- Low Q scores can increase false-positive variant calls, which can result in inaccurate conclusions and higher costs for validation experiments.

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

PHRED, Ewing & Green, 2008

Typical Quality Scores for Illumina

Phred Score	Quality Score	ASCII Code
25	089	Y
26	090	Z
27	091	[
28	092	\
29	093	]
30	094	^
31	095	_
32	096	‘
33	097	a
34	098	b
35	099	c

The difference between Solexa/Illumina scores and Phred Quality scores is documented here:
http://maq.sourceforge.net/qual.shtml

FastQC

- FastQC is free software under the GPLv3 (Andrews, 2010). You can download it from:

  http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
  http://www.youtube.com/watch?v=bz93ReOv87Y

- Perform QC checks on high throughput sequence data.

- Imports data from BAM, SAM, or FASTQ files and generates eleven summary plots

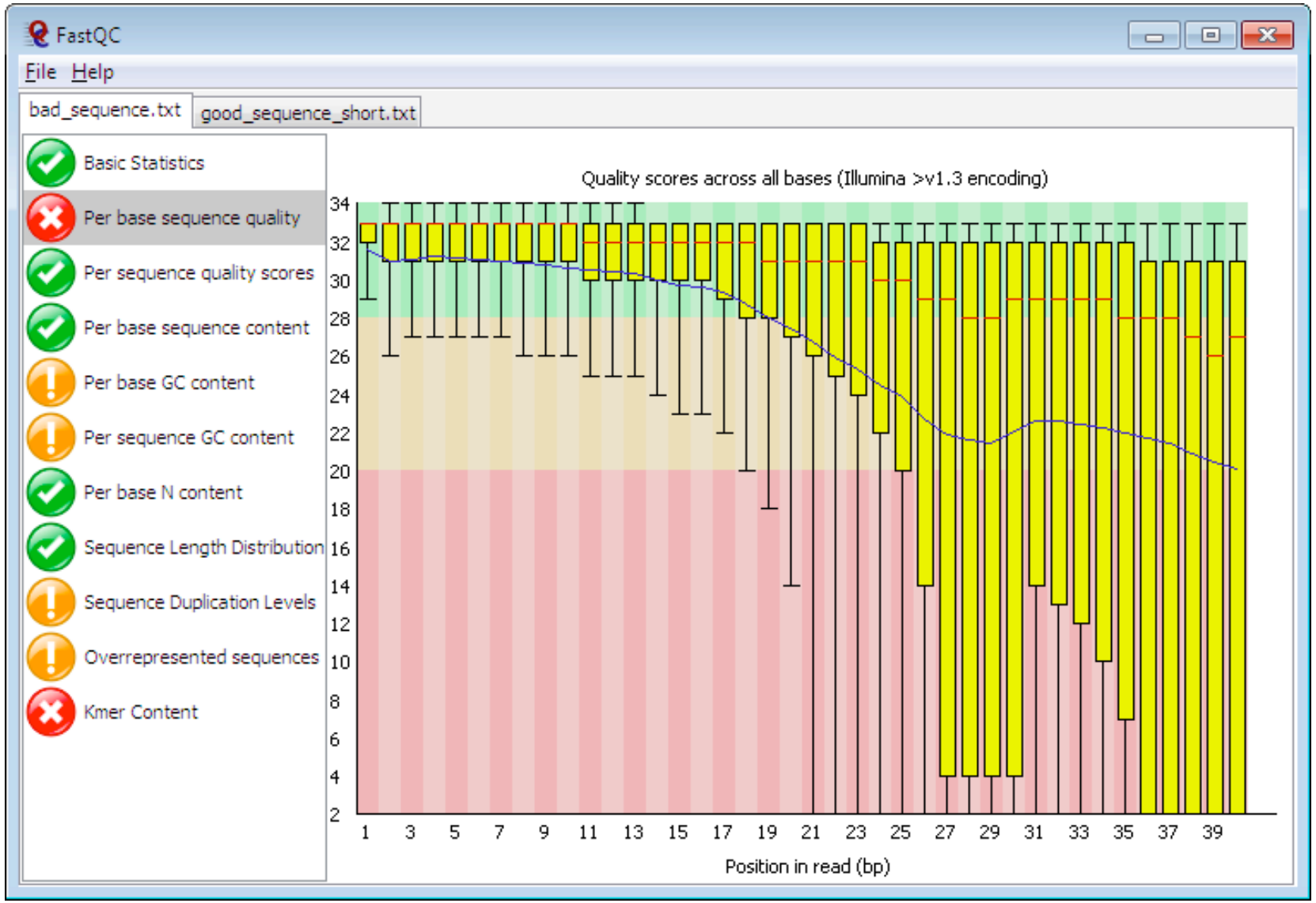
- Graphically presents **multiple metrics** for each dataset, including:
  - per base sequence quality score,
  - per base GC content
  - over-abundance of adaptors
  - over-represented sequences (to infer duplication rate).

- FastQC runs a series of tests and will flag up and potential problems with your data.

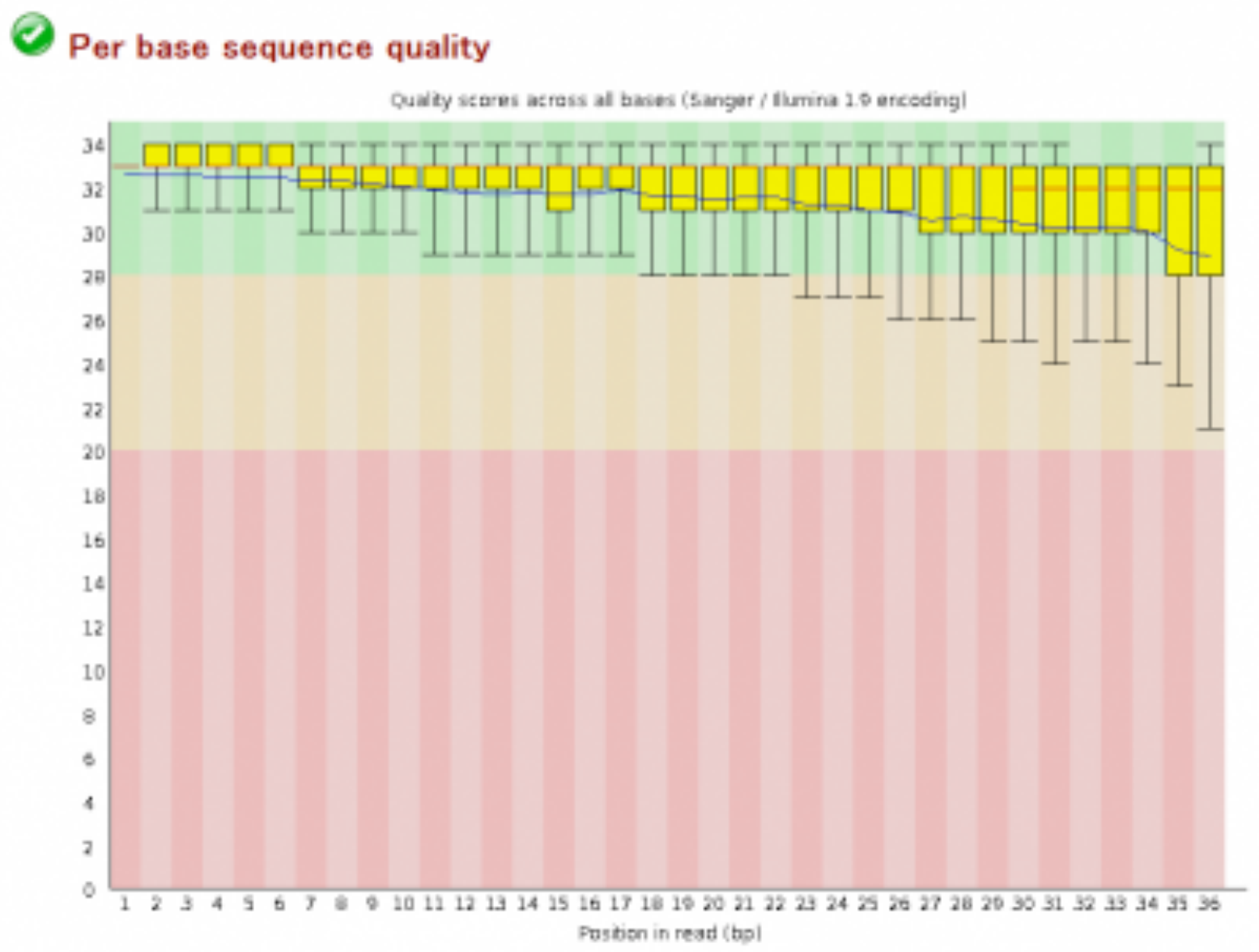
- Run as an interactive GUI application or run in an unattended offline mode where it generates HTML versions of its reports.

- Graphics: read length plots; read-quality plots; sequence duplication levels and many more

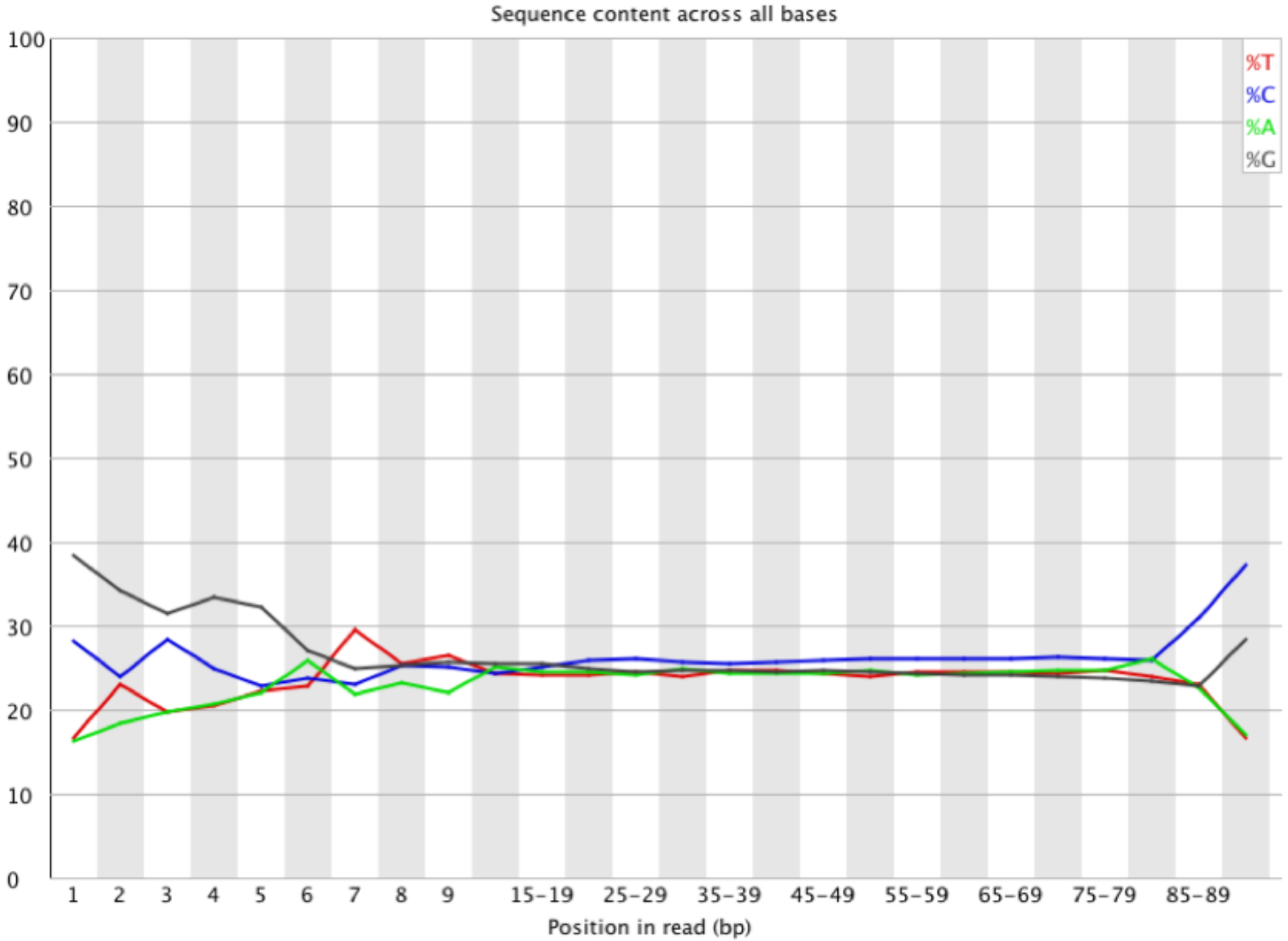
FastQC


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)

FastQC

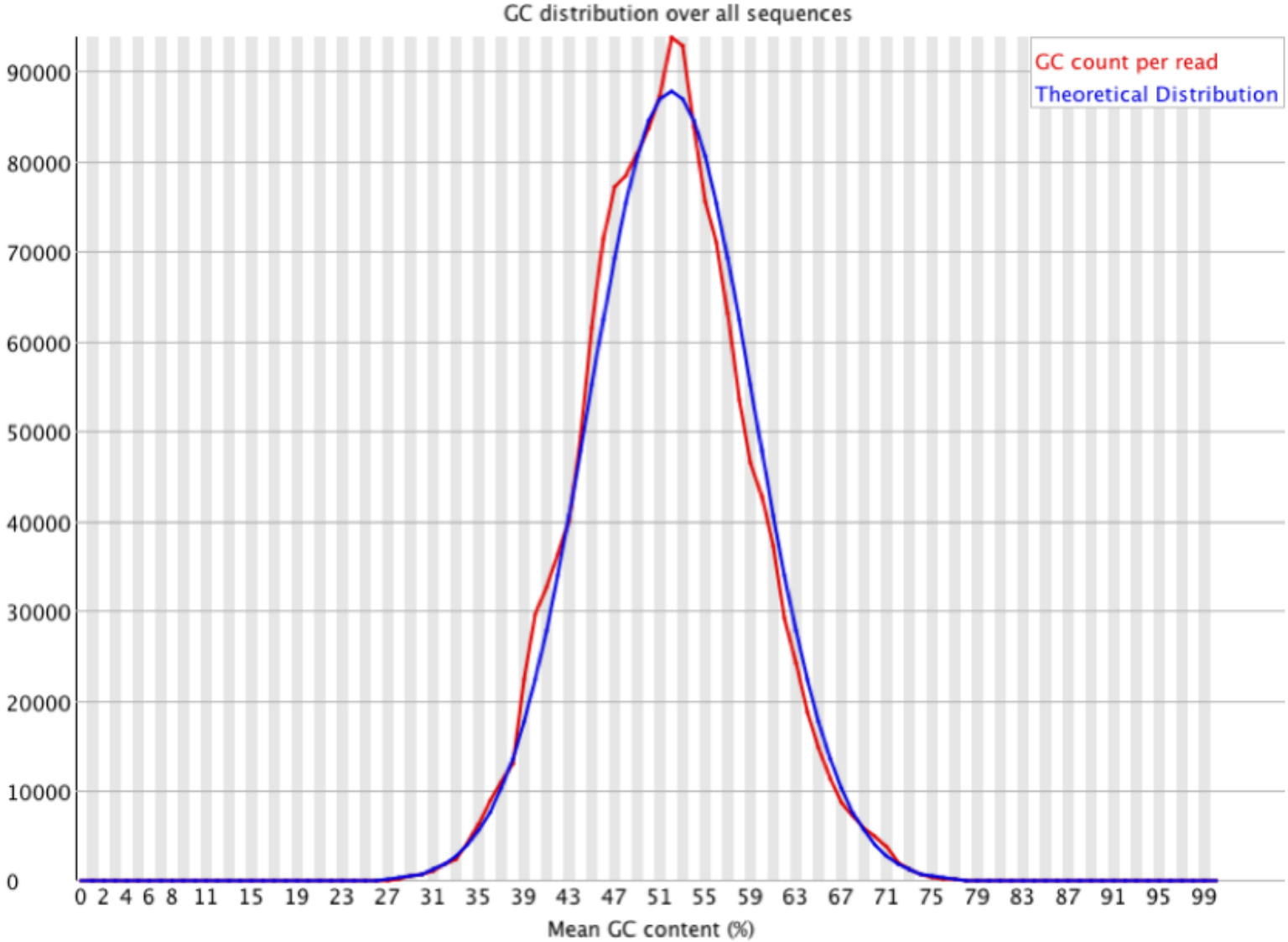

Per base sequence quality
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
Position in read (bp)

FASTQC:
Per base sequence content



Should see and equal distribution of the 4 bases which doesn't change with base position.

FASTQC:
Per sequence GC content



An ideal run would have very close matches between the distributions. Secondary peaks and large deviations from the theoretical distribution can indicate contamination.

QA: Multi-genome alignment (MGA) contamination screen

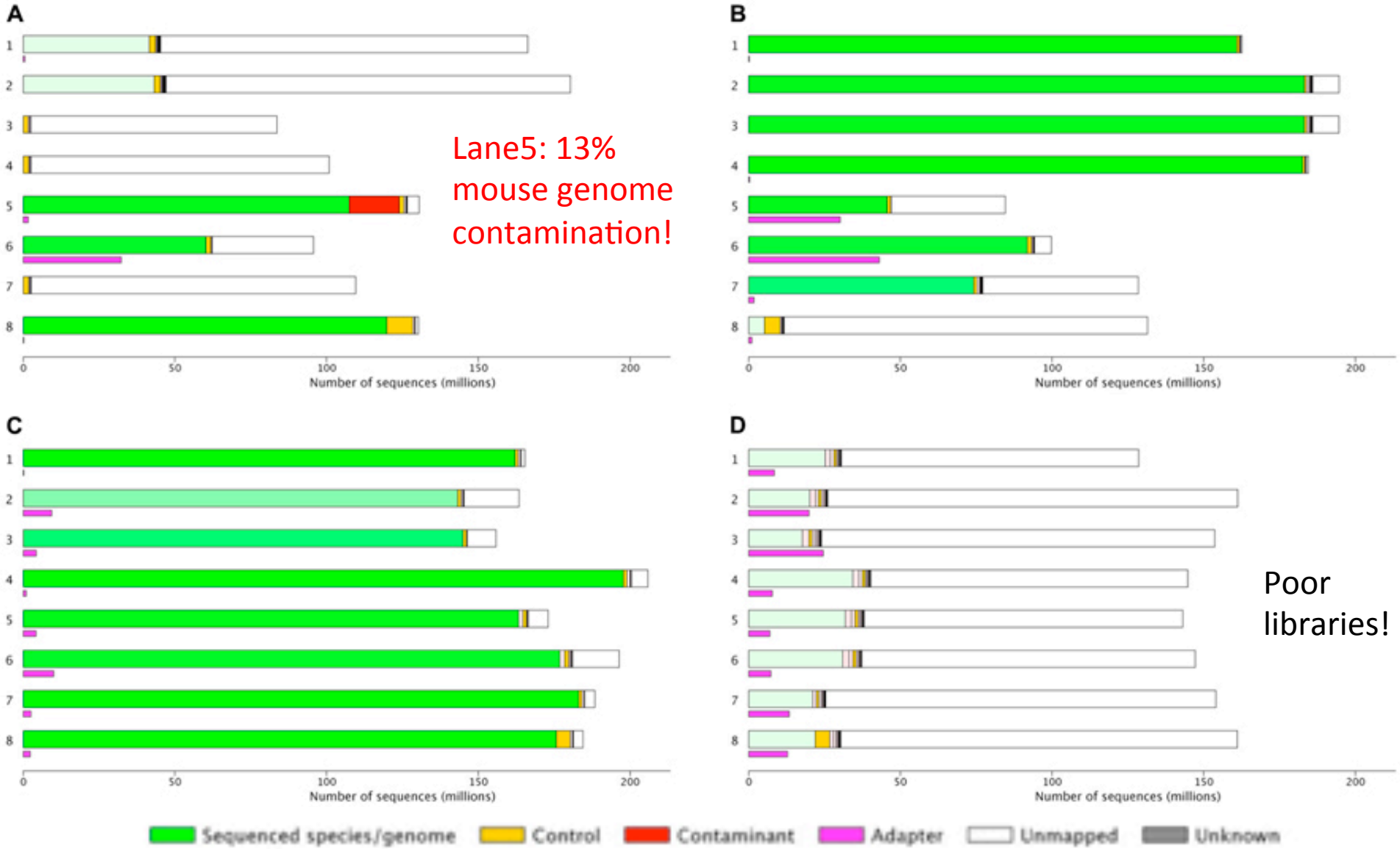
- Bionformatics & Genomics Core Facility, CI CRUK
  - Matthew Eldridge & James Hadfield

- Assessment for each lane:
  - Identify sequences from **another species**
  - Detect presence of **adapter sequences** ligated to the ends of sequence fragments.

- Detecting contamination in genomic and transcriptomic sequencing libraries:
  - Reads trimmed to 36b
  - Aligned to a set of possible contaminants - 23 reference genomes & 1000 bacteria, viruses & fungi using **bowtie** (Langmead et al., 2009)

- Align reads to a set of adaptor & primer sequences using **exonerate** (Slater & Birney, 2005)

QA: Multi-genome alignment (MGA) contamination screen



Hadfield & Eldridge (2014) Front Genet. 5:31

Pre-processing & QC

- **Trimming & Pattern Matching** (e.g., primer removal)
  - Good idea to trim your raw reads before assembly.
  - The primary reason for this is to remove poor quality reads that might reduce assembly speed and accuracy.
  - One popular Java tool for trimming raw reads is **Trimmomatic**:
    http://www.usadellab.org/cms/?page=trimmomatic
  - Bioconductor packages: **`ShortRead`** or **`Biostrings`**

- **General Quality Assessment & Reports**
  - **FASTQC**
  - Bioconductor packages: **`ShortRead`** and **`rsamtools`**

- **Marking PCR Duplicates:**
  - Marked in BAM files using **Picard**
  - Due to inherent mistakes in the sequencing technology (amplification biases), some reads will be exact copies of each other.

- **Alignment:**
  - Not a solved problem
  - Mapping reads to the reference genome
  - Accuracy of alignment significantly affects downstream analysis
  - **Filtering output:**
    - Read mapping > x number of times
  - Bioconductor package: **`GenomicRanges`**

Aligned reads: BAM files

- Binary form of SAM format taking about a quarter of the space.

- Convert SAM to BAM
  - **samTools** or **picard**
  - compresses the SAM file and can be indexed.
  - portions of the file can be accessed without the need to load the whole file.

- http://samtools.sourceforge.net/SAM1.pdf

Count Data
(RNA-seq & ChIP-seq)

- Count reads, not base-pairs
- Count each read at most once.
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Integrated Pipelines

- SOLID TM System Analysis Pipeline Tool
- CLCBio Genomic workbench
- DNAnexus – cloud computing
- Galaxy Server
- ERANGE: Full package for RNASeq and chipSeq data analysis

- **R & Bioconductor** workflows – “pipelinable”

Bioinformatics Workflow

- Image Extraction
- Base Calling, quality scoring
- Align reads to known sequence OR each other
- Assemble Reads
- Analysis of genes, regions
- Coverage, quantification
- Annotation

R & Bioconductor

- **Robert Gentleman** & **Ross Ihaka,** 1996

- Complete statistical package and programming language
- Efficient functions and data structures for data analysis
- Powerful graphics
- Access to fast growing number of analysis packages
- Most widely used language in bioinformatics
- Is standard for data mining and biostatistical analysis

- Technical advantages:
  - free, open-source, available for all OSs
  - Interfaces with Perl, Python, Java, C, XML

- Active user community

http://cran.r-project.org/

Literature on 

Books

- simpleR - Using R for Introductory Statistics (Verzani, 2004)
  http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

- Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Gentleman et al., 2005)
- R programming for Bioinformatics - Gentleman (2005)
- An Introduction to R - Venables WN et al (2005)
- R Graphics - Murrell (2009)

Online Guides

- Kickstarting R:
  http://cran.r-project.org/doc/contrib/Lemon-kickstart/index.html
- R & Bioconductor Manual:
  http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual

CambR Meeting

Cambridge R user group

Organised by Mark Dunning & Laurent Gatto

https://groups.google.com/forum/?fromgroups=#!forum/cambridge-r-user-group

CambR meeting

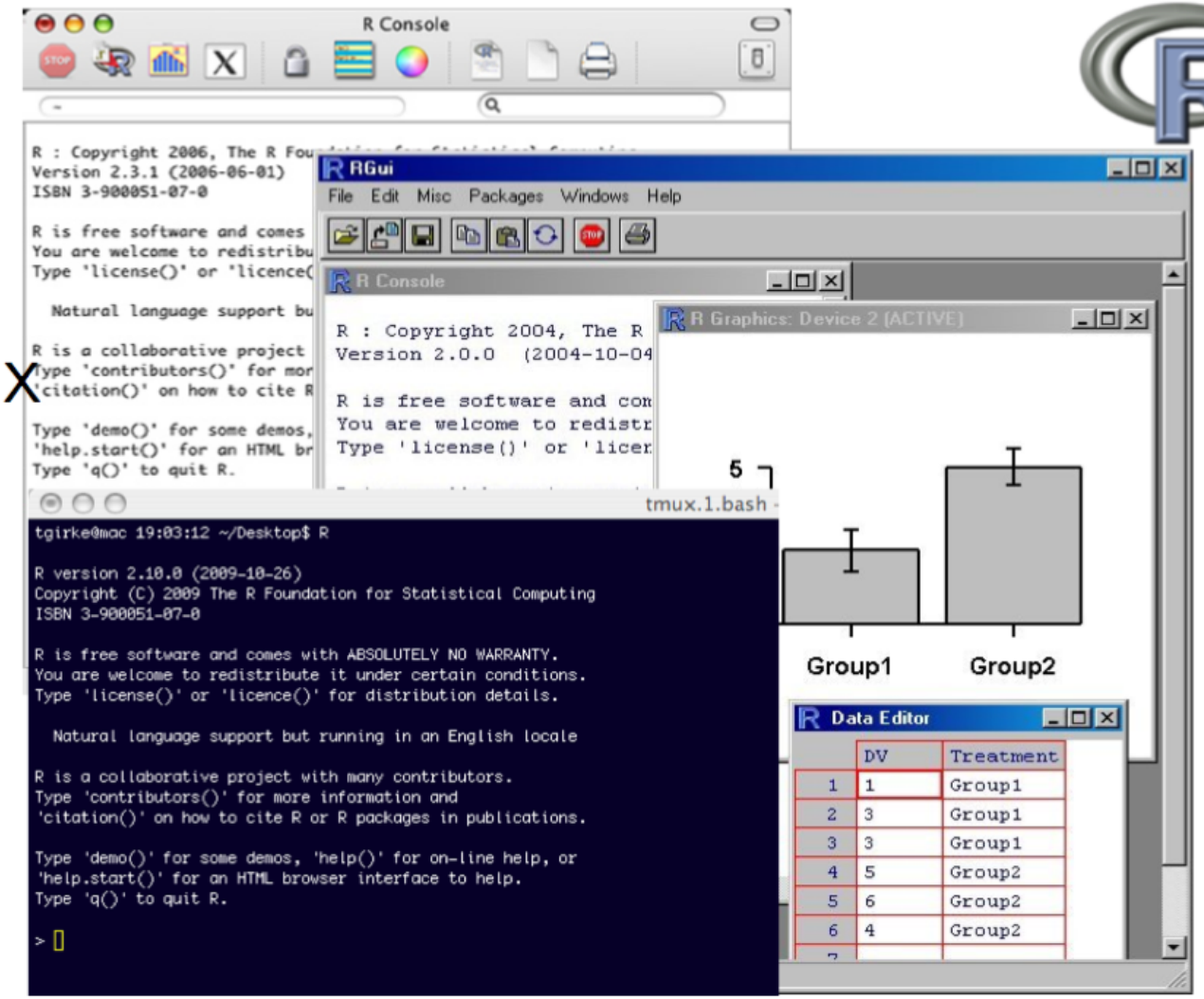
When: Monday, 29th October 2012 - doors open at 6:30 pm, talk(s) start at 7:00pm

Where: The Fountain Inn [2], 12 Regent Street, Cambridge, CB2 1DB

Talk given by **Martin Morgan**, who leads the famous Bioconductor project.

http://blog.revolutionanalytics.com/2011/05/the-r-files-martin-morgan.html

Working Environment



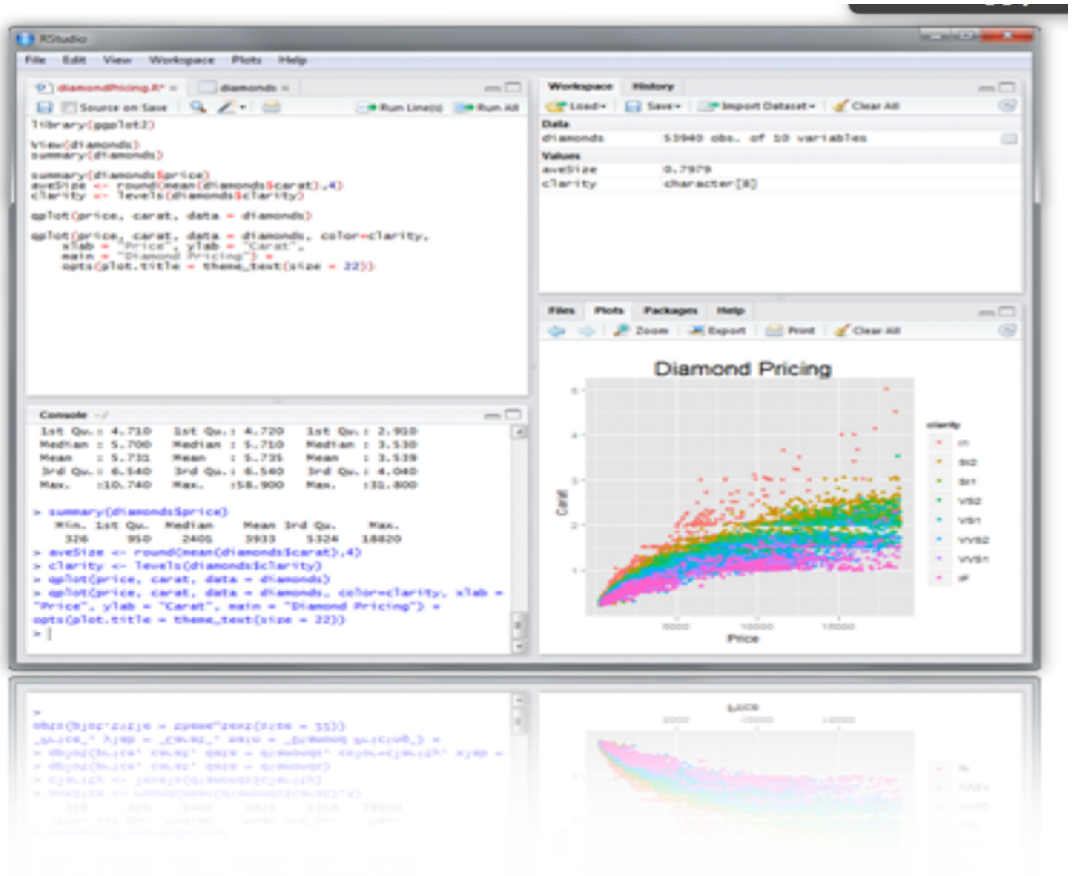
R Gui: OS X

Command-line R: Linux/OS X

R Gui: Windows

RStudio

- Alternative working environment for R.
- Integrated development environment (IDE) that works well for beginners and developers.



Thomas Girke

Package Depositories

- **CRAN** (>3500 packages) general data analysis
- **Bioconductor** (>600 packages) bioscience data analysis
- **Omegahat** (>30 packages) programming interfaces

Graphics in R

- **Powerful** environment for visualizing scientific data
- **Integrated** graphics and statistics infrastructure
- Vast number of R packages with **graphics utilities**
- **Publication quality** graphics
- Fully **programmable**
- Highly **reproducible**
- **LATEX** & **Sweave** support

Documentation on Graphics in R

- **General**
  - Graphics Task Page
    http://cran.r-project.org/web/views/Graphics.html
  - R Graph Gallery
    http://research.stowers-institute.org/efg/R/
  - R Graphical Manual
    http://rgm3.lab.nig.ac.jp/RGM/
  - Paul Murrell's book R (Grid) Graphics

- **Interactive Graphics**
  - rggobi (Ggobi)
    http://www.ggobi.org/
  - Iplots
    http://www.rosuda.org/iplots/
  - Open GL (rgl)
    http://rgl.neoscientists.org/gallery.shtml

Graphics Environments

- **Viewing and saving graphics in R**
  - On-screen graphics
  - postscript, pdf, svg
  - jpeg, png, wmf, tiff, ...

- **Four major graphic environments**
  - Low-level infrastructure
    - R Base Graphics (low- and high-level)
    - grid
  - High-level infrastructure
    - Lattice
    - ggplot2

Reporting in R

- **Tools:**
  - Sweave; KnitR; Shiny; Rpubs
- **Reproducibility**
  - Make your research more reproducible.
- **Efficiency**
  - Statistical output is automatically incorporated into your report.
- **Reliability**
  - The integration of analyses with the report reduces the chance of errors entering in through copying and pasting of statistical output into documents.
- **Education & Communication**
  - By providing data analysis code for a report, this teaches others how to do similar analyses.

Sweave

What is it?

Sweave is a tool that allows to embed the R code for complete data analyses in latex documents.

Create dynamic reports, which can be updated automatically if data or analysis change.

Where can I get it?

The Sweave software itself is part of every R installation, see

```
help("Sweave", package="utils")
```

to get started.

http://www.stat.uni-muenchen.de/~leisch/Sweave/

Bioconductor

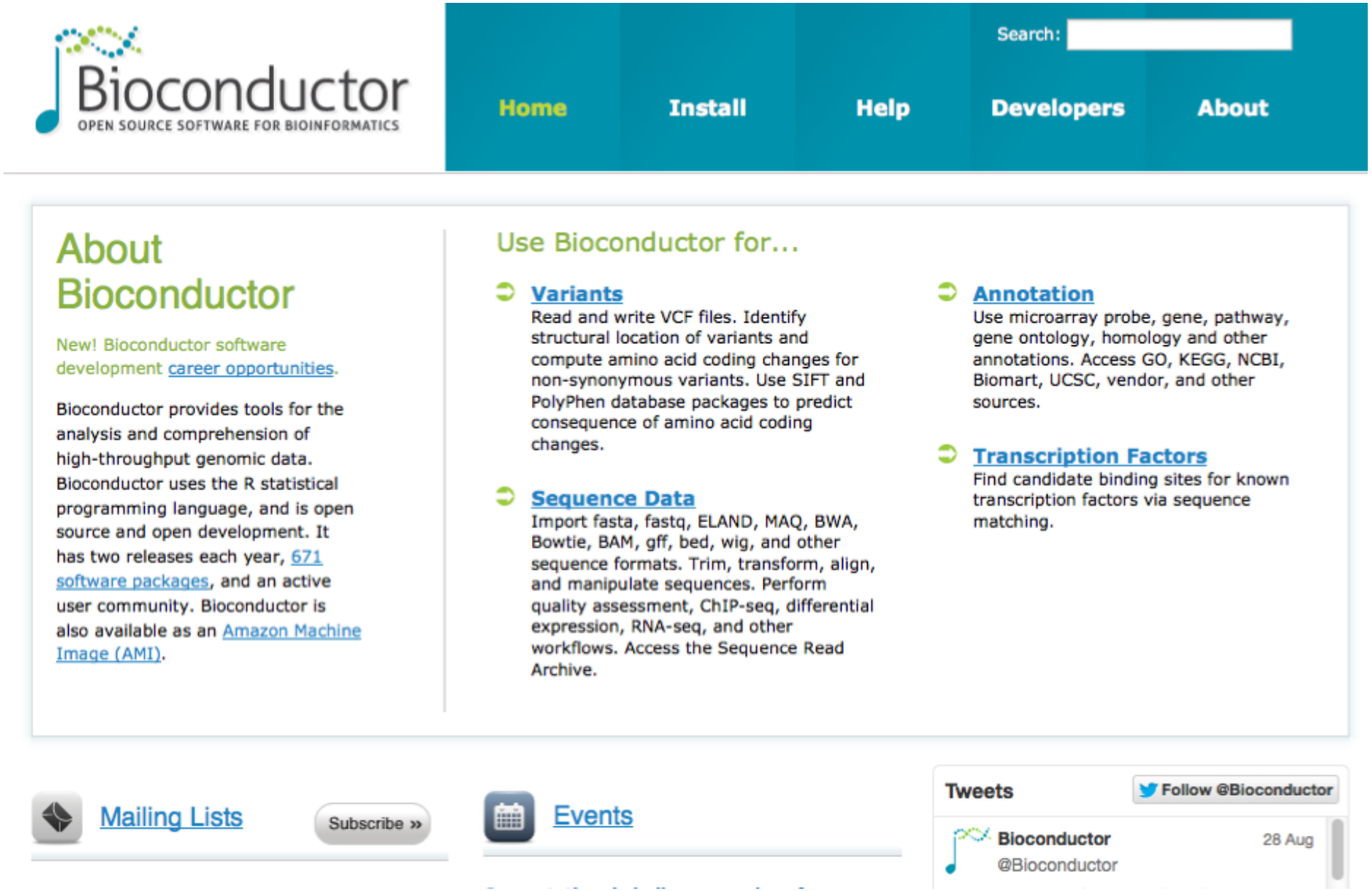
- Open source software for bioinformatics
- Bioconductor is a series of R packages
- Core emphasis on reproducible research, good documentation and training, re-usable data structures, designed to work with different variations of data
- Questions about the analysis of array data using Bioconductor can be posted on their mailing list. This is a very informative mailing list for the analysis of data from a wide variety of high throughput genomic technologies.

  http://www.bioconductor.org/docs/mailList.html

  http://www.bioconductor.org/

For a quick install of a subset of the most common packages:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite()
```



http://www.bioconductor.org/

Bioconductor Features

- Introductory **workflows.**
- A **manifest** of Bioconductor packages (BiocViews).
- **Annotation** (data bases of relevant genomic information).
- **Mailing lists**, including searchable archives (primary source of help).
- **Course/conference** information, including extensive reference material.
- **Package developer resources**, including guidelines for creating and submitting new packages.

HTP-seq Packages

Concept	Packages
Data representation	*IRanges*, *GenomicRanges*, *GenomicFeatures*, *Biostrings*, *BSgenome*, *girafe*.
Input / output	*ShortRead* (fastq), *Rsamtools* (bam), *rtracklayer* (gff, wig, bed), *VariantAnnotation* (vcf), *R453Plus1Toolbox* (454).
Annotation	*GenomicFeatures*, *ChIPpeakAnno*, *VariantAnnotation*.
Alignment	*gmapR*, *Rsubread*, *Biostrings*.
Visualization	*ggbio*, *Gviz*.
Quality assessment	*qrqc*, *seqbias*, *ReQON*, *htSeqTools*, *TEQC*, *Rolexa*, *ShortRead*.
RNA-seq	*BitSeq*, *cqn*, *cummeRbund*, *DESeq*, *DEXSeq*, *EDASeq*, *edgeR*, *gage*, *goseq*, *iASeq*, *tweeDEseq*.
ChIP-seq, etc.	*BayesPeak*, *baySeq*, *ChIPpeakAnno*, *chipseq*, *ChIPseqR*, *ChIPsim*, *CSAR*, *DiffBind*, *MEDIPS*, *mosaics*, *NarrowPeaks*, *nucleR*, *PICS*, *PING*, *REDseq*, *Repitools*, *TSSi*.
Variants	*VariantAnnotation*, *VariantTools*, *gmapR*
SNPs	*snpStats*, *GWASTools*, *hapFabia*, *GGtools*
Copy number	*cn.mops*, *genoset*, *CNAnorm*, *exomeCopy*, *seqmentSeq*.
Motifs	*MotifDb*, *BCRANK*, *cosmo*, *cosmoGUI*, *MotIV*, *seqLogo*, *rGADEM*.
3C, etc.	*HiTC*, *r3Cseq*.
Microbiome	*phyloseq*, *DirichletMultinomial*, *clstutils*, *manta*, *mcaGUI*.
Work flows	*QuasR*, *easyRNASeq*, *ArrayExpressHTS*, *Genominator*, *oneChannelGUI*, *rnaSeqMap*.
Database	*SRAdb*.

HTP-seq Packages

Concept	Packages
Data representation	*IRanges*, *GenomicRanges*, *GenomicFeatures*, *Biostrings*, *BSgenome*, *girafe*.
Input / output	*ShortRead* (fastq), *Rsamtools* (bam), *rtracklayer* (gff, wig, bed), *VariantAnnotation* (vcf), *R453Plus1Toolbox* (454).
Annotation	*GenomicFeatures*, *ChIPpeakAnno*, *VariantAnnotation*.
Alignment	*gmapR*, *Rsubread*, *Biostrings*.
Visualization	*ggbio*, *Gviz*.
Quality assessment	*qrqc*, *seqbias*, *ReQON*, *htSeqTools*, *TEQC*, *Rolexa*, *ShortRead*.
RNA-seq	*BitSeq*, *cqn*, *cummeRbund*, *DESeq*, *DEXSeq*, *EDASeq*, *edgeR*, *gage*, *goseq*, *iASeq*, *tweeDEseq*.
ChIP-seq, etc.	*BayesPeak*, *baySeq*, *ChIPpeakAnno*, *chipseq*, *ChIPseqR*, *ChIPsim*, *CSAR*, *DiffBind*, *MEDIPS*, *mosaics*, *NarrowPeaks*, *nucleR*, *PICS*, *PING*, *REDseq*, *Repitools*, *TSSi*.
Variants	*VariantAnnotation*, *VariantTools*, *gmapR*
SNPs	*snpStats*, *GWASTools*, *hapFabia*, *GGtools*
Copy number	*cn.mops*, *genoset*, *CNAnorm*, *exomeCopy*, *seqmentSeq*.
Motifs	*MotifDb*, *BCRANK*, *cosmo*, *cosmoGUI*, *MotIV*, *seqLogo*, *rGADEM*.
3C, etc.	*HiTC*, *r3Cseq*.
Microbiome	*phyloseq*, *DirichletMultinomial*, *clstutils*, *manta*, *mcaGUI*.
Work flows	*QuasR*, *easyRNASeq*, *ArrayExpressHTS*, *Genominator*, *oneChannelGUI*, *rnaSeqMap*.
Database	*SRAdb*.

R Basics: A refresher

Installation of R & Add-on Packages

- Install **R** for your operating system from:
  - http://cran.at.r-project.org

- Install **RStudio** from:
  - http://www.rstudio.com/ide/download

- Installation of **CRAN** Packages

```
> install.packages(c("pkg1", "pkg2"))
> install.packages("pkg.zip", repos=NULL)
```

- Installation of **Bioconductor** Packages

```
> source("http://www.bioconductor.org/
biocLite.R")
> biocLite()
> biocLite(c("pkg1", "pkg2"))
```

R Basic Syntax

General R command syntax

```
> object <- function_name(arguments)
> object <- object[arguments]
```

Finding help

```
> ?function_name
```

Load a library

```
> library("my_library")
```

Lists all functions defined by a library

```
> library(help="my_library")
```

Load library manual (PDF file)

```
> vignette("my_library")
```

Data Types I

Numeric data: 1, 2, 3

```
> x <- c(1, 2, 3); x
[1] 1 2 3
> is.numeric(x)
[1] TRUE

> as.character(x)
[1] "1" "2" "3"
```

Character data: ”a”, ”b”, ”c”

```
> x <- c("1", "2", "3"); x
[1] "1" "2" "3"
> is.character(x)
[1] TRUE
> as.numeric(x)
[1] 1 2 3
```

Data Types II

Complex data

```
> c(1, "b", 3)
[1] "1" "b" "3"
```

Logical data

```
> x <- 1:10 < 5 >x
[1] TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE
FALSE FALSE
> !x
[1] FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE
TRUE TRUE
> which(x) # Returns index for the 'TRUE' values in logical vector
[1] 1 2 3 4
```

Data Objects: Vectors and Factors

Vectors (1D)

```
> myVec <- 1:10; names(myVec) <- letters[1:10]
> myVec[1:5]
abcde
12345
> myVec[c(2,4,6,8)]
bdfh
2468
> myVec[c("b", "d", "f")]
bdf
246
```

Factors (1D): vectors with grouping information

```
> factor(c("dog", "cat", "mouse", "dog", "dog", "cat"))
[1] dog cat mouse dog dog cat
Levels: cat dog mouse
```

Data Objects: Matrices, Data Frames and Arrays

Matrices (2D): two dimensional structures with data of same type

```
> myMA <- matrix(1:30, 3, 10, byrow = TRUE)
> class(myMA)
[1] "matrix"
> myMA[1:2,]
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    1    2    3    4    5    6    7    8    9    10
[2,]   11   12   13   14   15   16   17   18   19    20
```

Data Frames (2D): two dimensional structures with variable data types

```
> myDF <- data.frame(Col1=1:10, Col2=10:1)
> myDF[1:2, ]
  Col1 Col2
1    1   10
2    2    9
```

Arrays: data structure with one, two or more dimensions

```
> array(1:3, c(2,4))
```

Data Objects: Lists & Functions

Lists: containers for any object type

```
> myL <- list(name="Fred", wife="Mary", no.children=3, child.ages=c(4,7,9))
> myL
$name
[1] "Fred"
$wife
[1] "Mary"
$no.children
[1] 3
$child.ages
[1] 4 7 9

> myL[[4]][1:2] [1] 4 7
```

Functions: piece of code

```
> myfct <- function(arg1, arg2, ...) {
    + function_body
+}
```

General Subsetting Rules

Subsetting by positive or negative index/position numbers

```
> myVec <- 1:26; names(myVec) <- LETTERS
> myVec[1:4]
 A B C D
 1 2 3 4
```

Subsetting by same length logical vectors

```
> myLog <- myVec > 10
> myVec[myLog]
  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
```

Subsetting by field names

```
> myVec[c("B", "K", "M")]
 B K  M
 2 11 13
```

Calling a single column or list component by its name with the $ sign

```
> iris$Species[1:8]
[1] setosa setosa setosa setosa setosa setosa setosa setosa
Levels: setosa versicolor virginica
```

Thomas Girke

Combining Objects

The c function combines vectors and lists

```
> c(1, 2, 3)
[1] 1 2 3
> x <- 1:3; y <- 101:103
> c(x, y)
[1] 1 2 3 101 102 103
```

The cbind and rbind functions can be used to append columns and rows, respecively.

```
> ma <- cbind(x, y)
> ma xy
[1,] 1 101
[2,] 2 102
[3,] 3 103
> rbind(ma, ma)
     x   y
[1,] 1 101
[2,] 2 102
[3,] 3 103
[4,] 1 101
[5,] 2 102
[6,] 3 103
```

Accessing Name Slots and Dimensions of Objects

Length and dimension information of objects

```
> length(iris$Species)
[1] 150
> dim(iris)
[1] 150 5
```

Accessing row and column names of 2D objects

```
> rownames(iris)[1:8]
[1] "1" "2" "3" "4" "5" "6" "7" "8"
> colnames(iris)
[1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width" "Species"
```

Return name field of vectors and lists

```
> names(myVec)
[1] "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" "M" "N" "O" "P" "Q" "R" "S
> names(myL)
[1] "name" "wife" "no.children" "child.ages"
```

Sorting Objects

The function sort returns a vector in ascending or descending order

```
> sort(10:1)
[1] 1 2 3 4 5 6 7 8 9 10
```

The function order returns a sorting index for sorting an object

```
> sortindex <- order(iris[,1], decreasing = FALSE)
> sortindex[1:12]
[1]14 9 39 43 42 4 7 23 48 3 30 12
> iris[sortindex,][1:2,]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
14          4.3         3.0          1.1         0.1  setosa
9           4.4         2.9          1.4         0.2  setosa
> sortindex <- order(-iris[,1]) # Same as decreasing=TRUE
```

Sorting on multiple columns

```
> iris[order(iris$Sepal.Length, iris$Sepal.Width),][1:2,]
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
14          4.3         3.0          1.1         0.1  setosa
9           4.4         2.9          1.4         0.2  setosa
```

Basic Operators and Calculations

Comparison operators: ==, ! =, <, >, <=, >=

```
> 1==1
[1] TRUE
```

Logical operators: AND: &, OR: |, NOT: !

```
> x <- 1:10; y <- 10:1
>x>y&x>5
[1] FALSE FALSE FALSE FALSE FALSE TRUE TRUE TRUE TRUE TRUE
```

Calculations: to look up math functions, see Function Index:

http://cran.at.r-project.org/doc/manuals/R-intro.html#Function-and-variable-index

```
>x+y
[1] 11 11 11 11 11 11 11 11 11 11
> sum(x)
[1] 55
> mean(x)
[1] 5.5
> apply(iris[1:6,1:3], 1, mean)
123456 3.333333 3.100000 3.066667 3.066667 3.333333 3.666667
```

Some Great R Functions I

The unique() function to make vector entries unique

```
> length(iris$Sepal.Length)
[1] 150
> length(unique(iris$Sepal.Length))
[1] 35
```

The table() function counts the occurrences of entries

```
> table(iris$Species)
 setosa versicolor virginica
     50         50        50
```

The aggregate() function computes statistics of data aggregates

```
> aggregate(iris[,1:4], by=list(iris$Species), FUN=mean,
na.rm=TRUE)
      Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
1      setosa        5.006       3.428        1.462       0.246
2  versicolor        5.936       2.770        4.260       1.326
3   virginica        6.588       2.974        5.552       2.026
```

Some Great R Function II

The %in% function returns the intersect between two vectors

```
> month.name %in% c("May", "July")
[1] FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE FALSE FALSE FALSE FALSE
```

The merge() function joins two data frames by common field entries, here row-names (by.x=0). To obtain only the common rows, change all=TRUE to all=FALSE. To merge on specific columns, refer to them by their position numbers or their column names.

```
> frame1 <- iris[sample(1:length(iris[,1]), 30), ]
> frame1[1:2,]
Sepal.Length Sepal.Width Petal.Length Petal.Width Species
110 7.2 3.6 6.1 2.5 virginica
30  4.7 3.2 1.6 0.2 setosa
> dim(frame1)
[1] 30 5
> my_result <- merge(frame1, iris, by.x = 0, by.y = 0, all = TRUE)
> dim(my_result)
[1] 150 11
```

R Programming: A refresher

Programming: `if` statements

Syntax

```
>if(cond1=true) { cmd1 } else
{ cmd2 }
```

`if` statement

```
>if(1==0) {
      print(1)
  } else {
      print(2)
  }
  [1] 2
```

Programming: `for` loops

Syntax

```
> for(variable in sequence) {
     statements
 }
```

For Loop with Condition

```
>x <- 1:10
>z <- NULL
>for(i in seq(along=x)) {
    if(x[i] < 5) {
        z <- c(z, x[i] - 1)
    } else {
        z <- c(z, x[i] / x[i])
    }
 }
>z
[1] 0 1 2 3 1 1 1 1 1 1
```

Programming: `while` loops

Syntax

```
> while(condition) statements
```

`while` Loop

```
>z <- 0
>while(z < 5) {
     z <- z + 2
     print(z)
 }
 [1] 2
 [1] 4
 [1] 6
```

Programming: `apply` loops

Syntax

```
> apply(X, MARGIN, FUN, ARGs)
```

Example for applying predefined mean function

```
>apply(iris[,1:3], 1, mean)
[1] 3.333333 3.100000 3.066667 3.066667 3.333333 3.666667
3.133333 3.300000
...
```

With custom function

```
>x <- 1:10
>test <- function(x) { # Defines some custom function
    if(x < 5) {
        x-1
    } else {
        x / x
    }
 }
```

Returns same result as previous for loop

```
>apply(as.matrix(x), 1, test)
[1] 0 1 2 3 1 1 1 1 1 1
```

Same as above but with a single line of code

```
apply(as.matrix(x), 1, function(x) { if (x<5) { x-1 } else
{ x/x } })
[1] 0 1 2 3 1 1 1 1 1 1
```

Thomas Girke

Programming: `tapply` loops

Syntax

```
>tapply(vector, factor, FUN)
```

Computes mean values of vector aggregates defined by factor

```
>tapply(as.vector(iris[,i]), factor(iris[,5]), mean)
>setosa versicolor  virginica
 0.246      1.326      2.026
```

The aggregate function provides related utilities

```
>aggregate(iris[,1:4], list(iris$Species), mean)
     Group.1 Sepal.Length Sepal.Width Petal.Length Petal.Width
1     setosa        5.006       3.428        1.462       0.246
2 versicolor        5.936       2.770        4.260       1.326
3  virginica        6.588       2.974        5.552       2.026
```

For Vectors and Lists: `lapply` and `sapply`

Syntax

```
lapply(X, FUN)
sapply(X, FUN)
```

Creates a sample list

```
>mylist <- as.list(iris[1:3,1:3])
>mylist
 $Sepal.Length
 [1] 5.1 4.9 4.7

 $Sepal.Width
 [1] 3.5 3.0 3.2

 $Petal.Length
 [1] 1.4 1.4 1.3
```

Compute sum of each list component and return result as list

```
>lapply(mylist, sum)
 $Sepal.Length
 [1] 14.7

 $Sepal.Width
 [1] 9.7

 $Petal.Length
 [1] 4.1
```

Compute sum of each list component and return result as vector

```
>sapply(mylist, sum)
Sepal.Length  Sepal.Width Petal.Length
        14.7          9.7          4.1
```

Thomas Girke

Acknowledgements

Bioinformatics Core Facility (Cambridge)
Thomas Carroll
Sarah Dawson
Mark Dunning
Silvia Halim
Suraj Menon
Rory Stark
Sarah Vowler
Matthew Eldridge

Computational Biology Group (Cambridge)
Nuno Barbosa-Morais
Natalie Thorne
Matt Ritchie
Andy Lynch
Christina Curtis
Benilton Carvalho
Oscar Rueda

Genomics Core Facility (Cambridge)
Michelle Osborne
Sarah Leigh-Brown
Hannah Haydon
Claire Fielding
Fatimah Madni
James Hadfield

Sanger Institute (Cambridge)
Natasha Karp

Stem Cell Institute (Cambridge)
Jelena Aleksic

UNIVERSITY OF CAMBRIDGE

MRC Medical Research Council

CANCER RESEARCH UK