ChIP-seq data analysis

04-05-12

Outlook

- Friday 04-05-12:
  - Next-generation sequencing
  - ChIP-seq
    - experimental design
  - ChIP-seq data analysis:
    - Mapping of sequenced reads to a reference geneome
    - Peak calling
    - Peak annotation
    - Discovery of transcrption factors sequence motifs
- Friday 11-05-12
  - Practical: ChIP-seq data analysis

Next generation sequencing course, 12th-14th March 2012

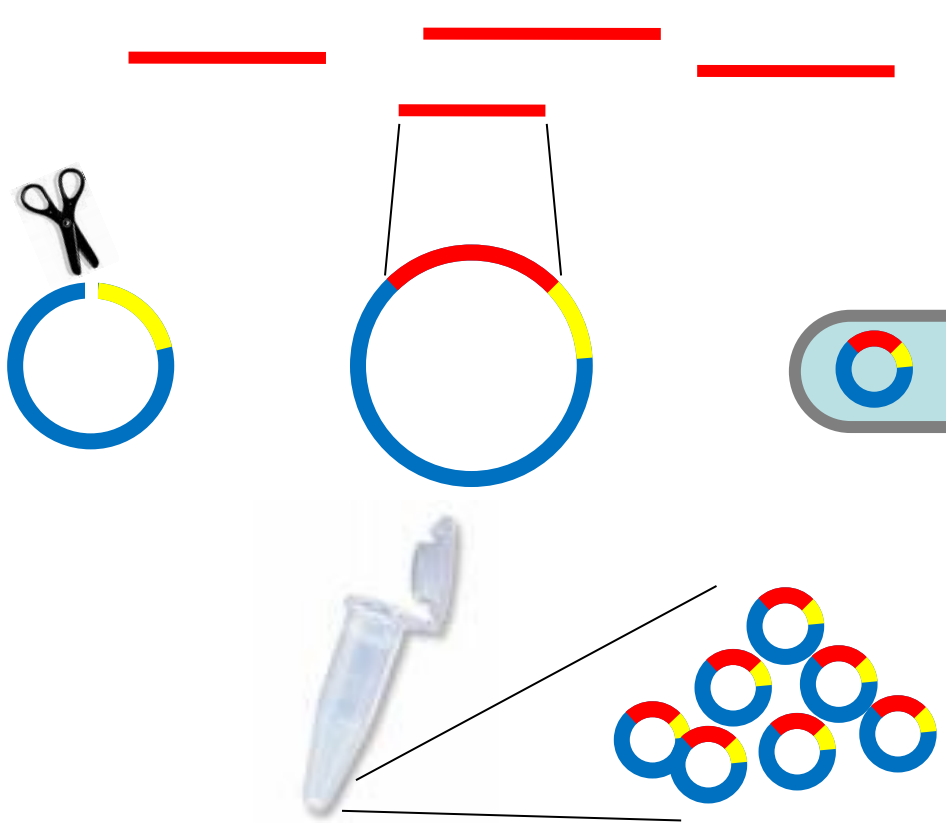
Harrold swerdlow, Head of R&D, WTSI
Remco loos and Myrto Kostadima, from EBI

wellcome trust sanger institute

Next-gen Rationale

Harrold swerdlow slide

Capillary Sample Prep



Fragment genome

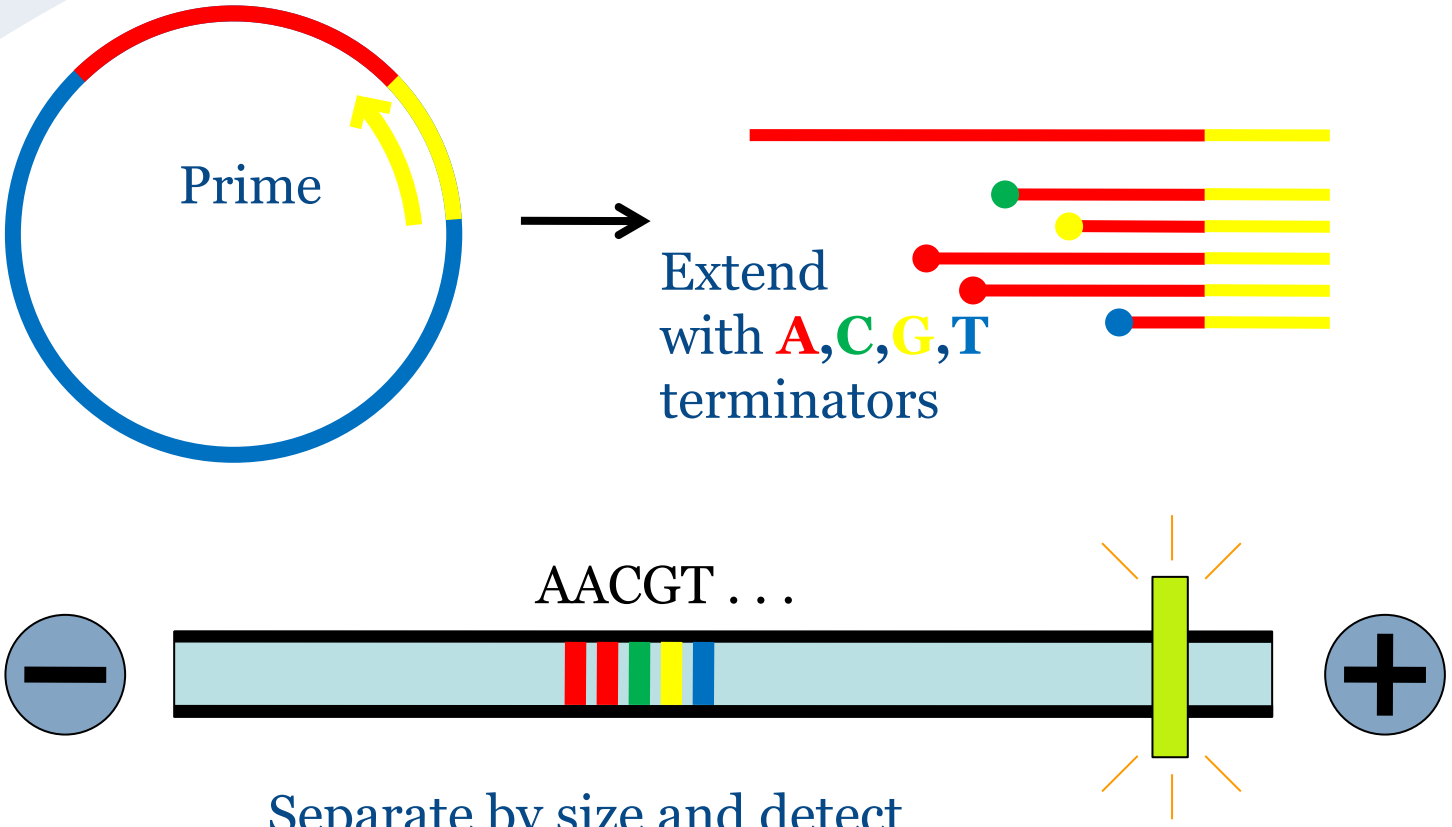
Clone into bacterial vector

Grow and purify

Harrold swerdlow slide

wellcome trust sanger institute

Capillary Sequencing



Harrold swerdlow slide

Capillary Reactions

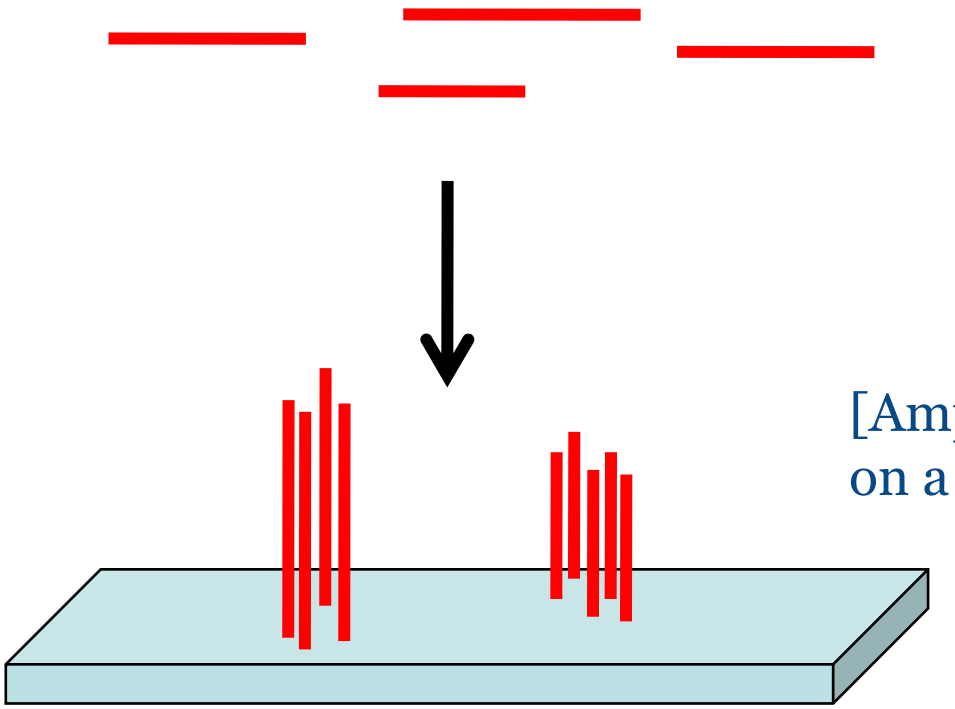
1 tube
1 template

→

1 capillary
1000 bases

Harrold swerdlow slide

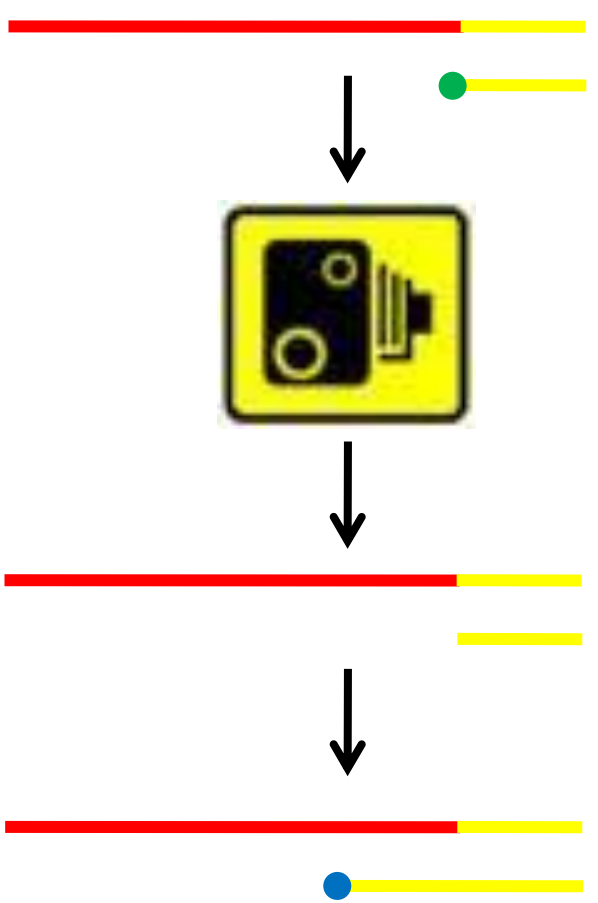
Next-Generation Sample Prep



[Amplify] fragments directly on a surface (bead, chip, etc.)

Harrold swerdlow slide

Sequencing by Synthesis



Extend by 1 base

Image

Reverse termination

Repeat

Harrold swerdlow slide

wellcome trust sanger institute

Next-Generation Reactions

1 feature
1 template

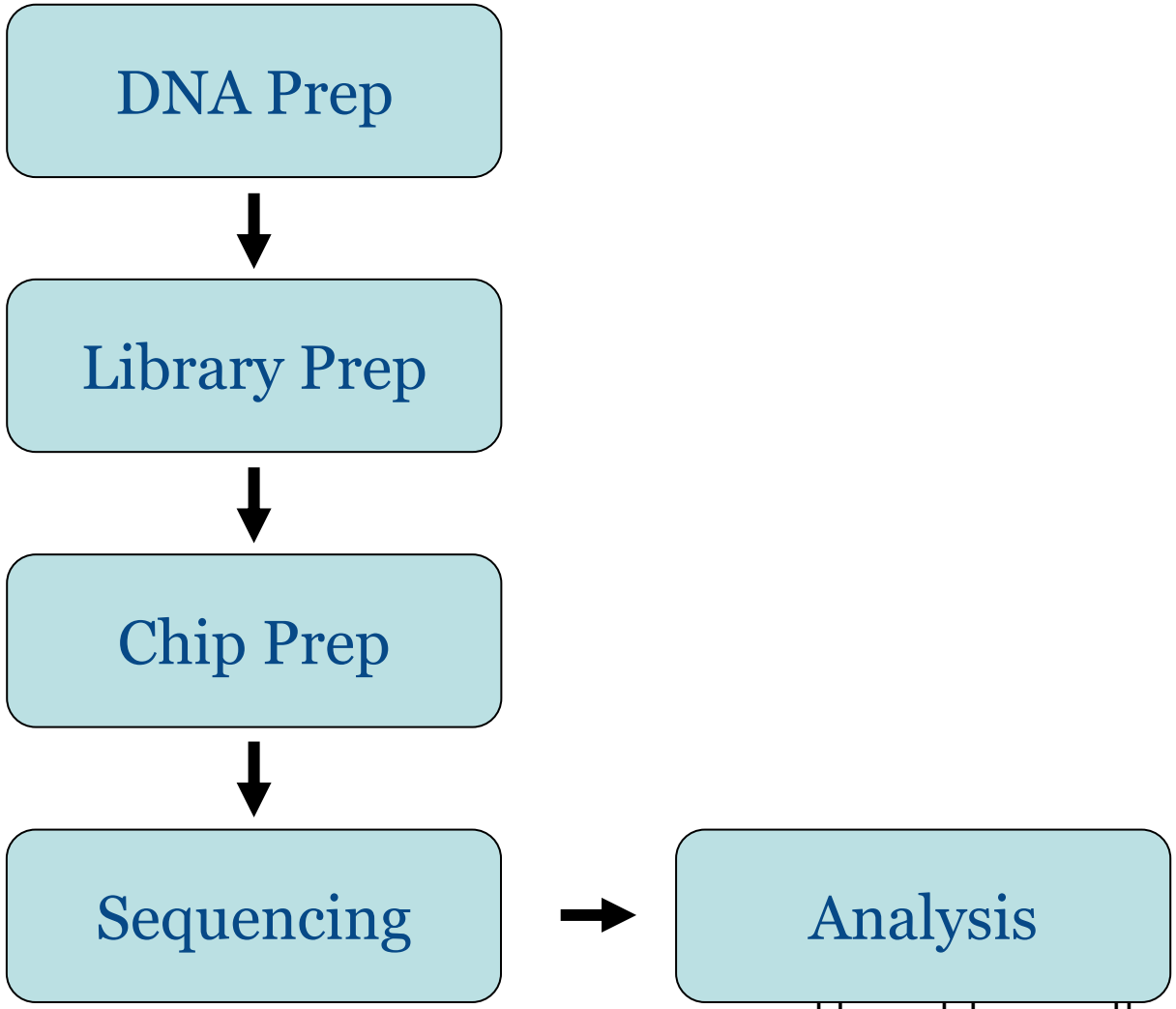
→

1 chip
gigabases

Harrold swerdlow slide

wellcome trust sanger institute

The Next-Generation Process



Harrold swerdlow slide

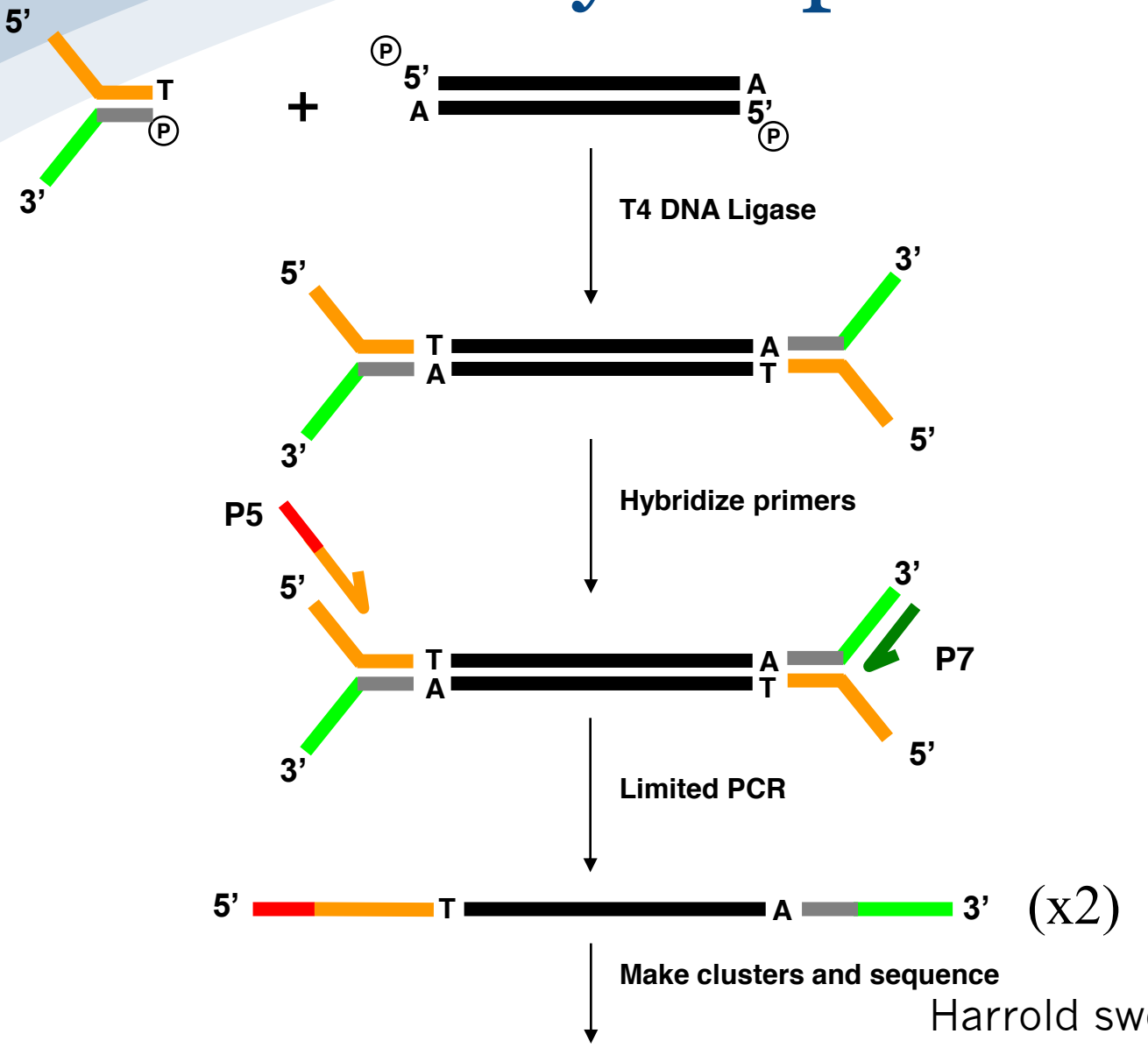
wellcome trust sanger institute

Illumina Technology

Harrold swerdlow slide

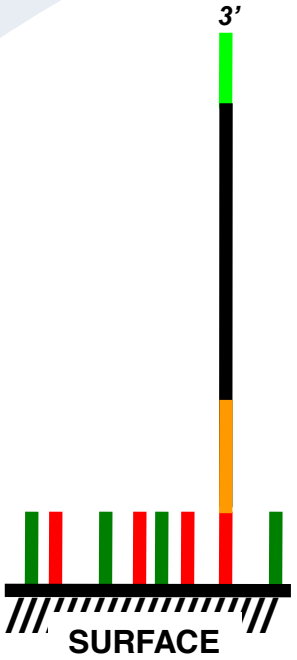
wellcome trust sanger institute

Library Prep

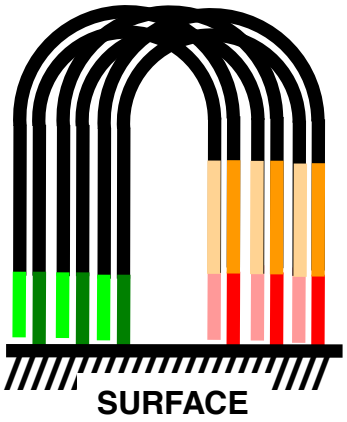


Harrold swerdlow slide

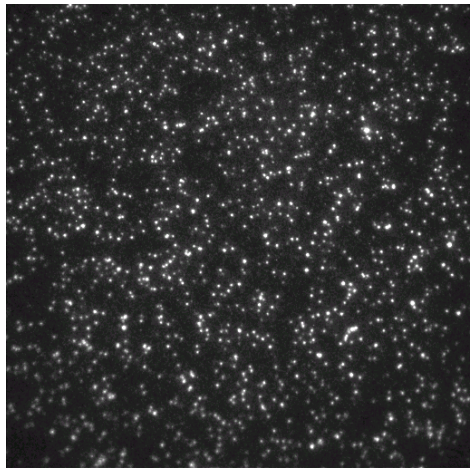
Cluster Amplification



Single-molecule array



Cluster ~1000 molecules



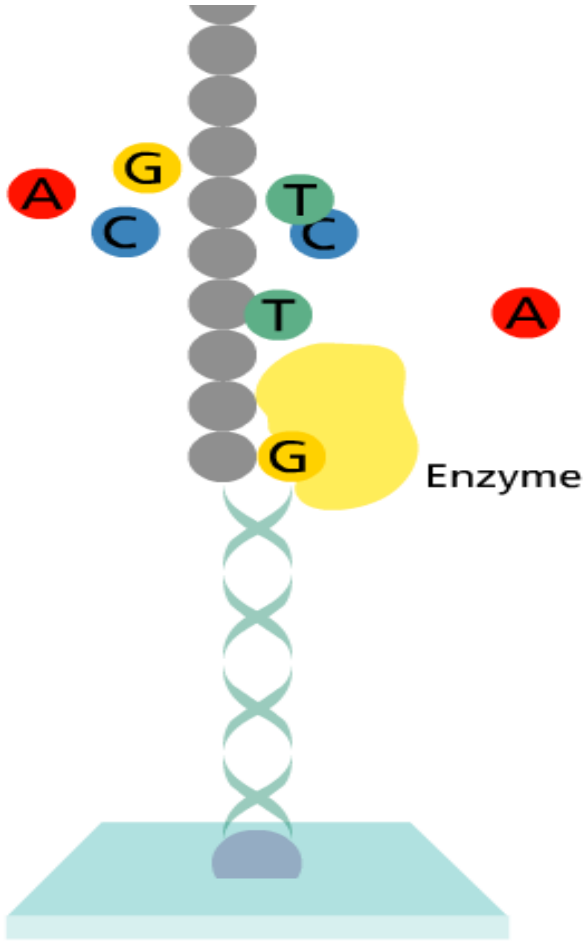
1 billion clusters on a single glass chip

Harrold swerdlow slide

wellcome trust sanger institute

Sequencing by Synthesis

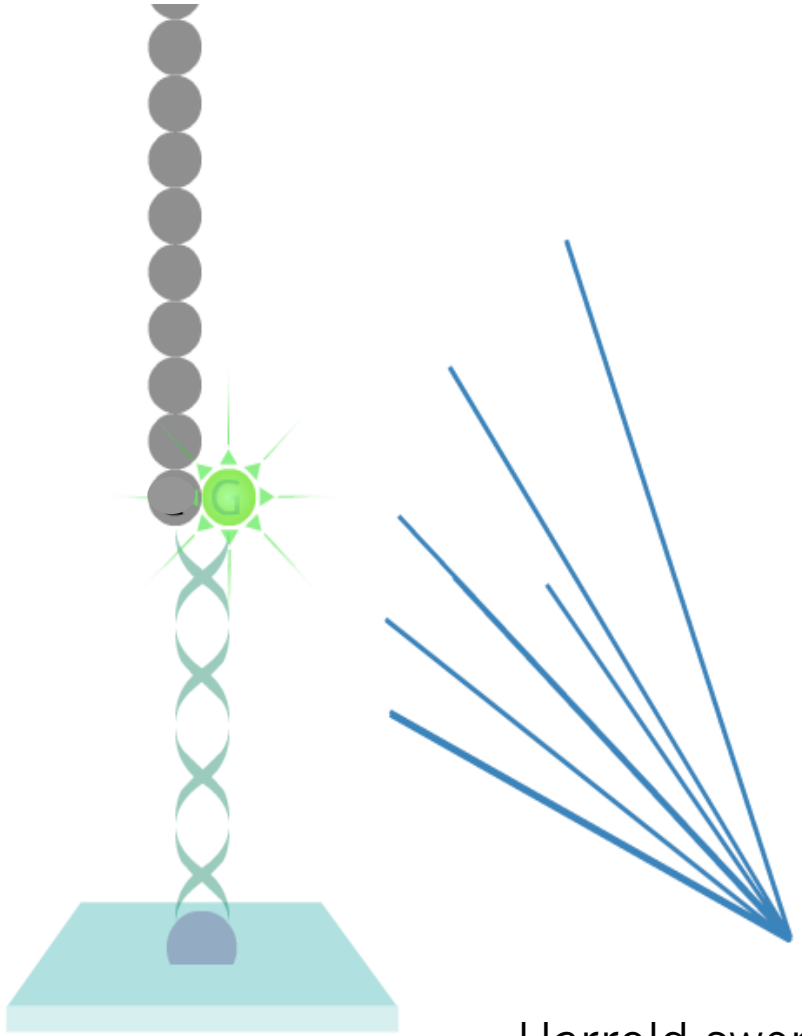
Cycle 1



Harrold swerdlow slide

Wash + Detect Fluorescence

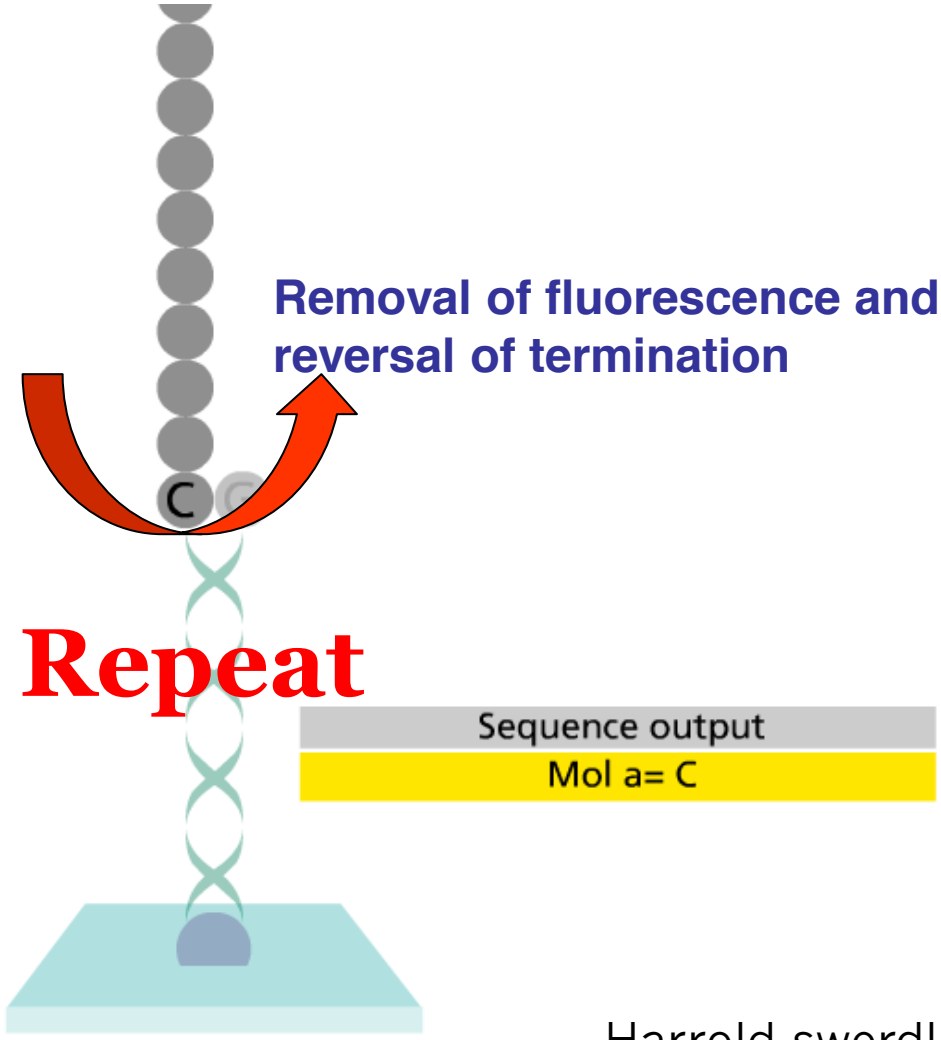
Cycle 1



Harrold swerdlow slide

Prepare for Next Cycle

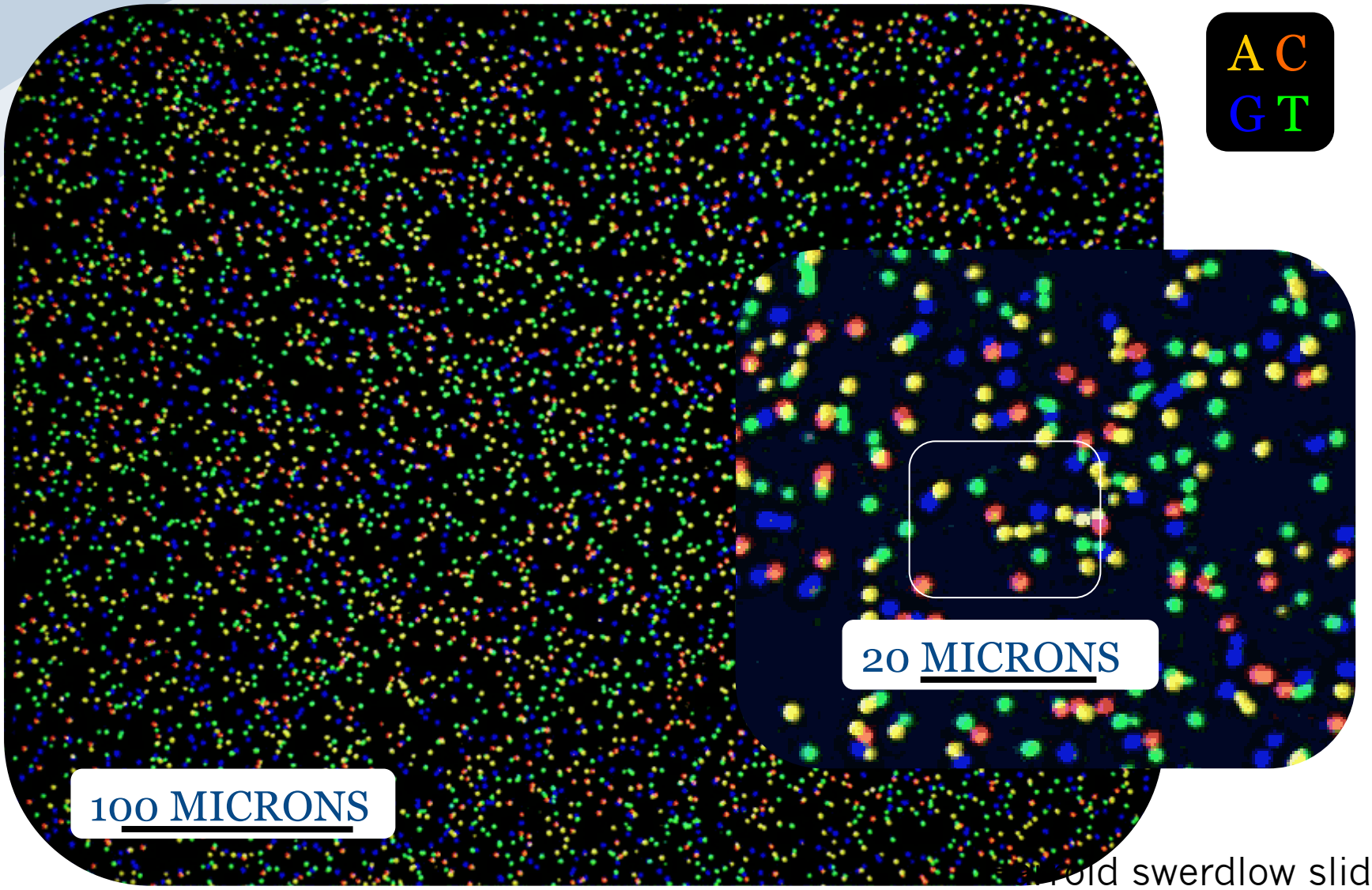
Cycle 1



Harrold swerdlow slide

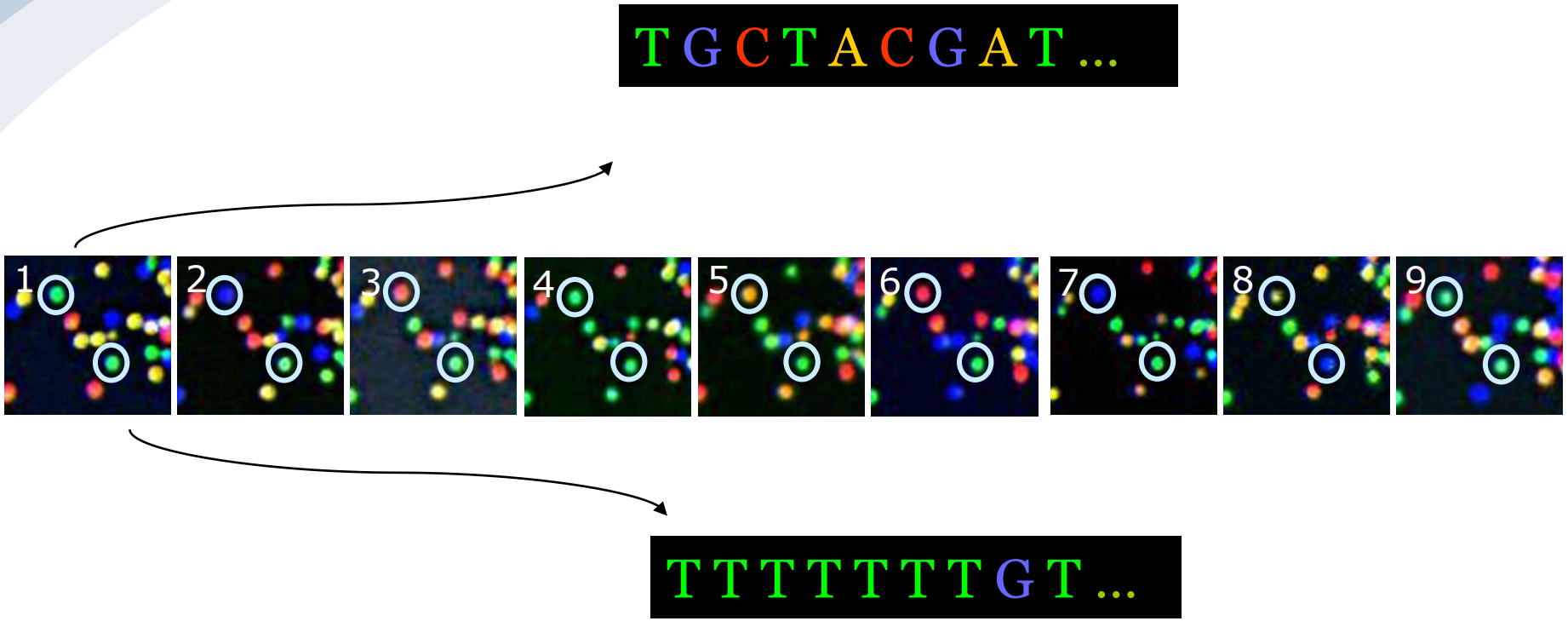
wellcome trust sanger institute

Four Colour Composite



old swerdlow slide

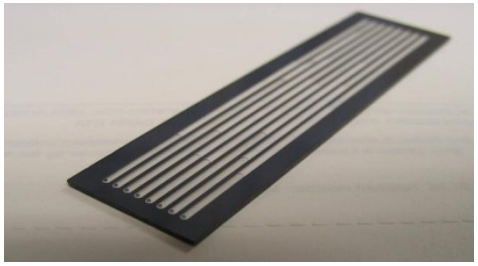
Base Calling From Raw Data



Harrold swerdlow slide

wellcome trust sanger institute

Billions of Bases of DNA Sequence (per instrument)



- 8 lanes per chip
- 48 tiles (6 swaths) per lane
- 4,000,000 clusters per tile
- 200 cycles (2 x 100) in 10 days
- 8 x 48 x 4,000,000 x 200 = 300 Gb
- 2 chips = 600 Gb / run = 6 Genomes

Harrold swerdlow slide

- Illumina solexa sequencing video !

Next-generation sequencing applications

- Genome applications:
  - ChIP-seq:TF binding sites, histone modifications, nucleosome positions mapping
  - Dnase-seq: DNA accessibility,
  - Methyl-seq: methylome characterisation
  - Variant discovery:SNPs,
  - De novo genome assembly

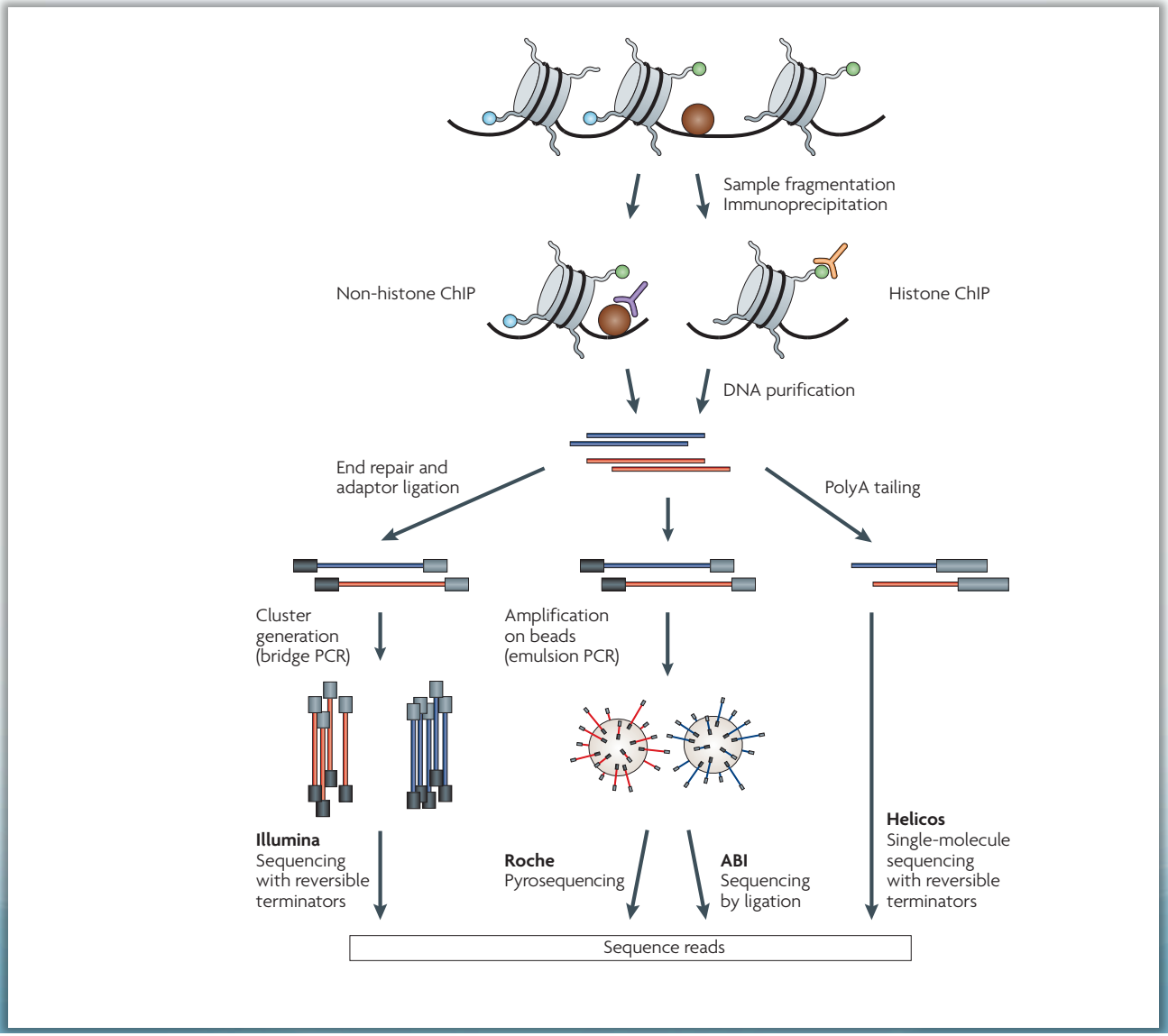
- Transcriptome applications:
  - Quantification of gene Expression
  - Differential gene expression
  - De novo transcript dicovery
  - Detection of abberant transcripts

ChIP-chip vs ChIP-seq

	ChIP-chip	ChIP-seq
Resolution	Array-specific	High - single nucleotide
Coverage	Limited by sequences on the array	Limited by “alignability” of reads to the genome, increases with read length
Repeat elements	Masked out	Many can be covered (40% of human genome is repetitive but 80% is uniquely mappable)
Cost	400-800$ per array (1-6M probes), multiple arrays needed for human genome	Around 1000$ per lane; 20-30M reads
Source of noise	Cross hybridization	Sequencing bias, GC bias, sequencing error
Amount of ChIP DNA required	High, few micrograms	Low 10-50ng
Dynamic range	Lower detection limit and saturation at high signal	Not limited
Multiplexing	Not possible	Possible

Remco loos slides

Overview of ChIP-seq experiments



Park J 2009, Nature Reviews, Genetics

ChIP-seq experimental design

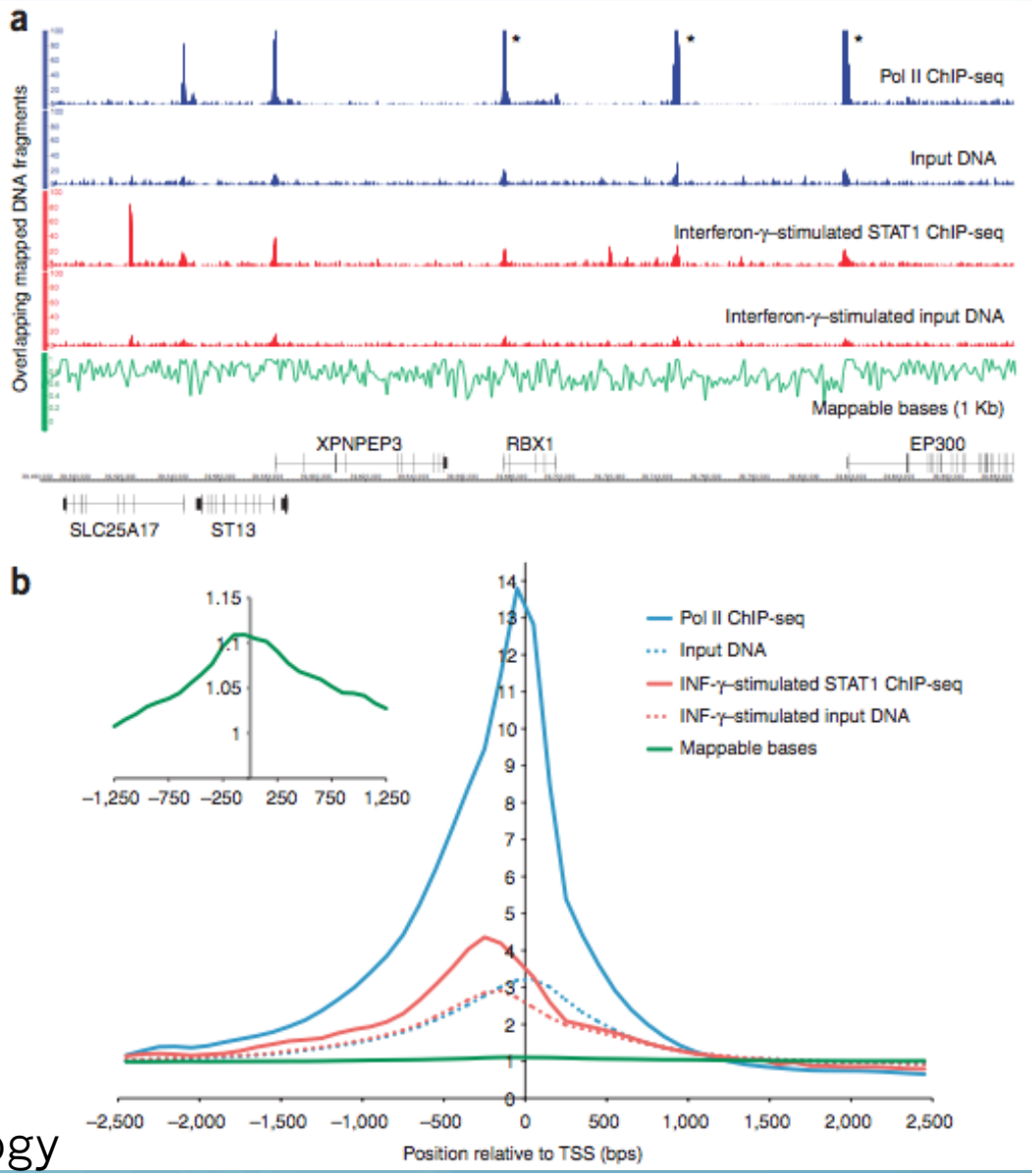
- Antibody quality
- Control experiment
- Depth of sequencing
- Multiplexing
- Sequencing options:
  - Paired-end or single-end reads
  - 36bp reads or longer

Antibody quality

- A sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason fo rfailed ChIP- seq experiments
  - Check your antibody ahead if possible. Western blotting to check the cross-reactivity of the antibody

Control experiment

- A ChIP-seq peak should be compared with the same region in a matched control
- Open chromatin regions are fragmented more easily than closed regions
- There is amplification and size selection bias during library preparation
- Repetitive sequences might seem to be enriched (inaccurate repeats copy number in the assembled genome)



Rozowski 2009, nature Biotechnology

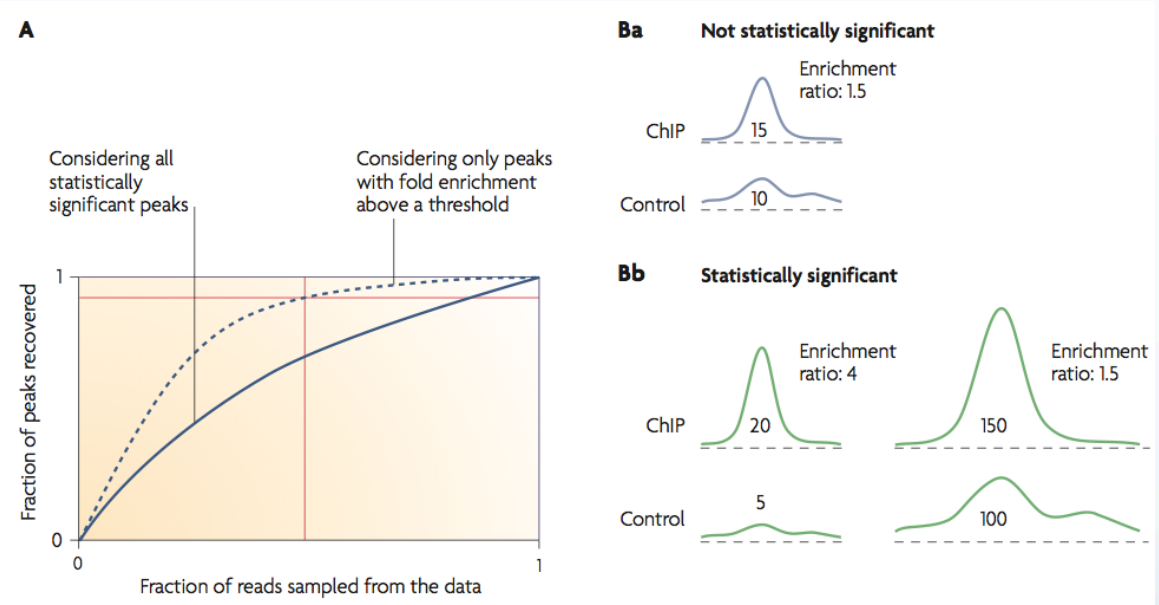
Control type

- Input DNA
- Mock IP - DNA obtained from IP without antibody
  - Very little material can be pulled down leading to inconsistent results of multiple mock IPs.
- Nonspecific IP - using an antibody against a protein that is not known to be involved in DNA binding
- There is no consensus on which is the most appropriate
- Sequencing a control can be avoided when looking at:
  - time points
  - differential binding pattern between conditions

Depth of sequencing

More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth

Number of putative target regions continues to increase significantly as a function of sequencing depth



Park J 2009, Nature Reviews, Genetics

With current sequencing technologies, one lane is usually sufficient

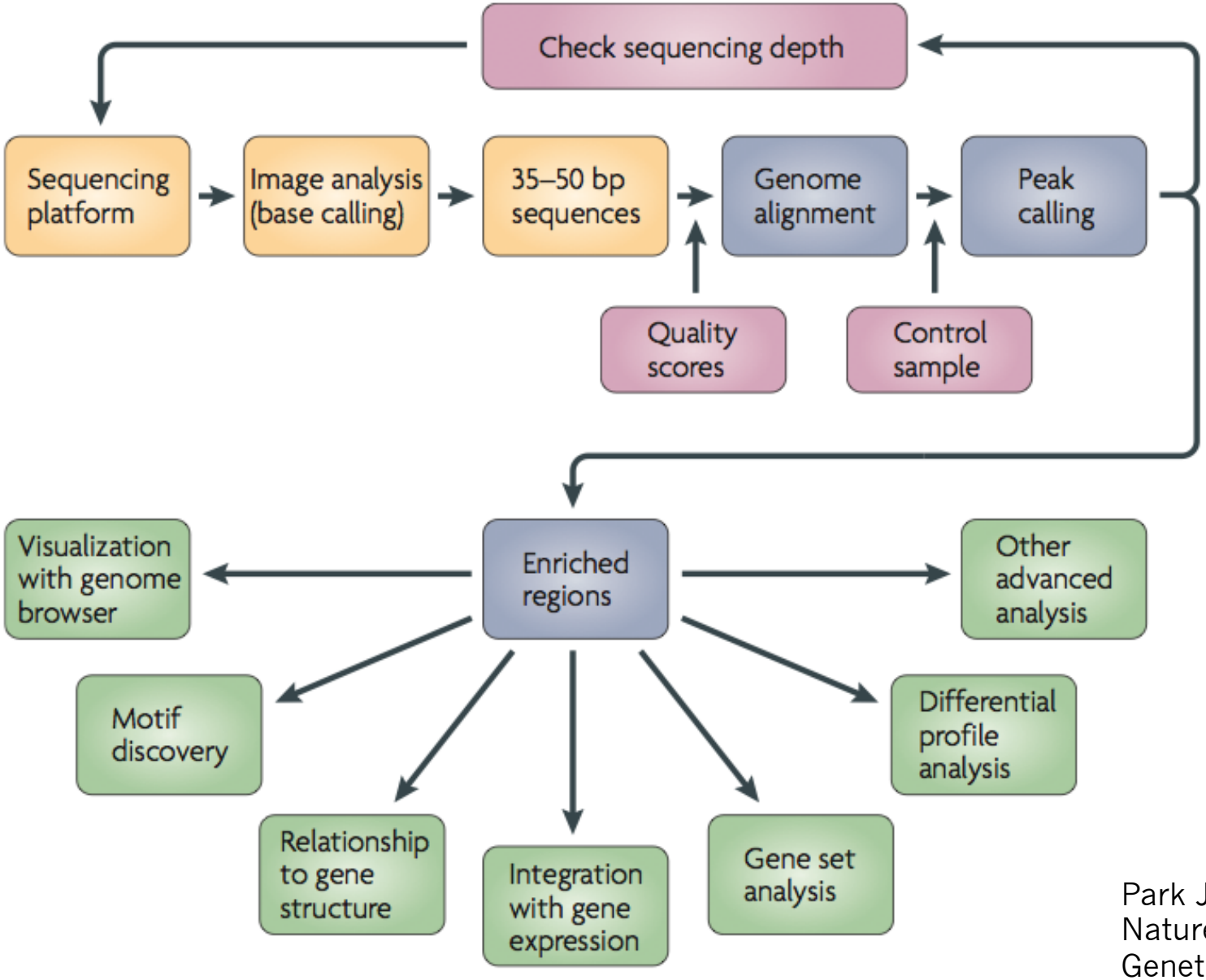
Saturation-MACS « diag » table

FC	# peaks	90%	80%	70%	60%	50%	40%	30%	20%
0-20	31530	75.01	55.98	39.58	26.01	15.35	7.43	2.64	0.51
20-40	5481	99.62	97.7	92.52	80.46	61.34	36.75	14.61	2.81
40-60	235	100	100	100	100	99.57	90.21	68.51	28.09
60-80	40	100	100	100	100	100	100	95	62.5
80-100	7	100	100	100	100	100	100	100	85.71
100-120	2	100	100	100	100	100	100	100	100
120-140	5	100	100	100	100	100	100	100	100
160-180	1	100	100	100	100	100	100	100	100

Sequencing options

- Pared-ends vs single-end:
  - DNA fragements are sequenced from both ends
  - Costs twice as mutch as single end sequencing
  - Increase « mappability » of reads specially in repetitive regions
  - For ChIP-seq, usually not worth the extra cost, unless you have a specific interest in repeat regions
- Short vs long reads:
  - For ChIP-seq of 36 bp single-end reads are sufficient

Overview of ChIP-seq analysis



Park J 2009, Nature Reviews, Genetics

Raw reads-fastq file

@HWI-EAS225_30EJMAAXX:6:1:1300:1234
GAAAATCACGGAAAATGAGAAATACACACTTTAGGA
+
;;;;;;;;;;;;;;;;;;;;;;;;;;;;888666 @HWI-EAS225_30EJMAAXX:
6:1:330:1573
GGATACAACAGAAGATCTCGGGAACGGACTCAGAAG
+
;;;;;;;;;;;;;;;1;;;;;;;1;;;;;488884 @HWI-EAS225_30EJMAAXX:
6:1:1079:806
GGCTTAGTAGTCCACCCTGGAGTTATGGATTGTGAA
+
;;48;4;84.4;;47;8;887;;49;;.4;8.1&8+ @HWI-
EAS225_30EJMAAXX:6:1:1775:216
GTTCAAGGTCACAGGAGATCCTGTCTCAAAACCACC
+
;88;;48;.;;;8;2;4;;;44;8)8;4+4++%8.4 @HWI-
EAS225_30EJMAAXX:6:1:703:1984
GAAGGTCTTCTCAGCCACGCCCCTGCCTCCTGCTCC
+
;;;;;;;;;;;;;;;;;;;;;;;;6;;7887876 @HWI-EAS225_30EJMAAXX:
6:1:1109:1520
GTGAGATGTTCAGGTAGAGACTAATGTAAGCGGTGA
+
;;;;;;;;;;;;;7:;;;;64;::;1;:::786716 @HWI-EAS225_30EJMAAXX:
6:1:999:1416
GTTAGACGCAGCTCATTAGGGAAAAACCTATCCCAT
+
;;;;;;.;;;;;;;;;;;;;;1;;;;(9;;866886

Remco loos slides

Fasq format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

6 - Flowcell lane

73 - Tile number

941,1973 - 'x','y'-coordinates of the cluster within the tile

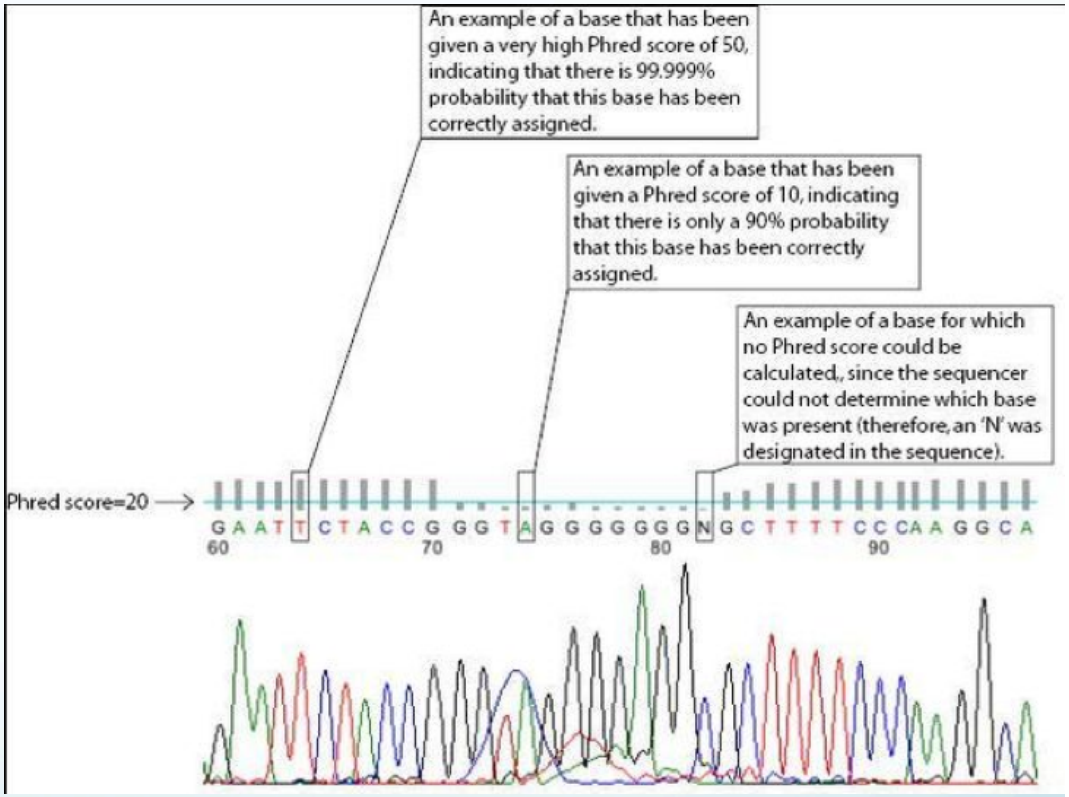
#0 - index number for a multiplexed sample (0 for no indexing)

/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Remco loos slides

Phred quality score

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %



Wikipedia

A Phred score of a base:
Q phred = -10 * log10($e)
where $e is the estimated probability of a base being wrong.
For example: If a base is estimated to have a 0.1% chance of being wrong, it gets a Phred score of 30

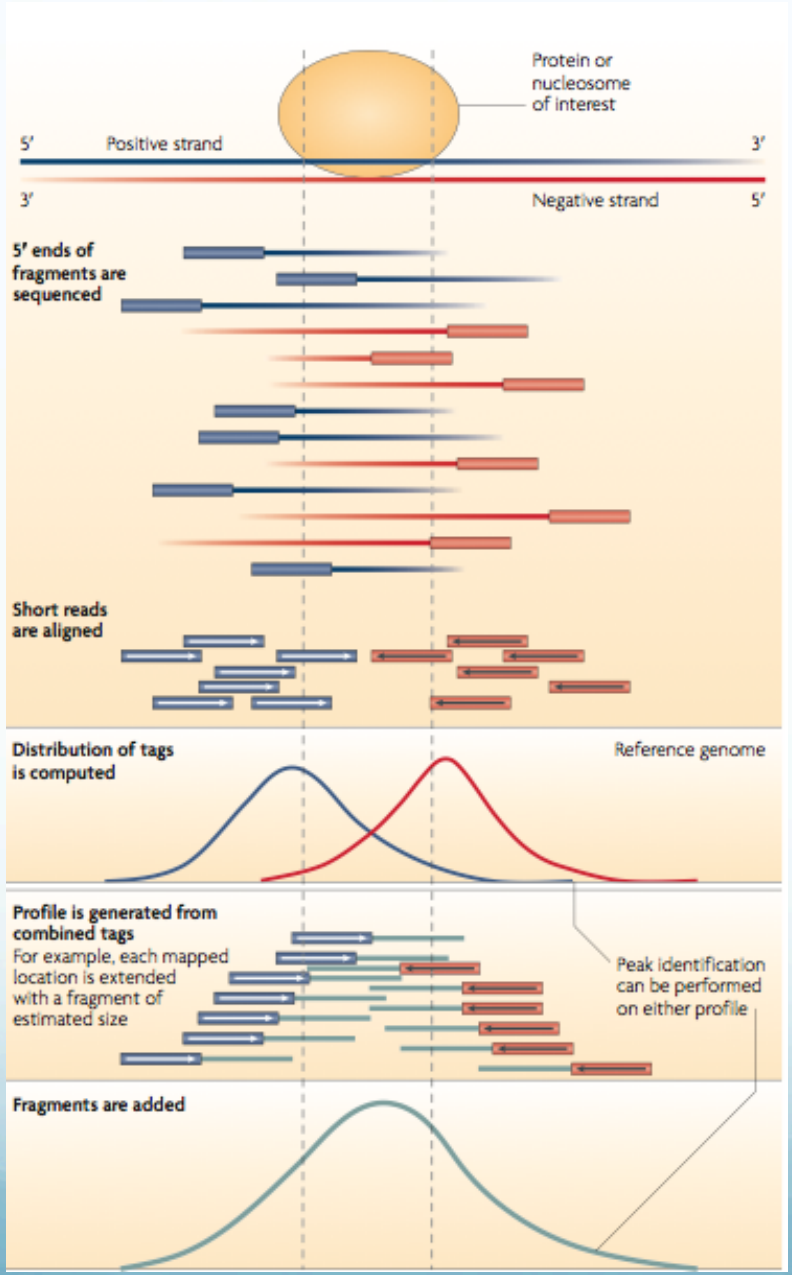
Mapping of sequenced reads

- ELAND-provided with Illumina sequencer
  - Limited reads length
  - Allow 2 substitutions
- MAQ
  - Uses quality values
  - Integrate consensus calling
- Bowtie
  - Ultrafast
  - Can work on workstations with < 2 Gb memory
- Many others: BWA, Novoalign, BFAST,...

Mapping challenges

- Enormous amount of short reads against large genomes
- Presence of repetitive regions, pseudogenes
- Mismatches:
  - Allow or not
  - SNP or sequencing errors
  - Insertion/deletion
- Multipe reads: reads that map to more than one genomic location
- Software challeges:
  - Balance between speed, precision and memory usage

Strand specific profile at enriched sites

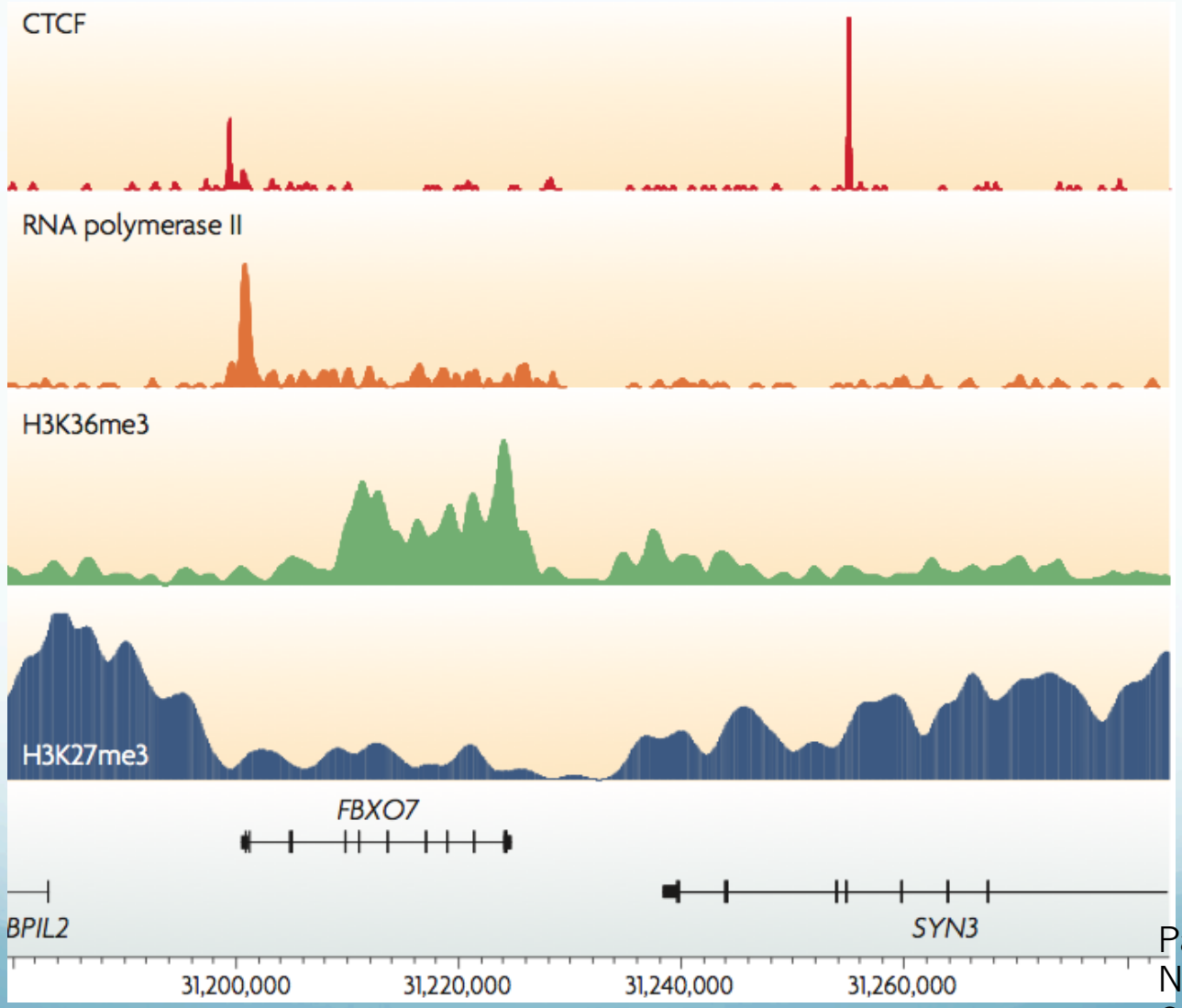


Park J 2009, Nature Reviews, Genetics

Peak calling

- CisGenome:
  - Peak criteria: number of reads in windows and number ChIP read minus control reads
- ERANGE:
  - High quality peak estimate
- MACS:
  - Poisson P value estimate
- Many others: FindPeaks, QuEST…

Peak calling-Challenges



Park J 2009, Nature Reviews, Genetics

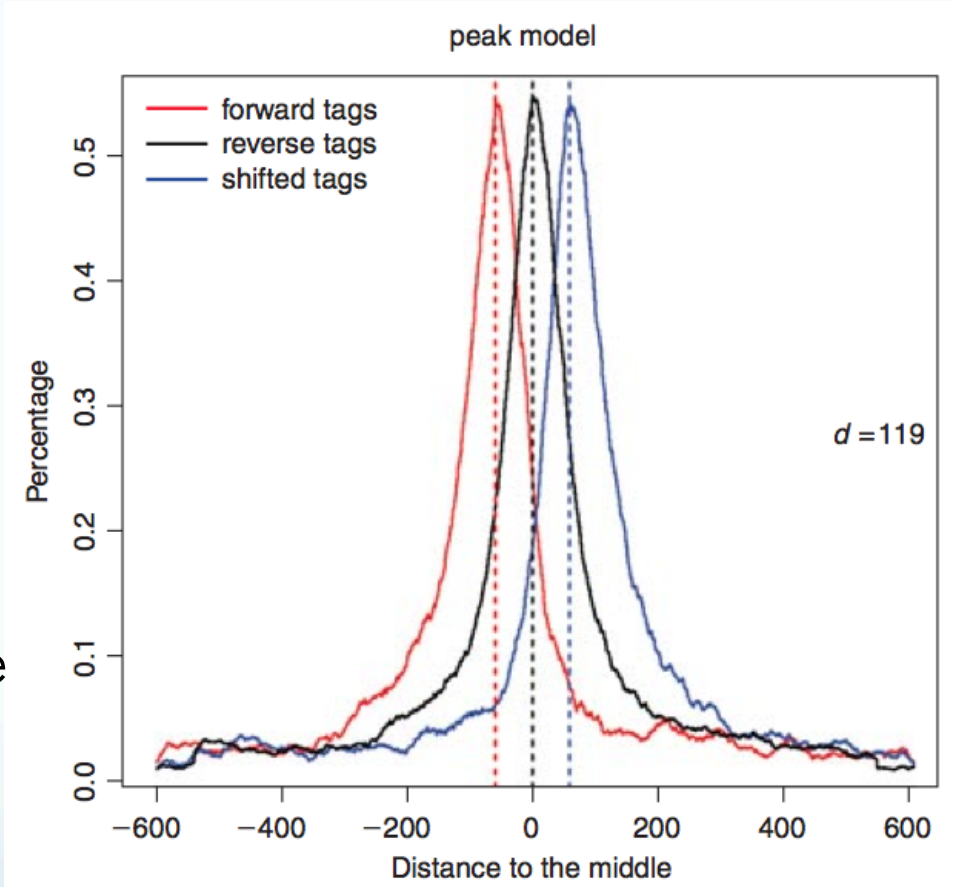
MACS tool

- Model the shift size between +/- strand tags:

- Scan the genome to find regions With tags more than mfold enriched Relative to random tag distribution

- Randomly sample 1000 of these (high quality peaks) and calculate the distance between the modes of their +/- peaks

- Shift all the tags by d/2 toward the 3' end



Feng 2011
Current protocols
in bioinformatics

Analysis downstream to peak calling

- Visualization - genome browser: Ensembl, UCSC, IGB
- Peak Annotation - finding interesting features surrounding peak regions: PeakAnalyzer
- Correlation with expression data
- Discovery of binding sequence motifs
  - Split peaks
  - Fetch summit sequences
  - Run motif prediction tool
- Gene Ontology analysis on genes that bind the same factor or have the same modification
- Correlation with SNP data to find allele-specific binding

Tools to install for the next session

- Bowtie (http://sourceforge.net/projects/bowtie-bio/files/latest/download)
- MACS (http://liulab.dfci.harvard.edu/MACS/index.html)
- PeakAnalyser (avalable at http://www.ebi.ac.uk/bertone/software)
- Java (http://www.java.com/fr/)