ChIP-seq

NGS Course

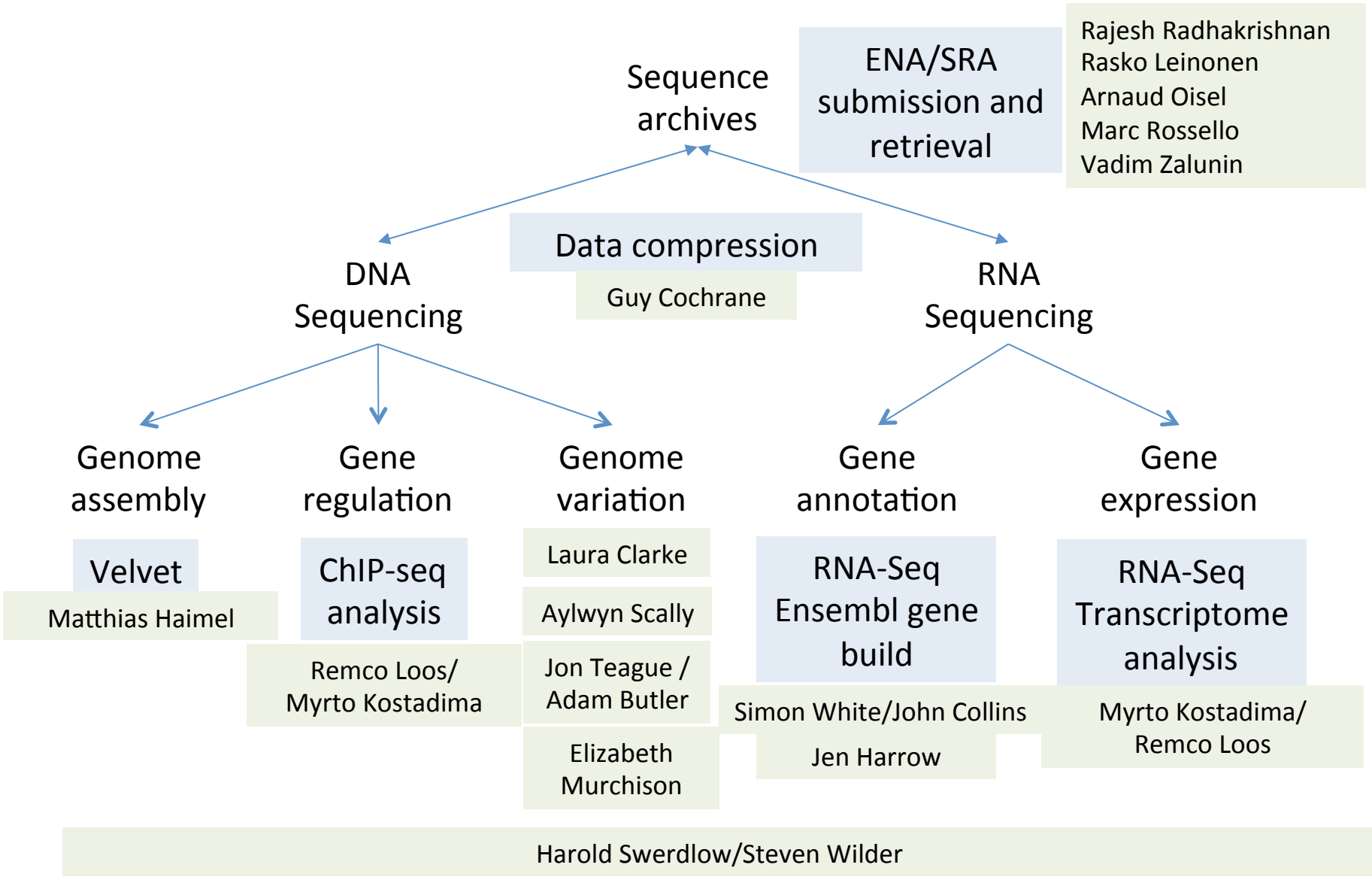
14th March 2012

Remco Loos

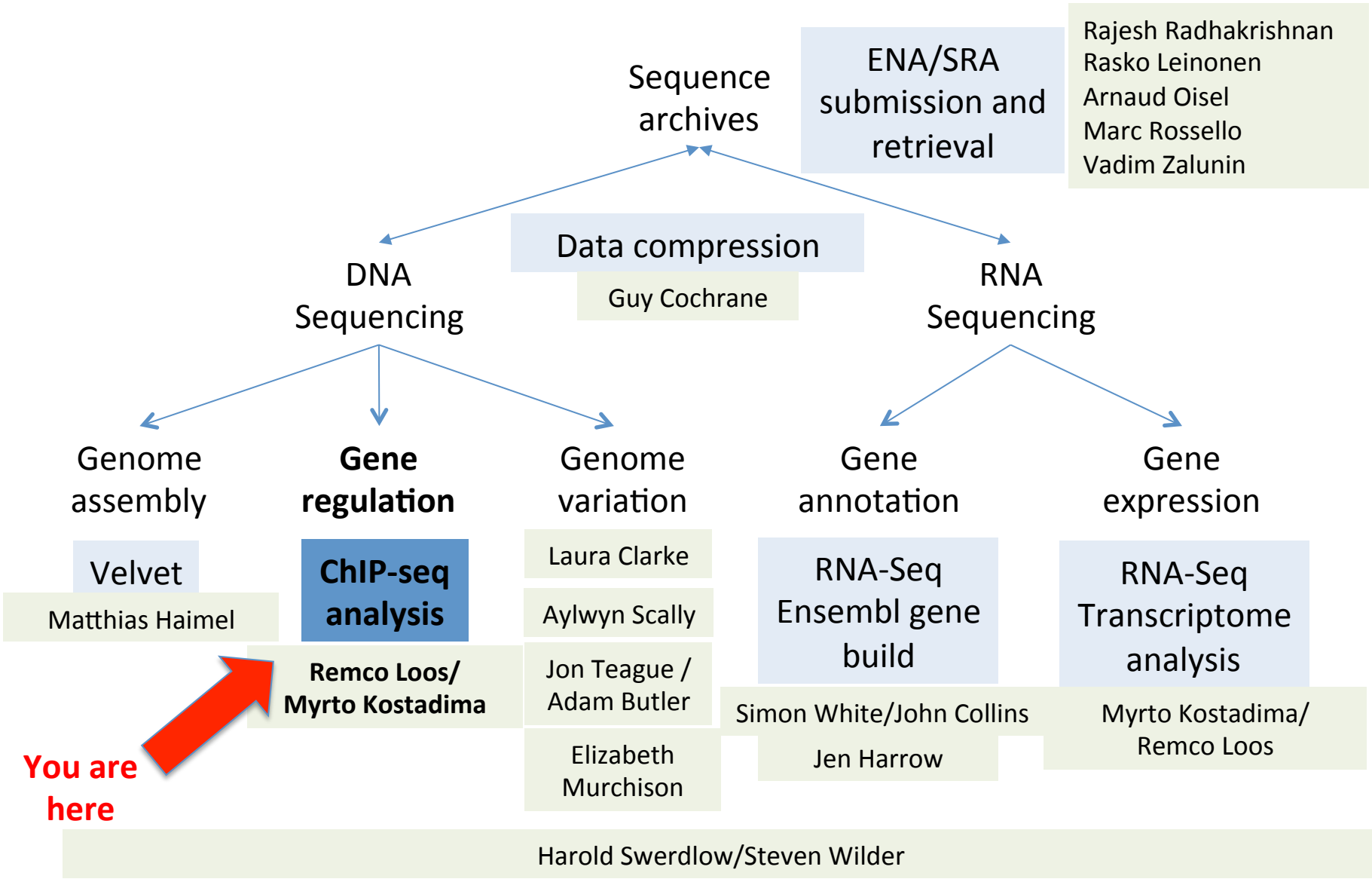
Myrto Kostadima

EMBL-EBI

Next generation sequencing course



Next generation sequencing course

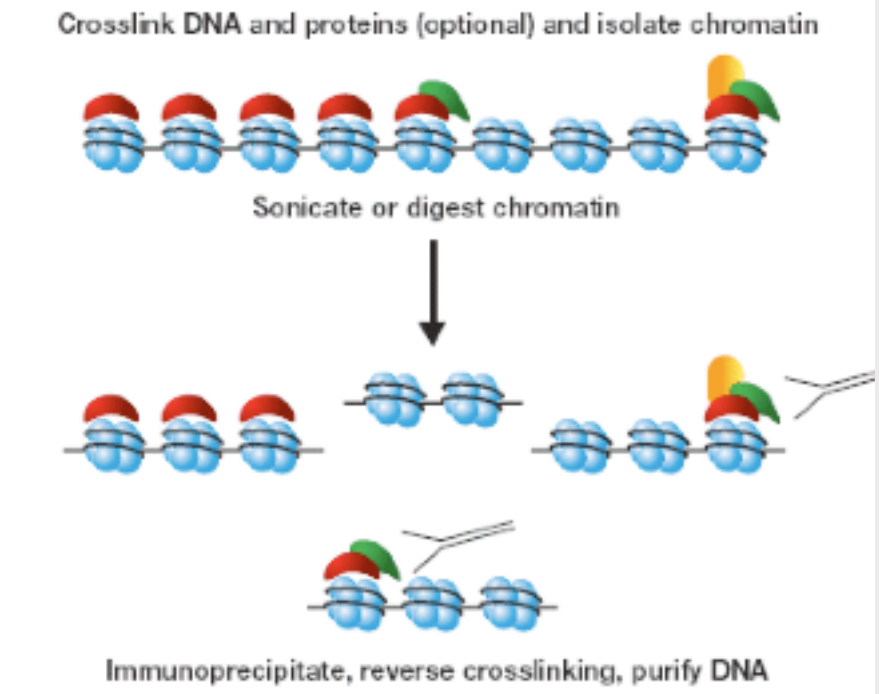


ChIP-seq Applications

- Protein-DNA interaction: Transcription factor binding locations, core transcriptional machinery
- Nucleosome positioning, histone modifications
- Chromatin states, DNA methylation

Chromatin ImmunoPrecipitation

- Transcription factors
- Nuclear receptors
- Polymerase
- Histones



- PCR with gene specific primers
- Hybridization on microarrays
- Sequencing

Lab procedures

	Transcription factor binding location	Map nucleosome positions or histone modifications
Crosslinking	Formaldehyde	Usually not
Fragmentation	Sonication (200-600bp)	MNase treatment
Immunoprecipitation	Antibody specific to protein of interest	Antibody specific for histone modification

Library construction

- Size selection ~150-300bp
- Adapter ligation
- Cluster generation (amplification)
- Sequence by synthesis

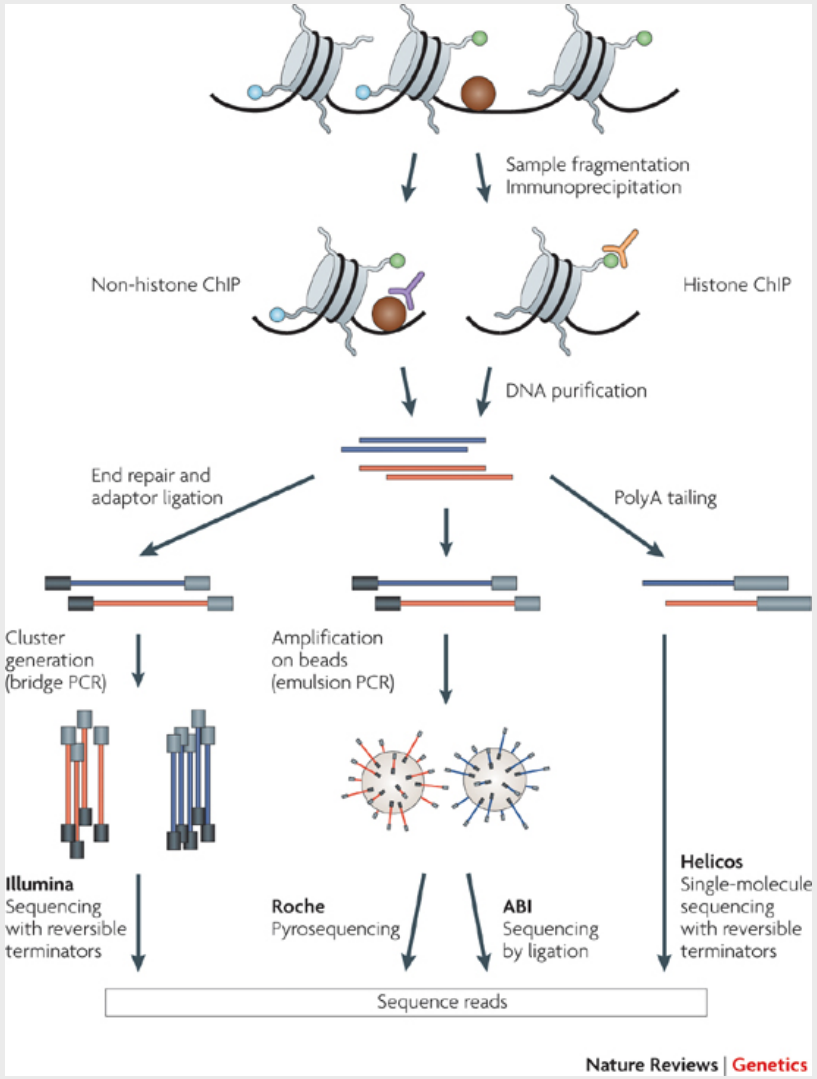
Chromatin ImmunoPrecipitation followed by Sequencing

- One of the early applications of NGS
- First studies published in 2007
  - Johnson et al (Science) - NRSF
  - Barski et al (Cell) - histone methylation
  - Robertson et al (Nature Methods) - STAT1
  - Mikkelsen et al (Nature) - histone modification
- Over 500 publications currently in PubMed

Historical slide: 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

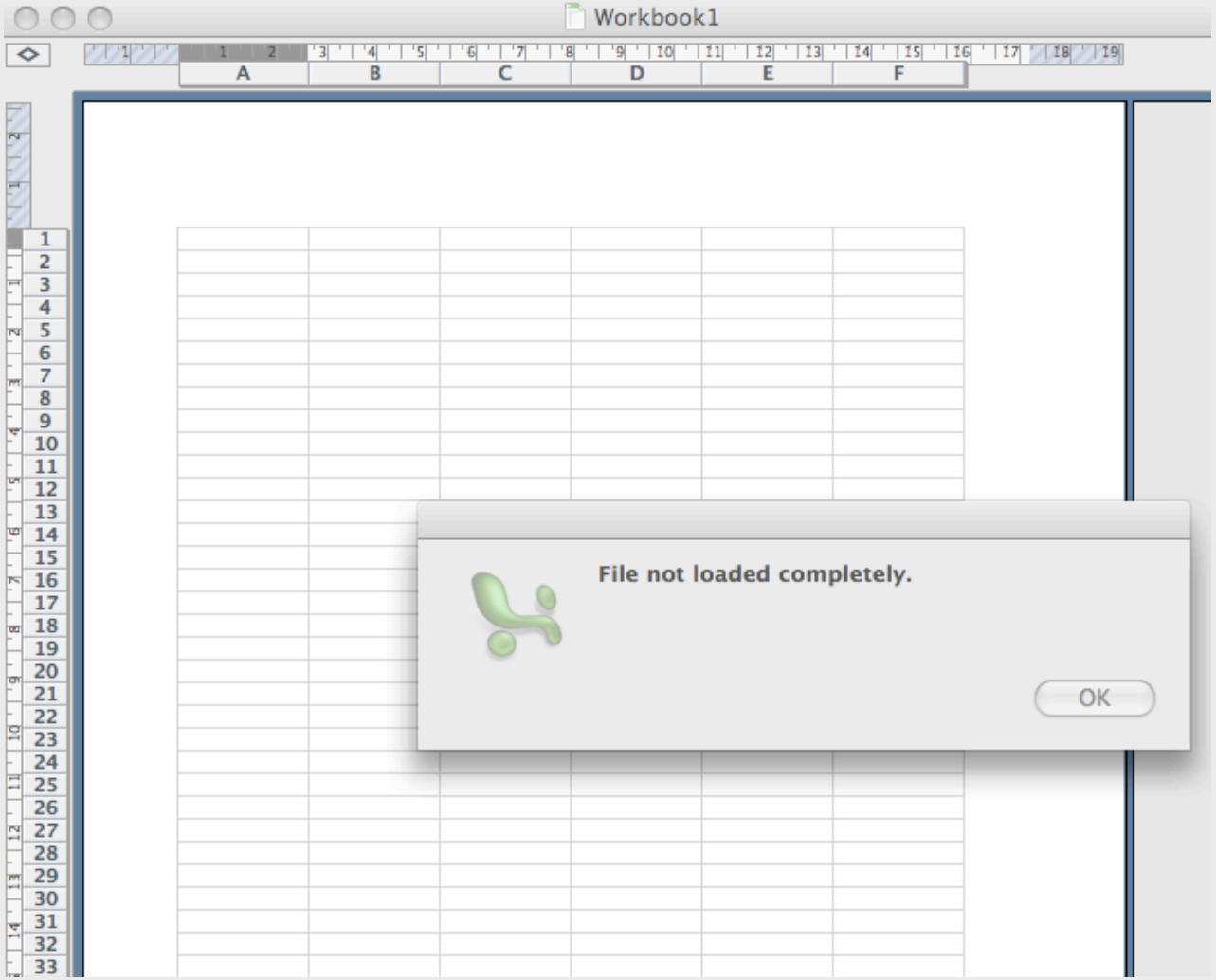
Overview of ChIP-seq experiment



Park J, 2009

EMBL-EBI

Main Challenge - Bioinformatics



EMBL-EBI

Experimental Design

- **Antibody quality**
- Control experiment
- Depth of sequencing
- Multiplexing
- Paired-end reads

Antibody quality

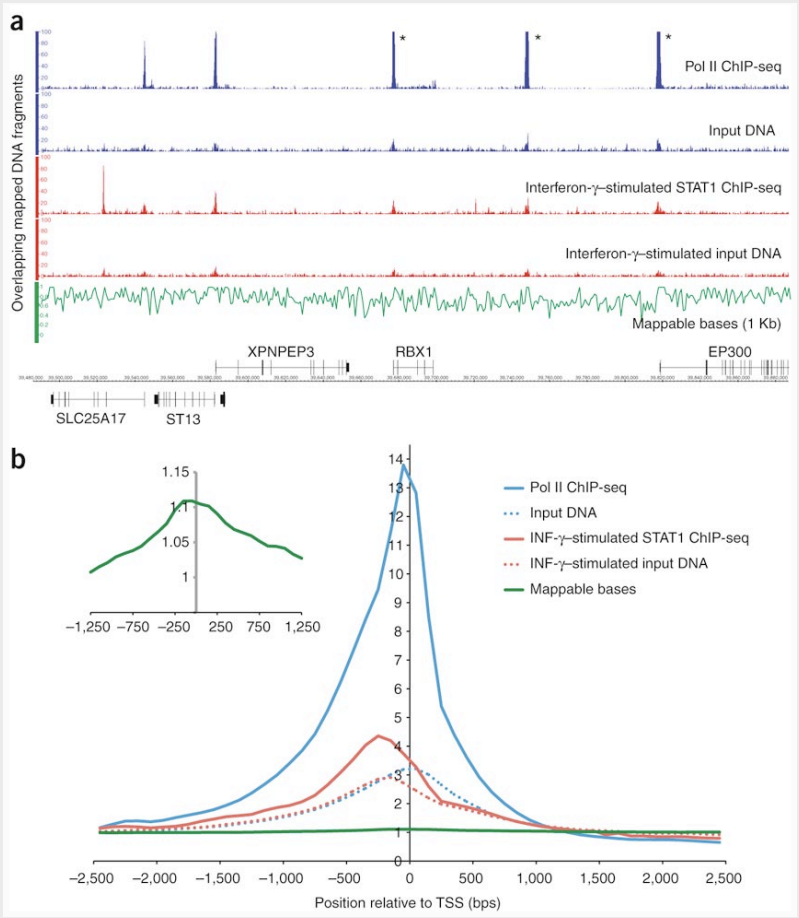
- Antibody quality - a sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason for failed ChIP-seq experiments
  - Check your antibody ahead if possible. Western blotting to check the reactivity of the antibody with unmodified and non-histone proteins.

Experimental Design

- Antibody quality
- **Control experiment**
- Depth of sequencing
- Multiplexing
- Paired-end reads

Why we need a control sample

- Open chromatin regions are fragmented more easily than closed regions.
- Repetitive sequences might seem to be enriched (inaccurate repeats copy number in the assembled genome).
- Uneven distribution of sequence tags across the genome
- A ChIP-seq peak should be compared with the same region in a matched control



Rozowsky, Nature Biotechnology, 2009

EMBL-EBI

Control type

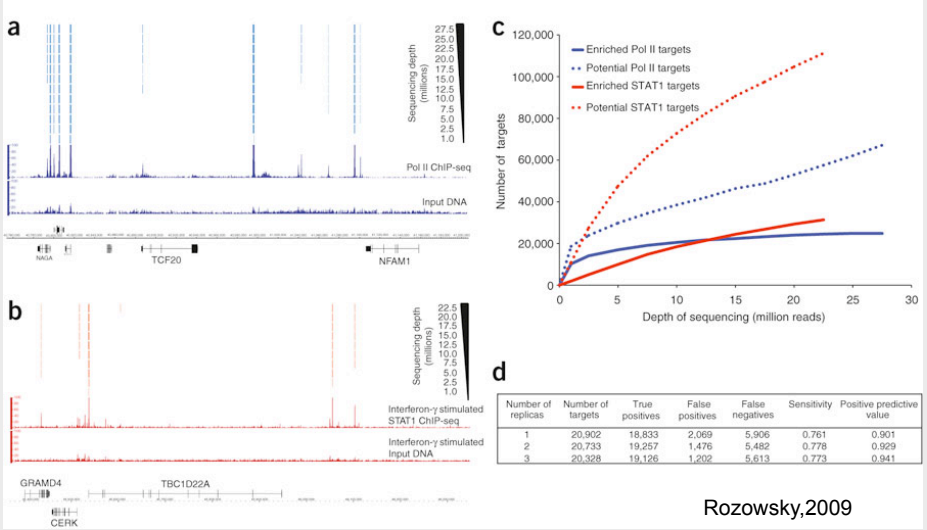
- Input DNA
- Mock IP - DNA obtained from IP without antibody
  - Very little material can be pulled down leading to inconsistent results of multiple mock IPs.
- Nonspecific IP - using an antibody against a protein that is not known to be involved in DNA binding
- There is no consensus on which is the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

Experimental Design

- Antibody quality
- Control experiment
- **Depth of sequencing**
- Multiplexing
- Paired-end reads

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
- GA1 generated 4-6M reads, GA2 12-15M reads, GA2X 18-30M, HiSeq & SOLiD up to 100 M
- With current sequencing technologies, one lane is usually sufficient



Rozowsky,2009

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

Experimental Design

- Antibody quality
- Control experiment
- Depth of sequencing
- **Multiplexing**
- Paired-end reads

Multiplexing

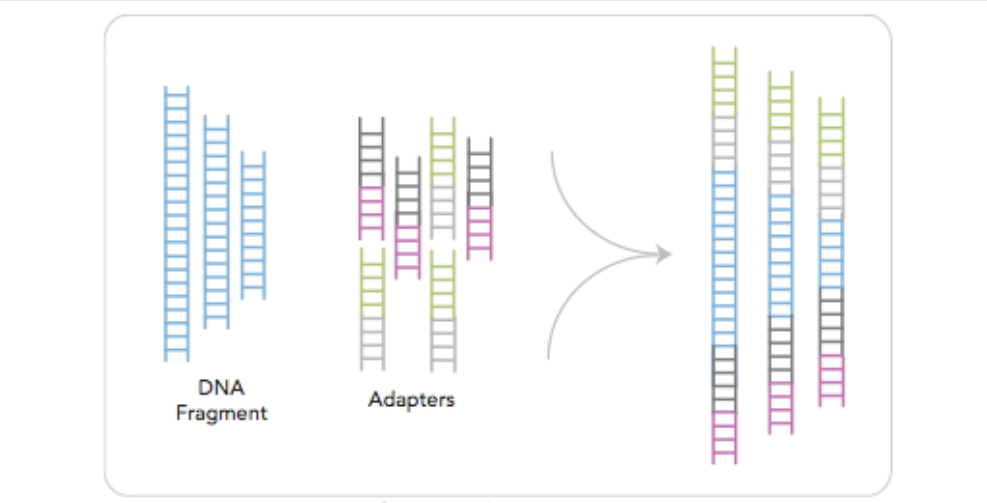
- Number of reads per run continue to increase
- The ability to sequence multiple samples at the same time becomes important, especially for small genomes
- Different barcode adaptors are ligated to different samples

Experimental Design

- Antibody quality
- Control experiment
- Depth of sequencing
- Multiplexing
- **Paired-end reads**

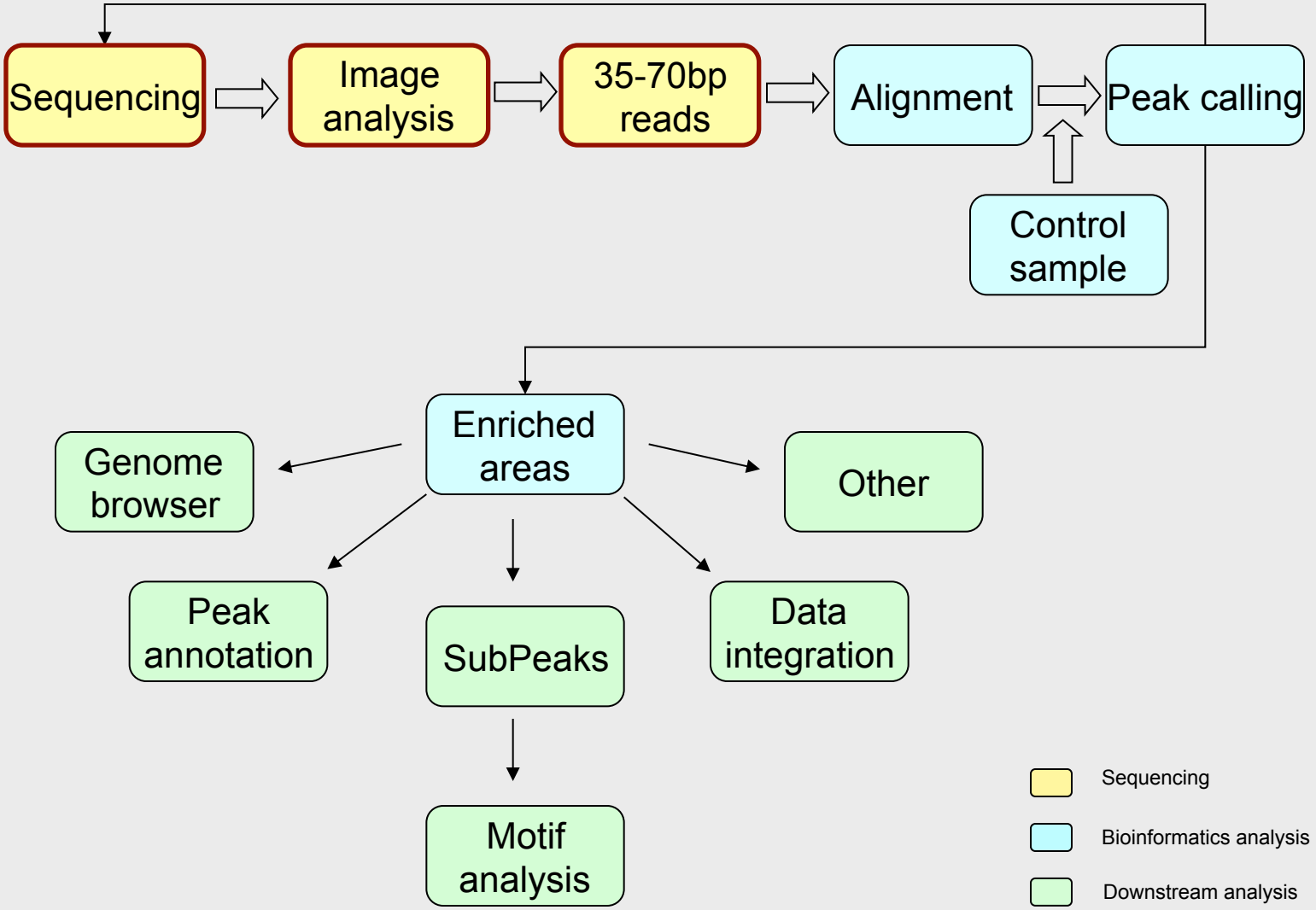
Paired-end sequencing

- Reads are sequenced from both ends
- Increase “mappability” - especially in repetitive regions
- Costs twice as much as single end reads



- For ChIP-seq, usually not worth the extra cost, unless you have a specific interest in repeat regions

Analysis - Overview



Sequencing

Sequence by synthesis

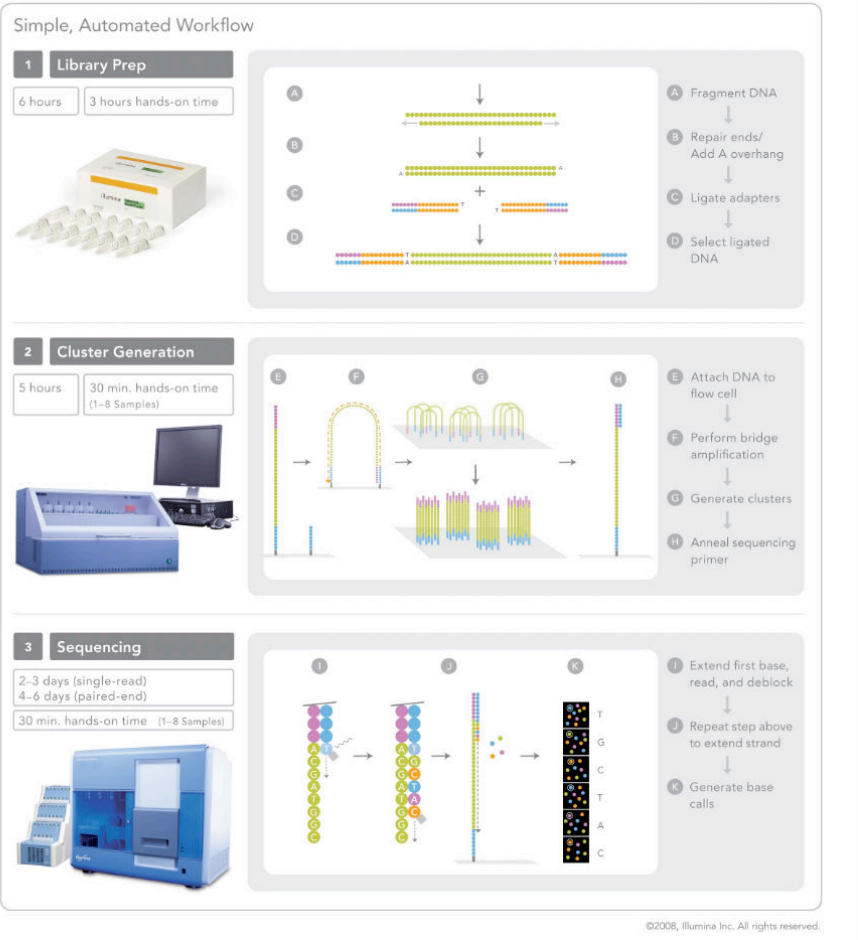
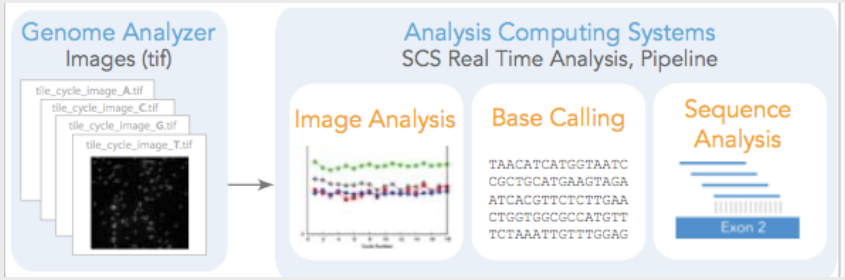


Image analysis



EMBL-EBI

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
```

Fastq 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)

Phred quality scores

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
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%

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.

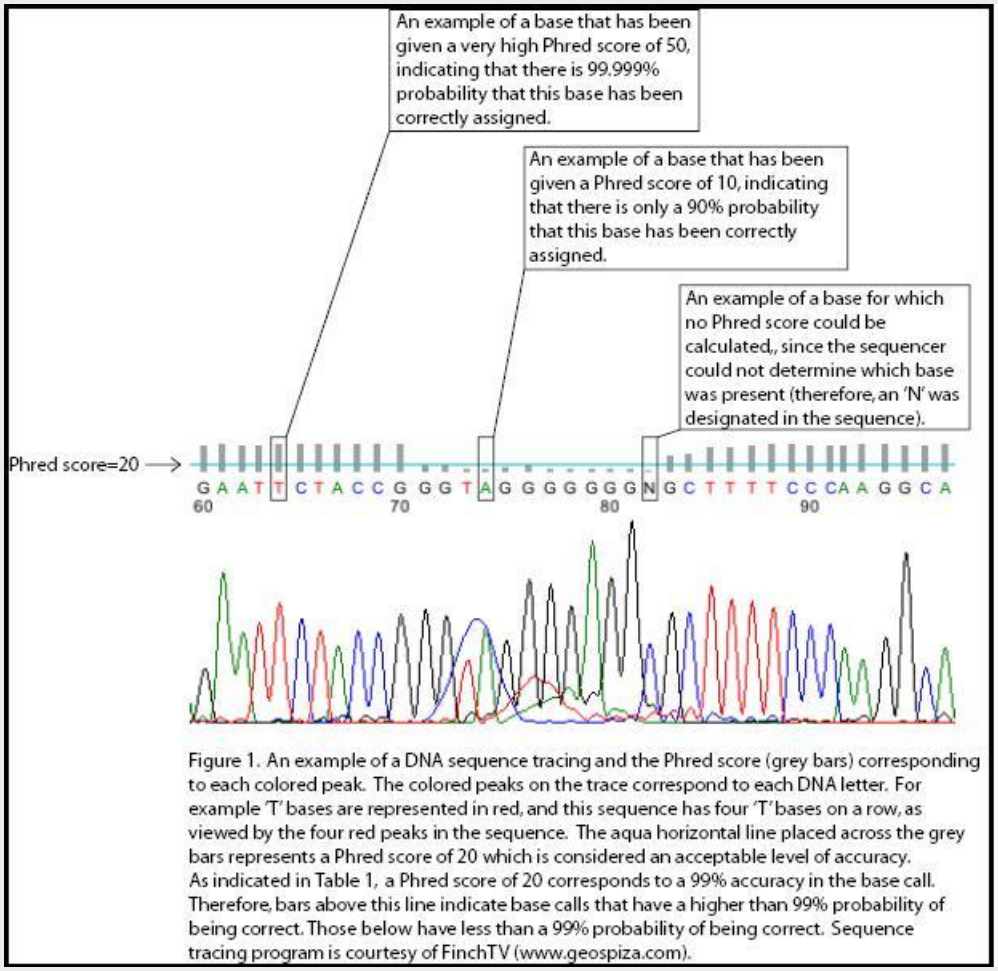
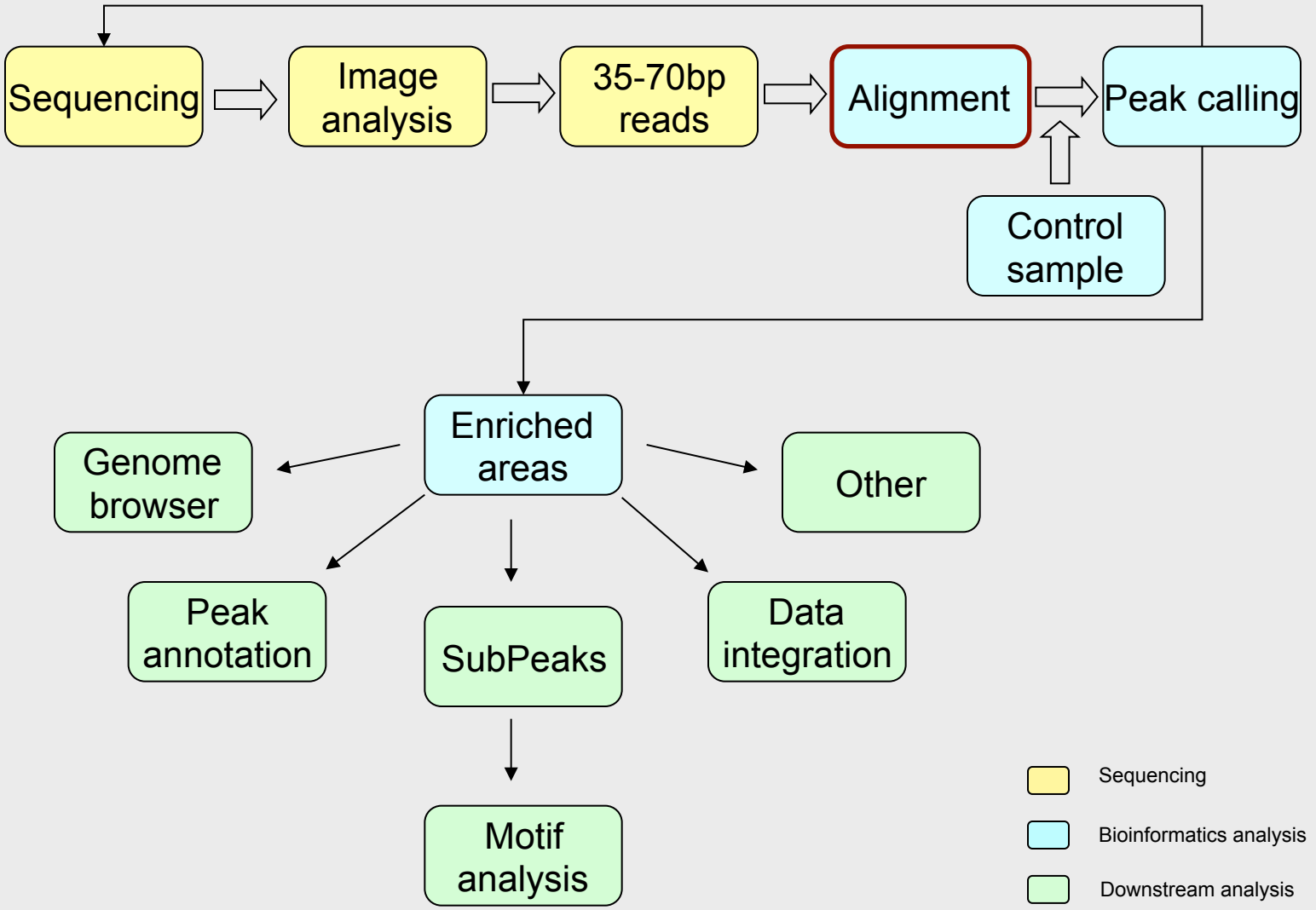


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

EMBL-EBI

Analysis - Overview



Mappability

- Not all of the genome is ‘available’ for mapping
- Align your reads to the unmasked genome

		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens	3,080.44	1,462.69	47.5%	2,451.96	79.6%

*Calculated based on 30nt sequence tags

Rozowsky, 2009

- For ChIP-seq, usually short reads are used (36bp)
- Limited gain in using longer reads (again, unless you have a specific interest in repeat regions)

Mapping Challenges

- Enormous amount of reads to align
- Done against large genome - needs pre-indexing structure and large memory
- Shorter reads length
- Mismatches
  - Sequencing errors - higher error rates than conventional sequencing methods
  - SNPs, insertion/deletions
- Repetitive regions
- Has to be fast and memory efficient

Mapping Softwares

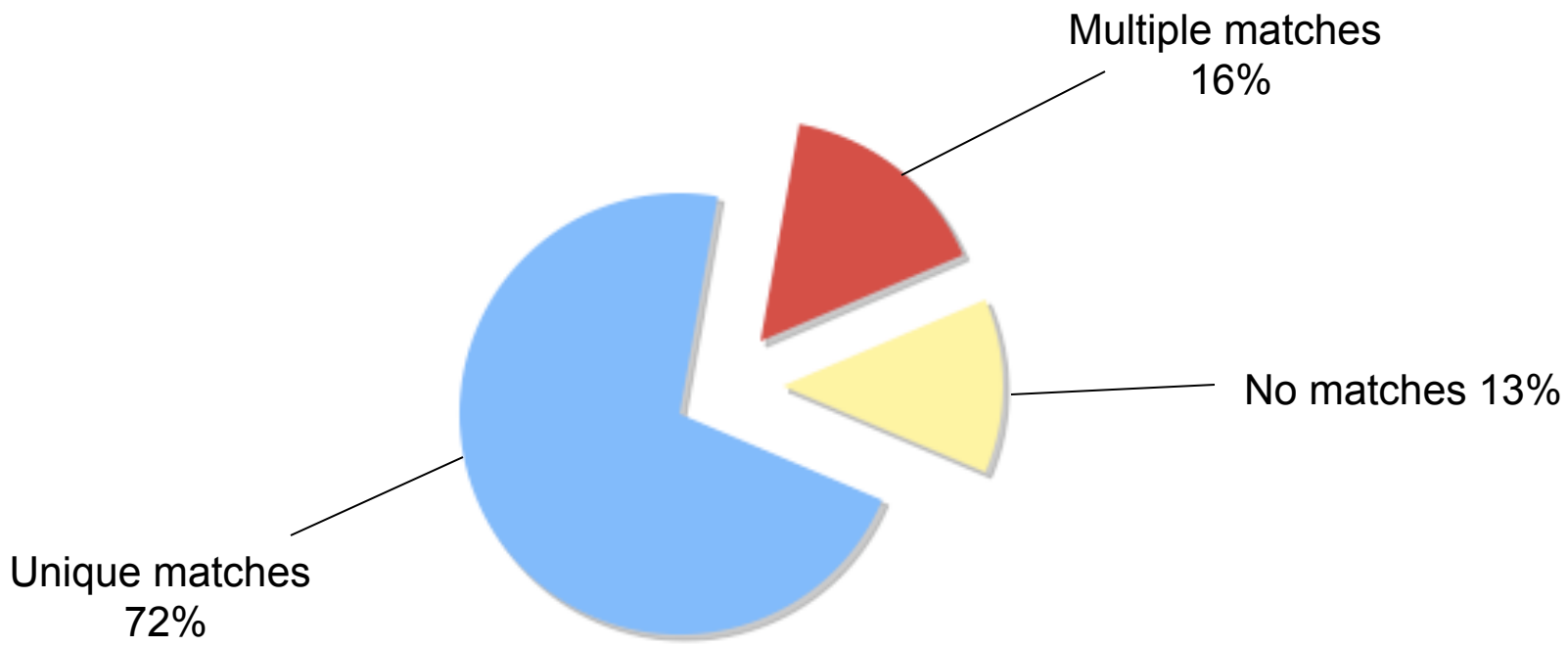
- BLAST - have been developed for conventional sequencing methods, do not take into account the shorter read lengths and the higher error rate.
- Therefore new programs have been developed, which deal with these new challenges.
- Each program usually addresses the problem slightly different
  - gapped versus un-gapped alignment
  - taking the quality value of each base into account.
  - How to treat non-unique reads
  - trim reads at the 3' end
  - find a balance between accuracy and speed of mapping
- Each program might be applicable for different applications and/or sequencing technologies.

Mapping Softwares

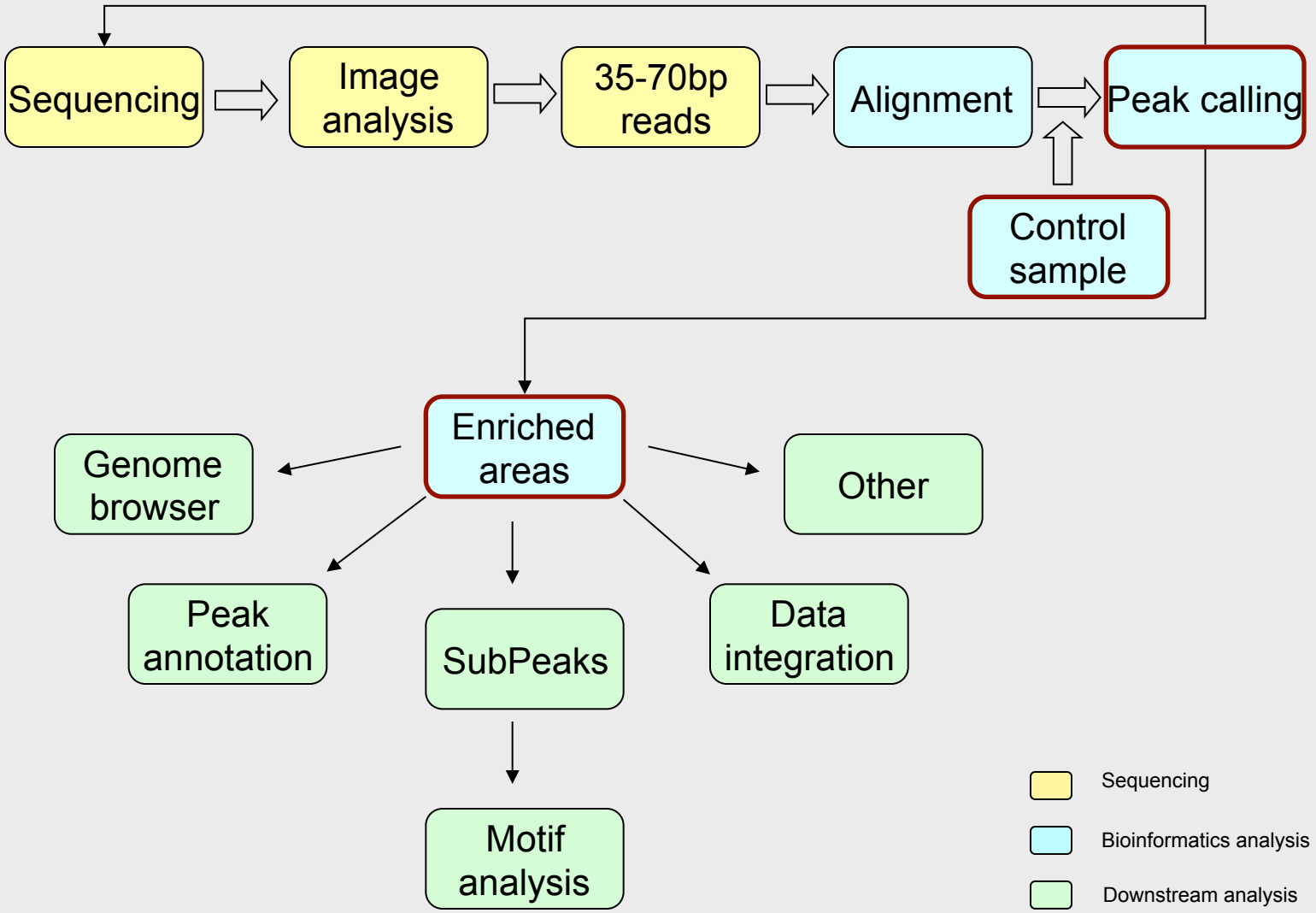
- ELAND - provided with Illumina sequencer
  - Limited reads length
  - Allow 2 substitutions
- MAQ (Li et al 2008)
  - Uses quality values
  - Integrate consensus calling
- Bowtie
  - Ultra fast
  - Can work on workstations with < 2Gb memory

Many others: BWA, Novoalign, BFAST,…

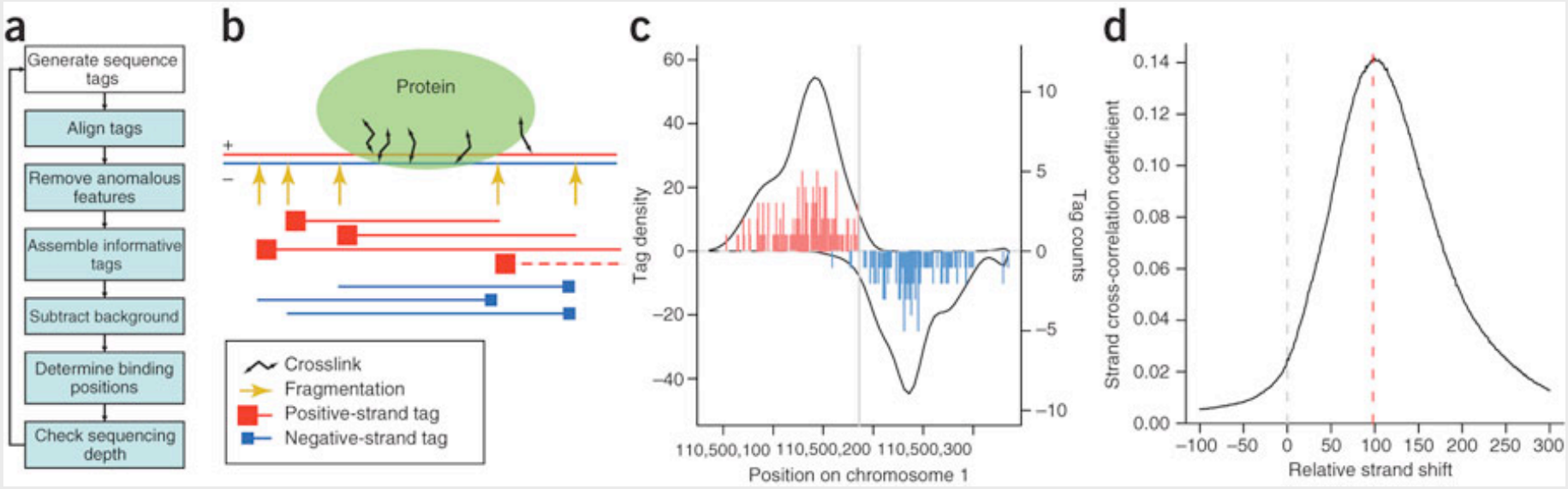
36-mers single-end reads mapped with Maq software



Analysis - Overview



Strand specific profile



Kharchenko, Nature Biotechnology, 2008

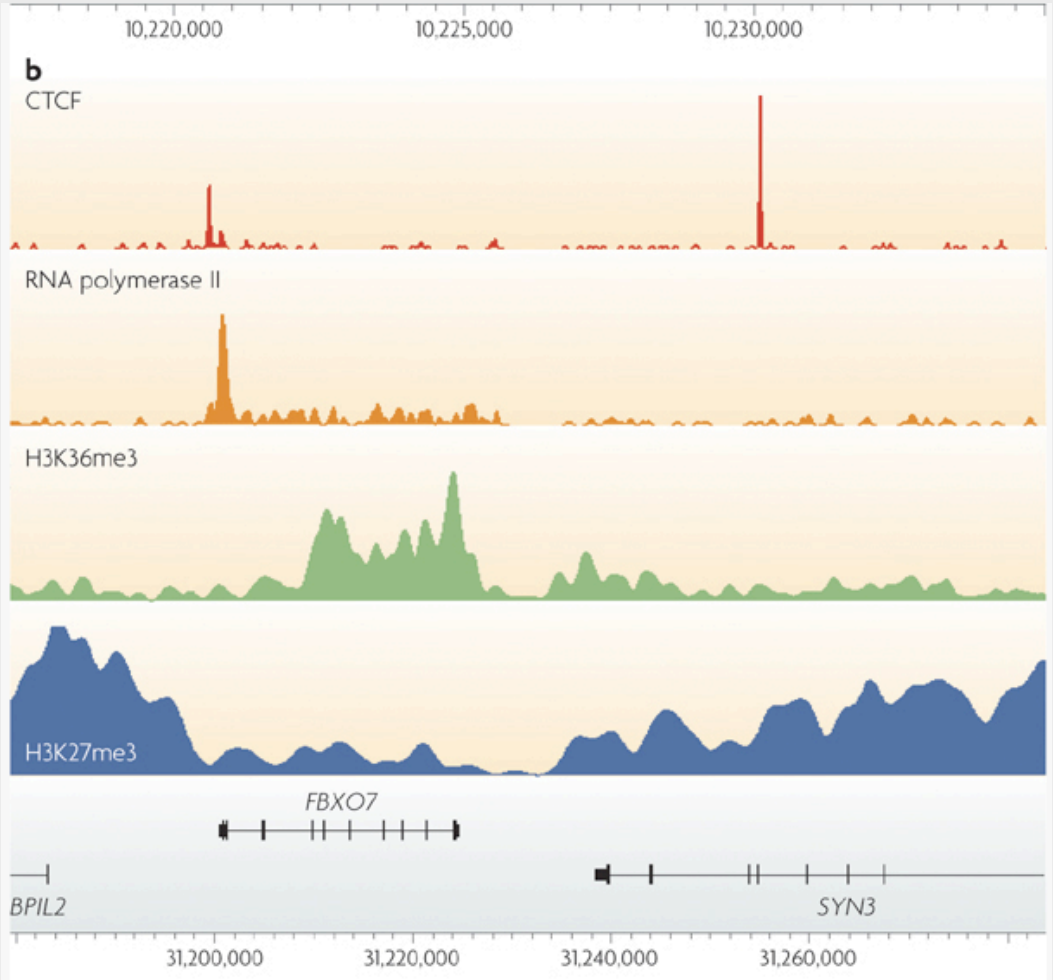
EMBL-EBI

Peak Calling

- Basic - regions are scored by the number of tags in a window of a given size. Then assess by enrichment over control and minimum tag density.
- Advanced - take advantage of the directionality of the reads.
- Advanced methods make more assumptions, making them less appropriate in certain cases

Peak Calling - Challenges

- Adjust for sequence alignability - regions that contain repetitive elements have different expected tag count
- Different ChIP-seq applications produce different type of peaks. Most current tools have been designed to detect sharp peaks (TF binding, histone modifications at regulatory elements)
- Alternative tools exist for broader peaks (histone modifications that mark domains - transcribed or repressed), e.g. SICER

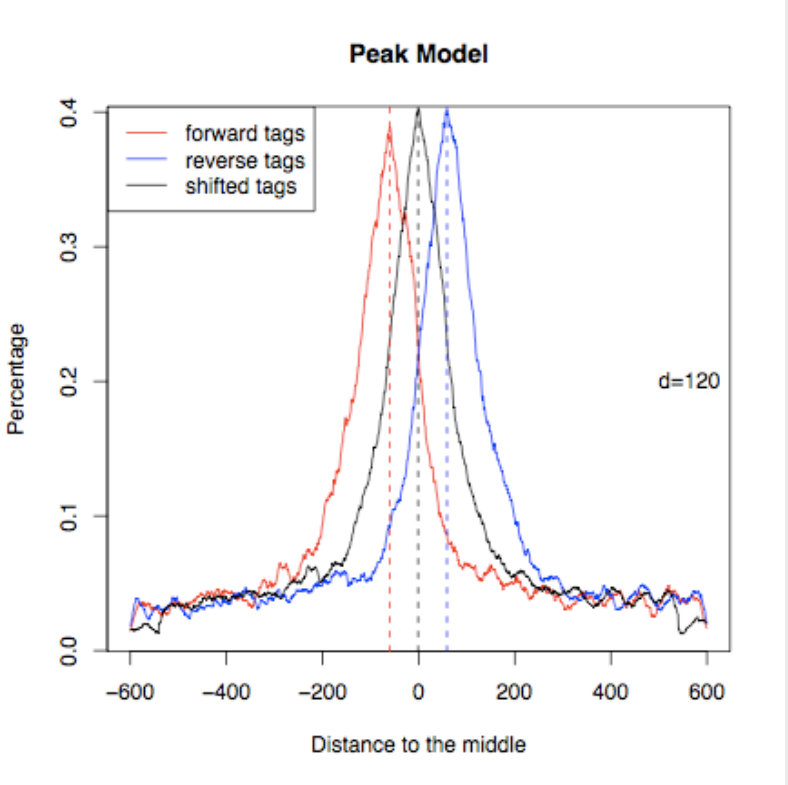


Park J, Nature Reviews Genetics, 2009

EMBL-EBI

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.



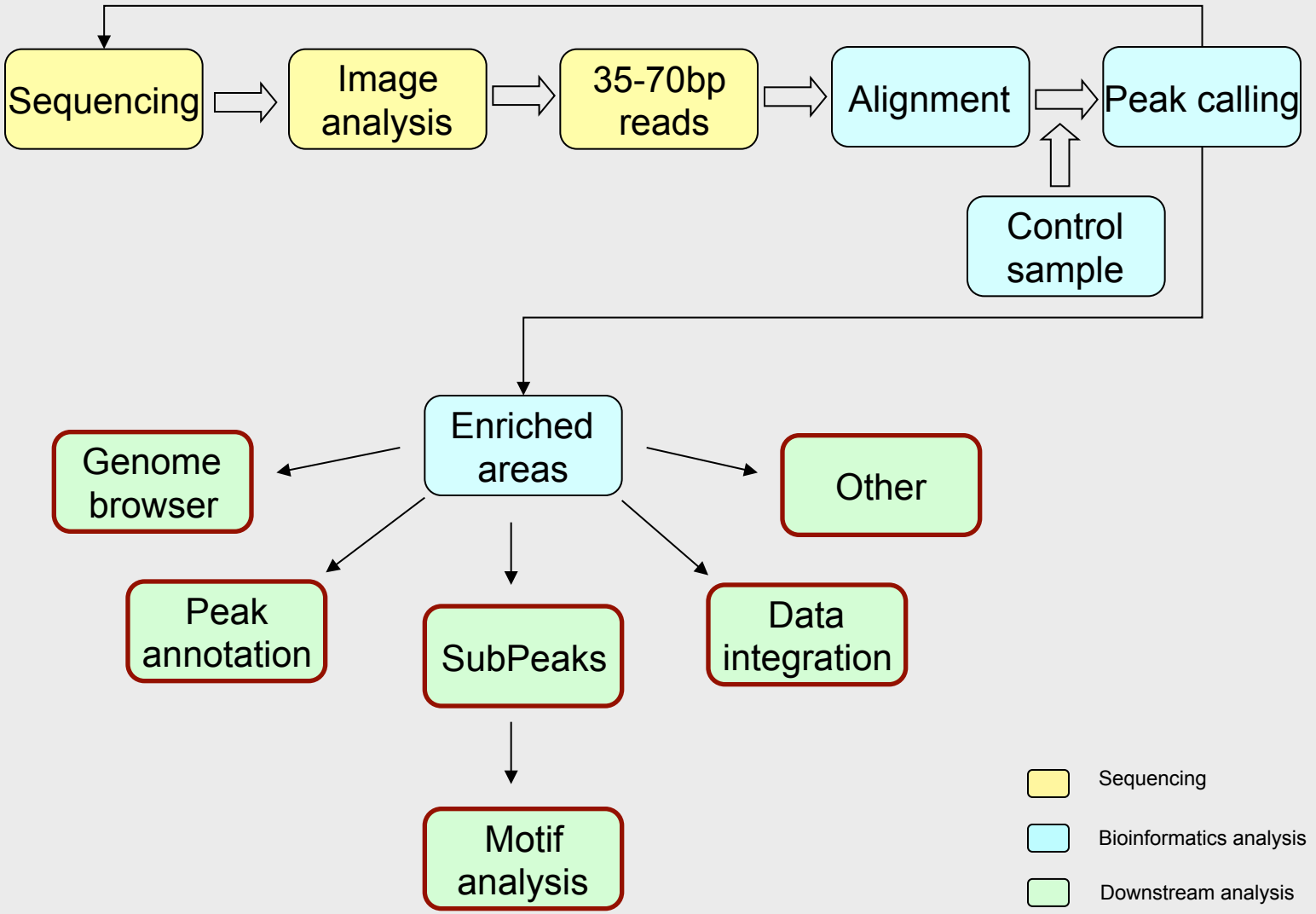
MACS - Peak detection

- Remove duplicate tags (in excess of what can be expected by chance)
- Slide window across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution, global background, p-value 10e-5)
- Merge overlapping peaks, and extend each tag d bases from its center
- Also looks at local background levels and eliminates peaks that are not significant with respect to local background
- Uses the control sample to eliminates peaks that are also present there

Non-peak based analysis

- Can be more appropriate for non-specific binding sites
- Does not involve calling significant peaks, and discarding the rest of the signal as noise
- Instead, the whole signal is used for analysis
- Global analysis, e.g. looking for enrichment at TSS
- Tools include CEAS and EpiChip

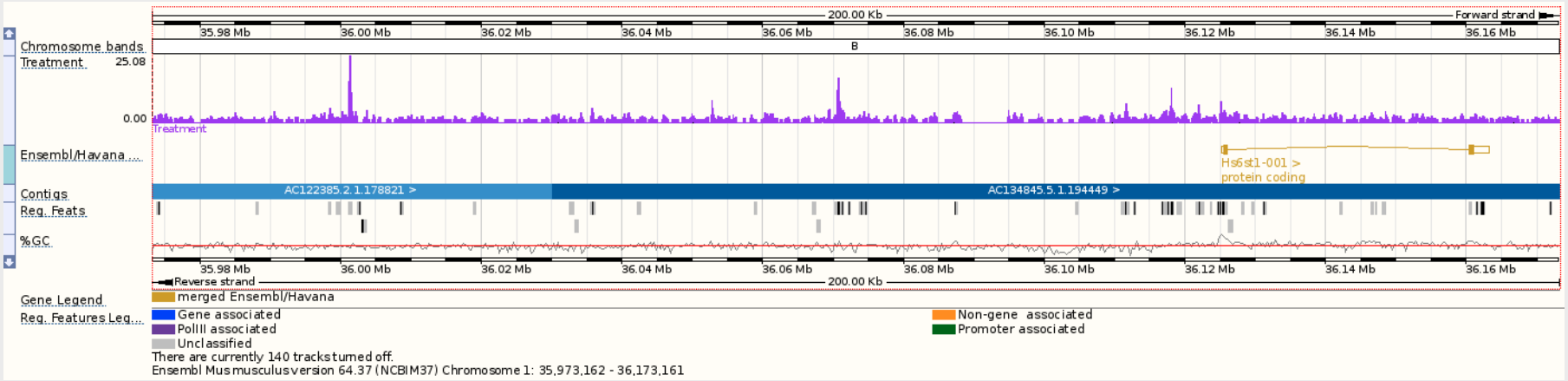
Analysis - Overview



Analysis downstream to peak calling

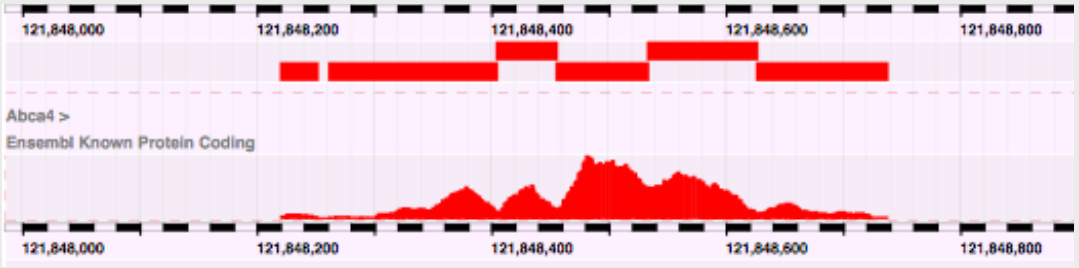
- Visualization - genome browser: Ensembl, UCSC, IGB
- Peak Annotation - finding interesting features surrounding peak regions: PeakAnalyzer
- Correlation with expression data
- Discovery of binding sequence motifs
  - Split peaks
  - Fetch summit sequences
  - Run motif prediction tool
- Gene Ontology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

Visualization in a genome browser



EMBL-EBI

Motif Analysis



GAATCCCACA TTTGCATAACAAAAG ACTCCTGGTG
CAGCTGCTCT TCTGCATAACAAAGG GTGGCCCTGC
CCGGTTTTTC TTTGCATAACAATAA GATCTGGCTA
TTATTCTCAC TTTGCATAGGAATGG GGCAGTTAGA
CACAGCCACA TTTGCATAACAGAAG CCGAGCCCGC
CTTGGGTGAA TTTGCAAGACAAAGG ACAATGATCA