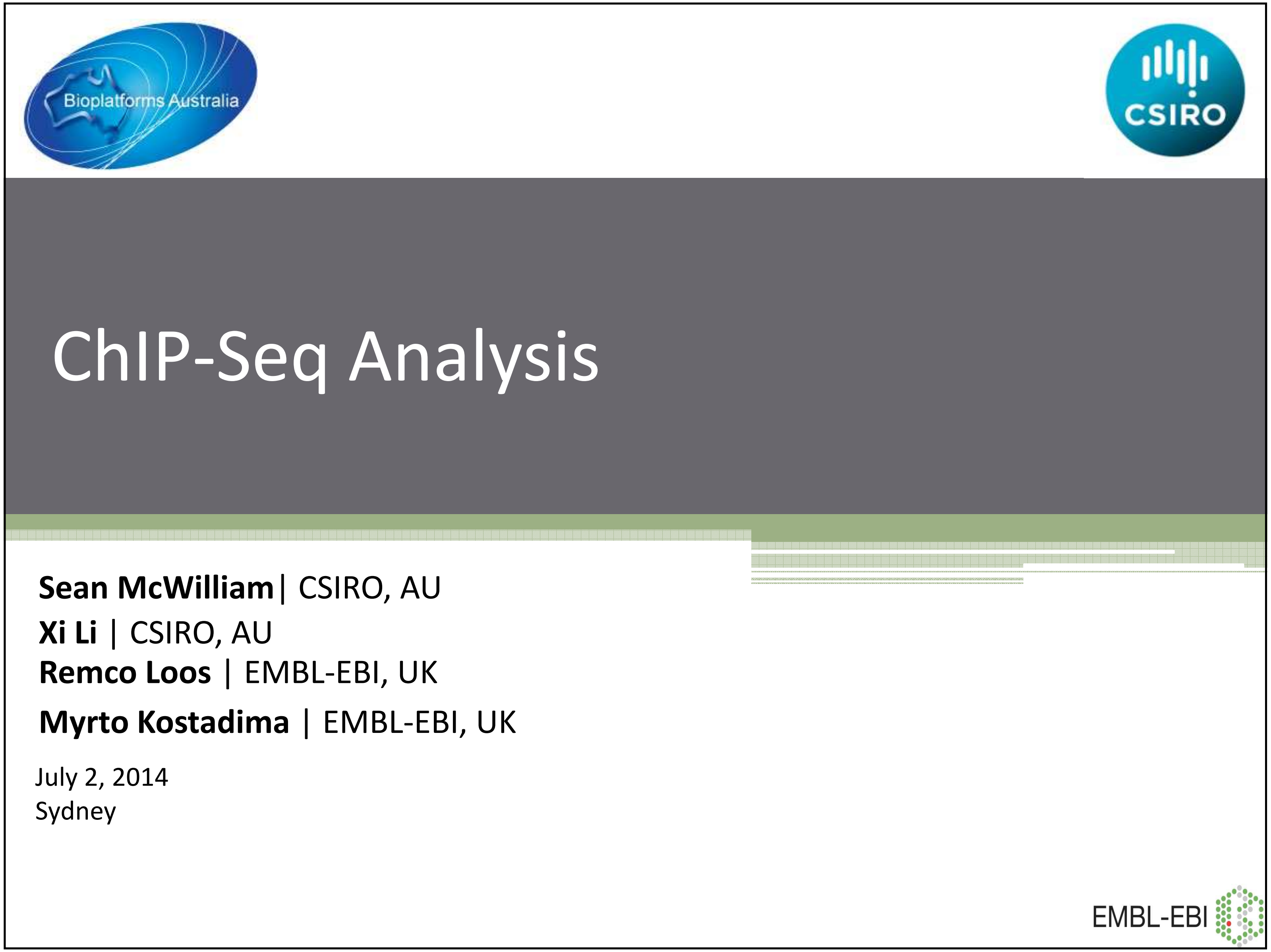
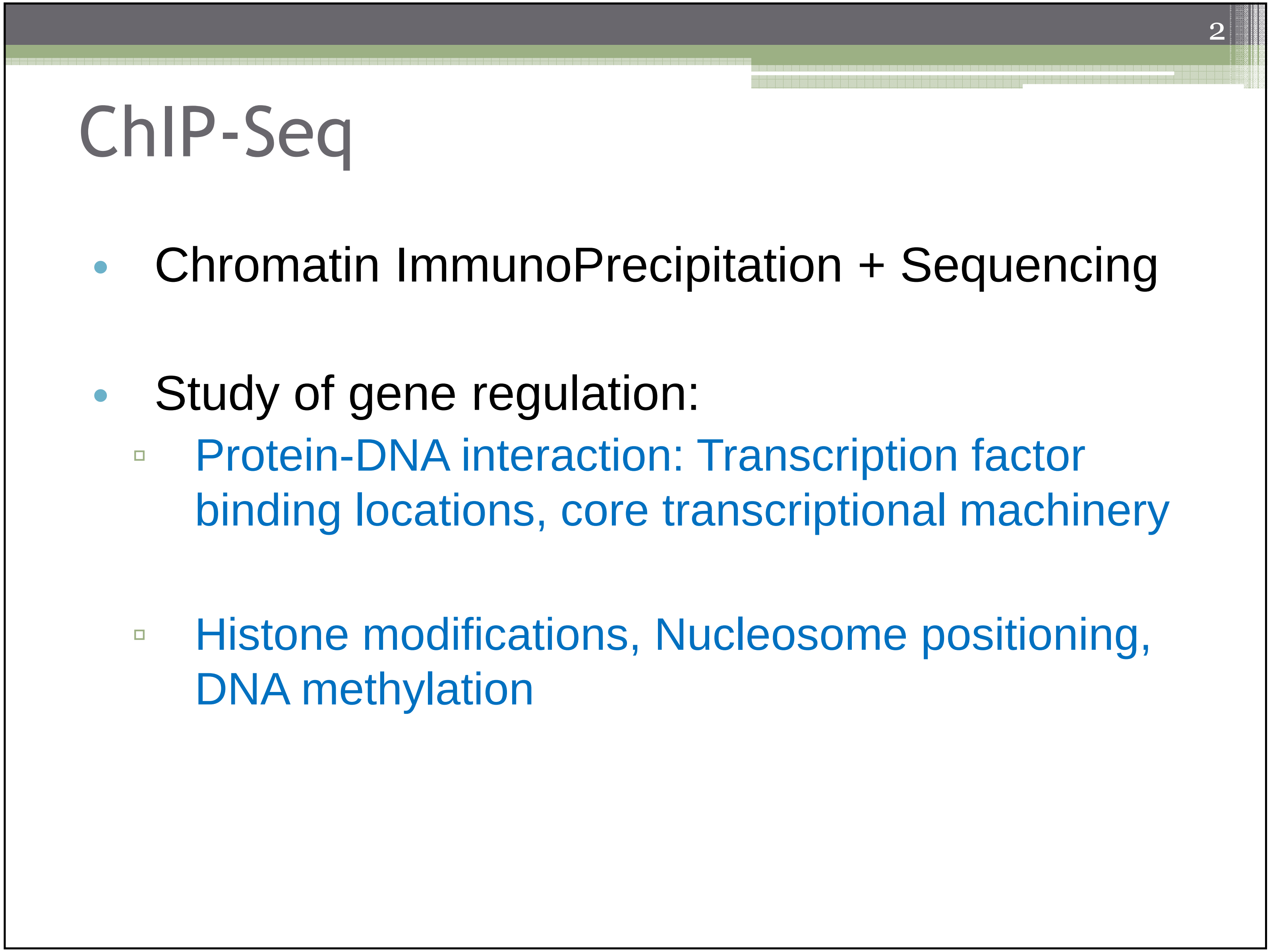


ChIP-Seq Analysis

Sean McWilliam| CSIRO, AU
Xi Li | CSIRO, AU
Remco Loos | EMBL-EBI, UK
Myrto Kostadima | EMBL-EBI, UK

July 2, 2014
Sydney



ChIP-Seq

- Chromatin ImmunoPrecipitation + Sequencing
- Study of gene regulation:
  - Protein-DNA interaction: Transcription factor binding locations, core transcriptional machinery
  - Histone modifications, Nucleosome positioning, DNA methylation

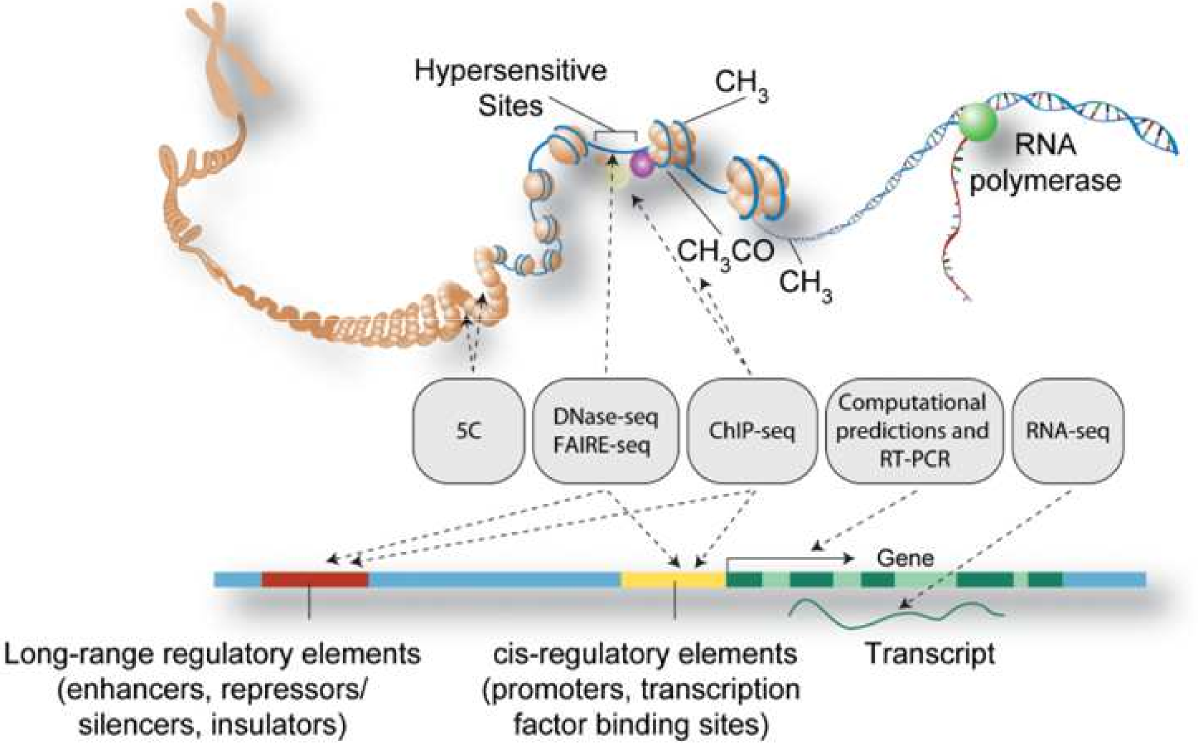
ChIP-Seq

- 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 1300 publications currently in PubMed
- Nearly 300 this year so far

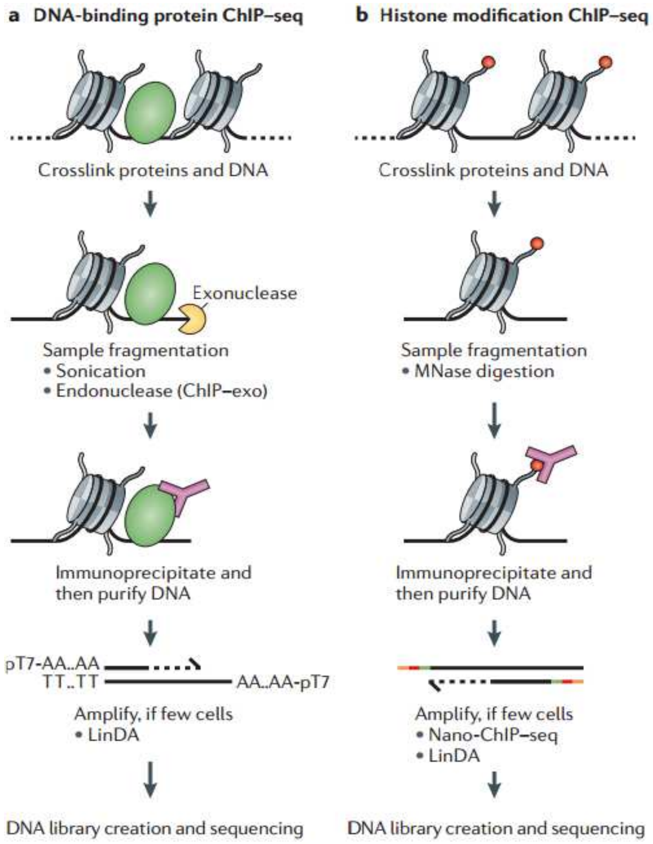
ChIP-Seq


Hypersensitive Sites
CH3
RNA polymerase
CH3CO
CH3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Transcript
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)

A User's Guide to the Encyclopedia of DNA Elements (ENCODE), 2011 PLOS | BIOLOGY

ChIP-Seq Lab procedures

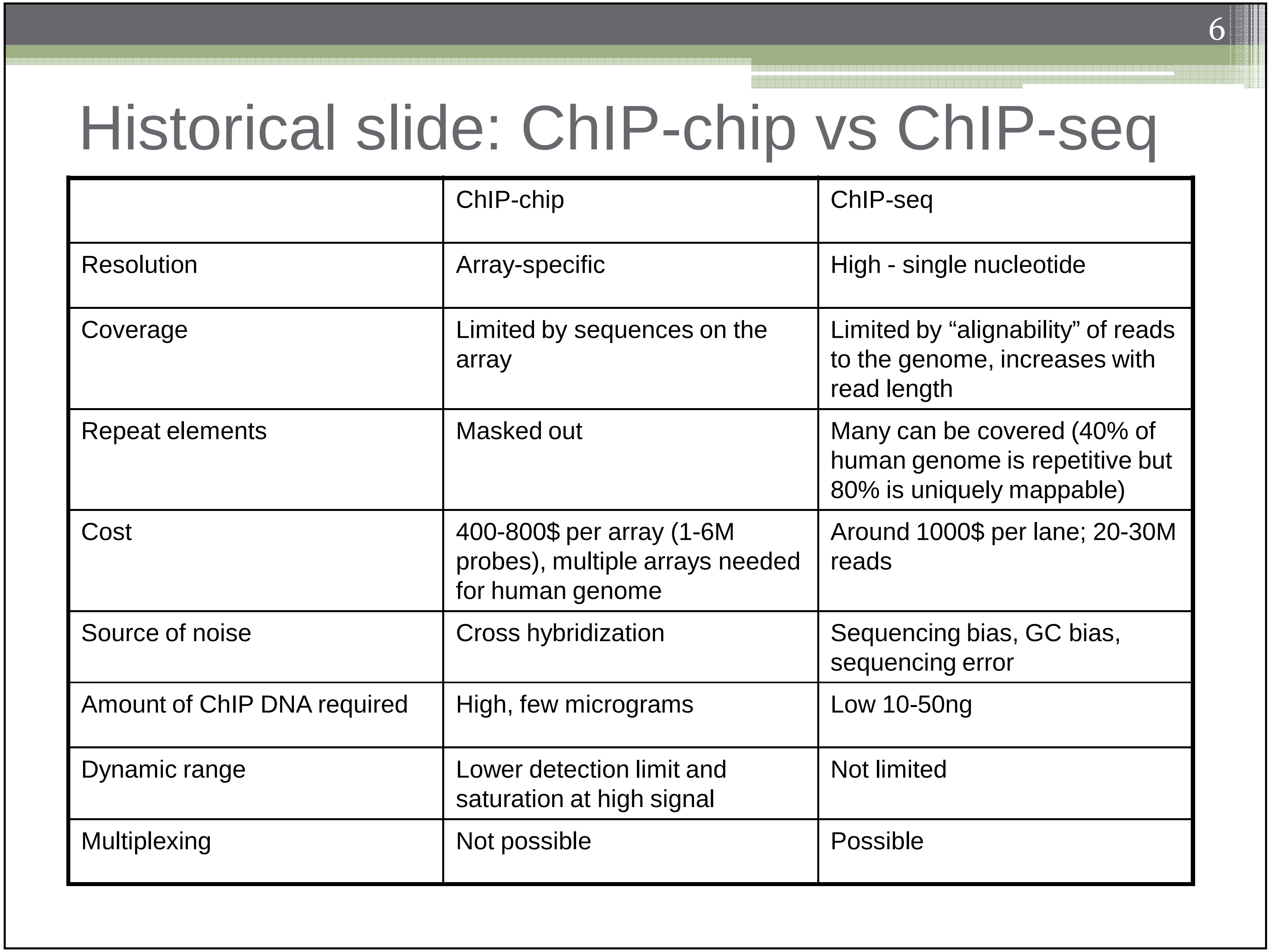
- Transcription factors
- Histones
- Nuclear receptors
- Polymerase


a DNA-binding protein ChIP–seq
b Histone modification ChIP–seq
Crosslink proteins and DNA
Crosslink proteins and DNA
Exonuclease
Sample fragmentation
• Sonication
• Endonuclease (ChIP–exo)
Sample fragmentation
• MNase digestion
Immunoprecipitate and then purify DNA
Immunoprecipitate and then purify DNA
pT7-AA..AA
TT..TT
AA..AA-pT7
Amplify, if few cells
• LinDA
Amplify, if few cells
• Nano-ChIP–seq
• LinDA
DNA library creation and sequencing
DNA library creation and sequencing

Furey, TS, 2012

Nature Reviews | Genetics

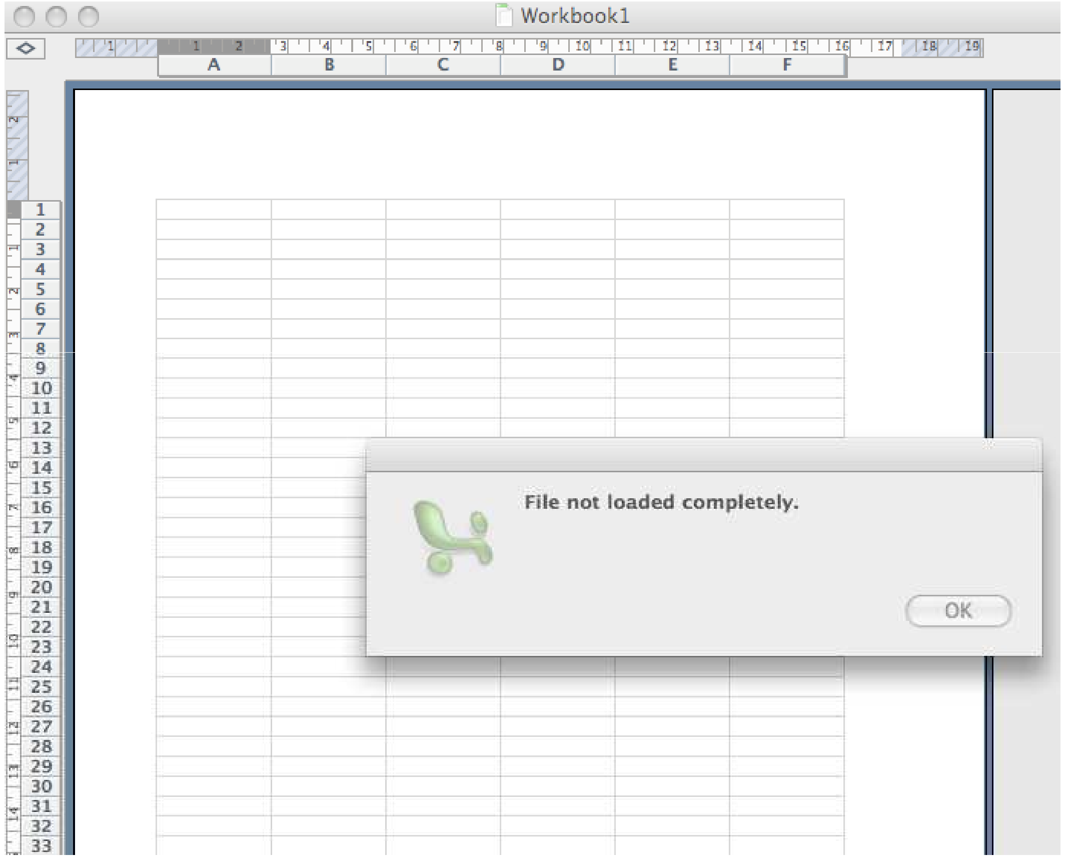
- PCR with gene specific primers
- Hybridization on microarrays
- Sequencing



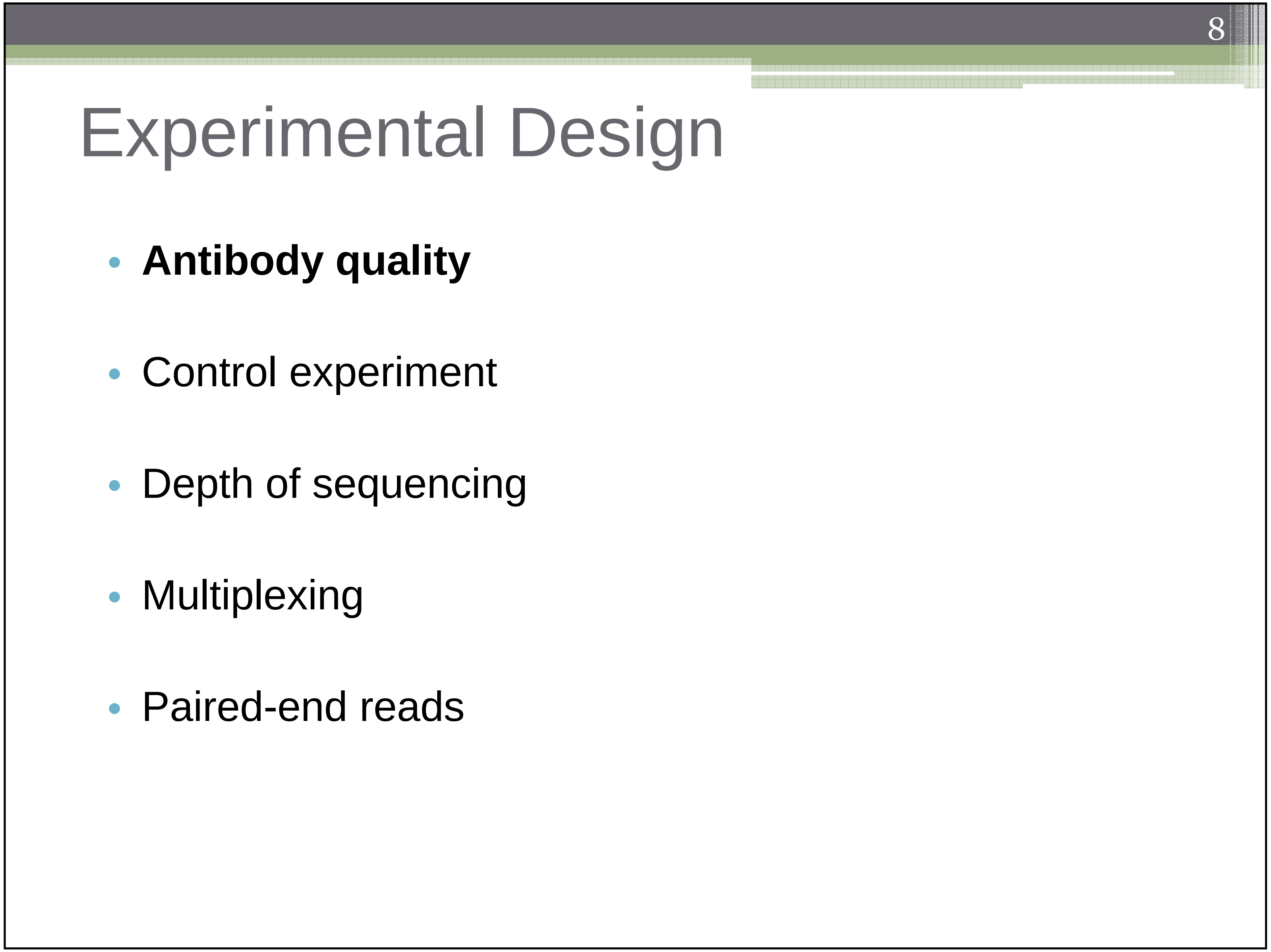
Historical slide: ChIP-chip vs ChIP-seq

	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

Main Challenge - Bioinformatics

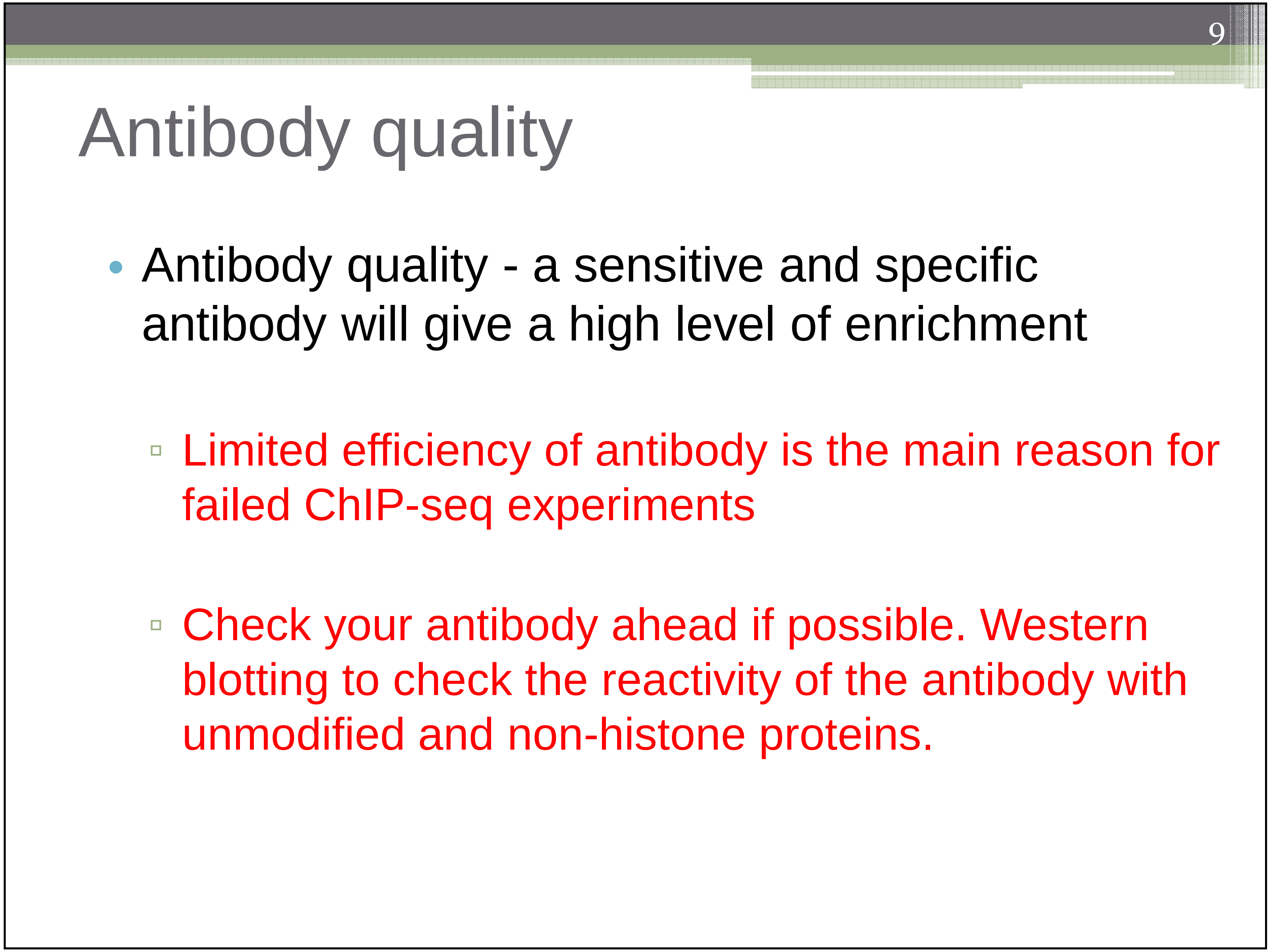


File not loaded completely.



Experimental Design

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



Antibody quality

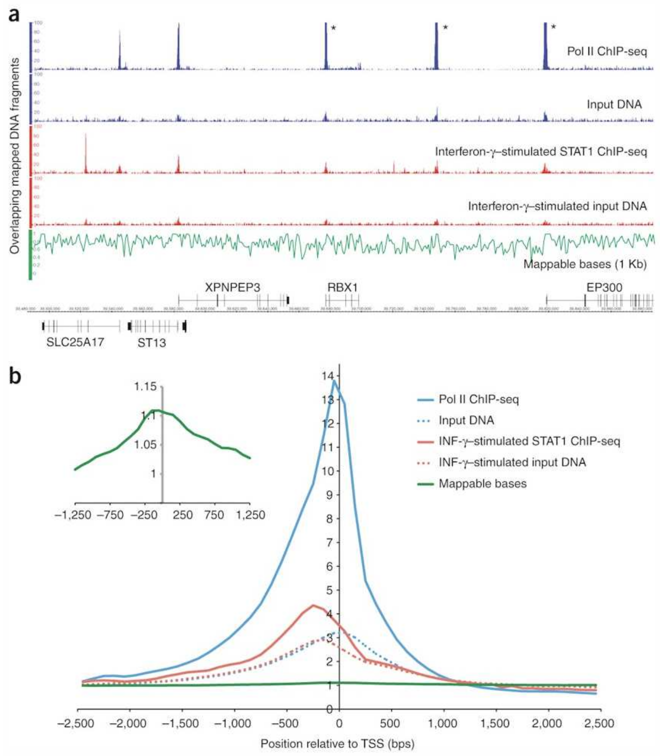
- Antibody quality - a sensitive and specific antibody will give a high level of enrichment
  - Limited efficiency of antibody is the main reason for failed ChIP-seq experiments
  - Check your antibody ahead if possible. Western blotting to check the reactivity of the antibody with unmodified and non-histone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

Experimental Design

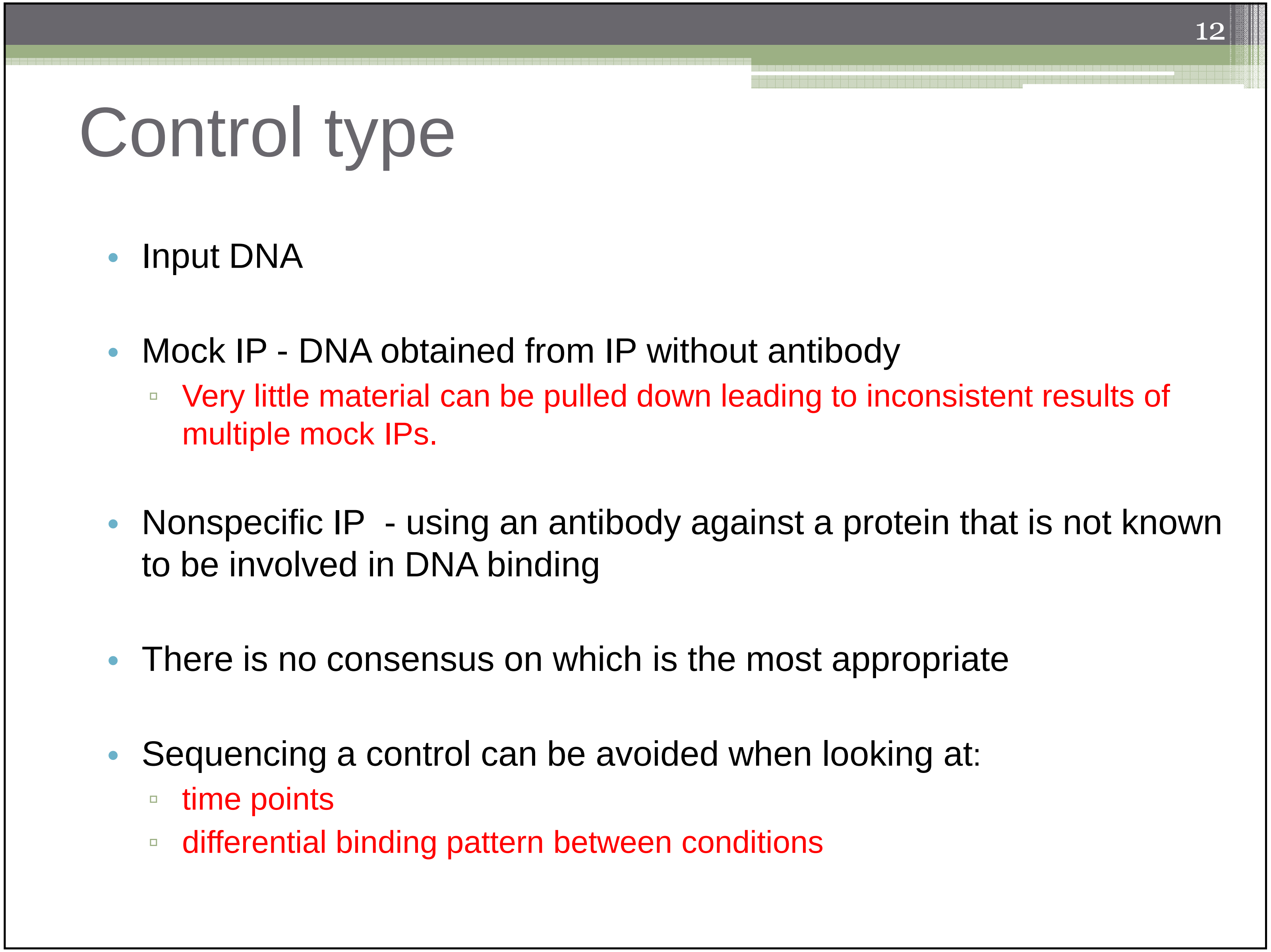
- Antibody quality
- **Control experiment**
- Depth of sequencing
- Multiplexing
- Paired-end reads

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

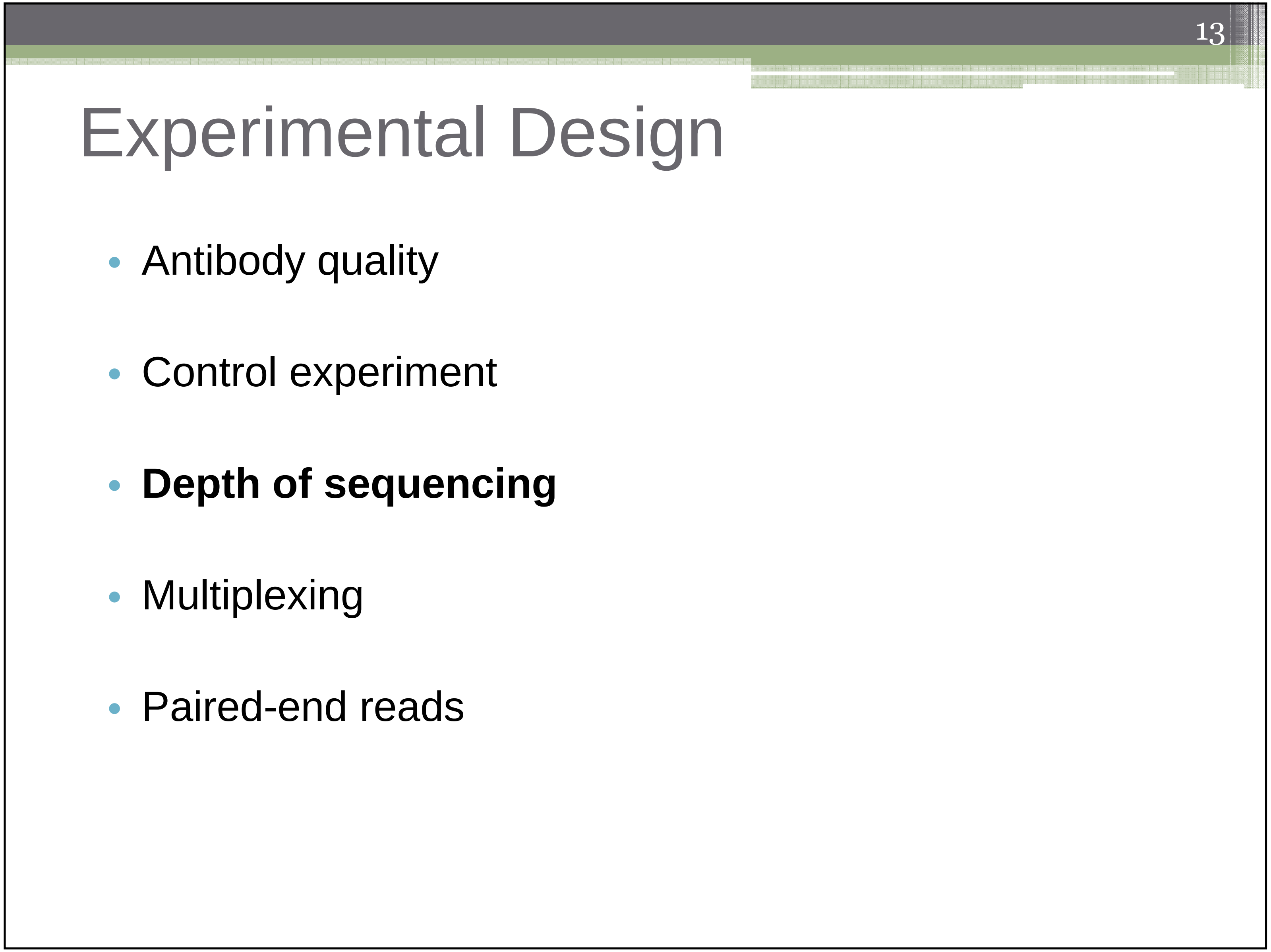


Rozowsky, Nature Biotechnology, 2009



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

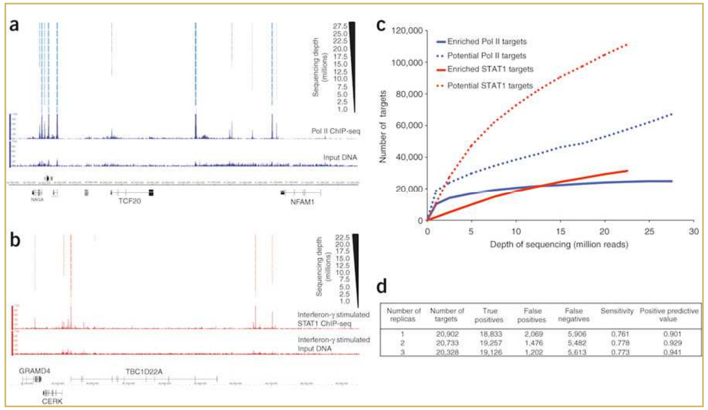


Experimental Design

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

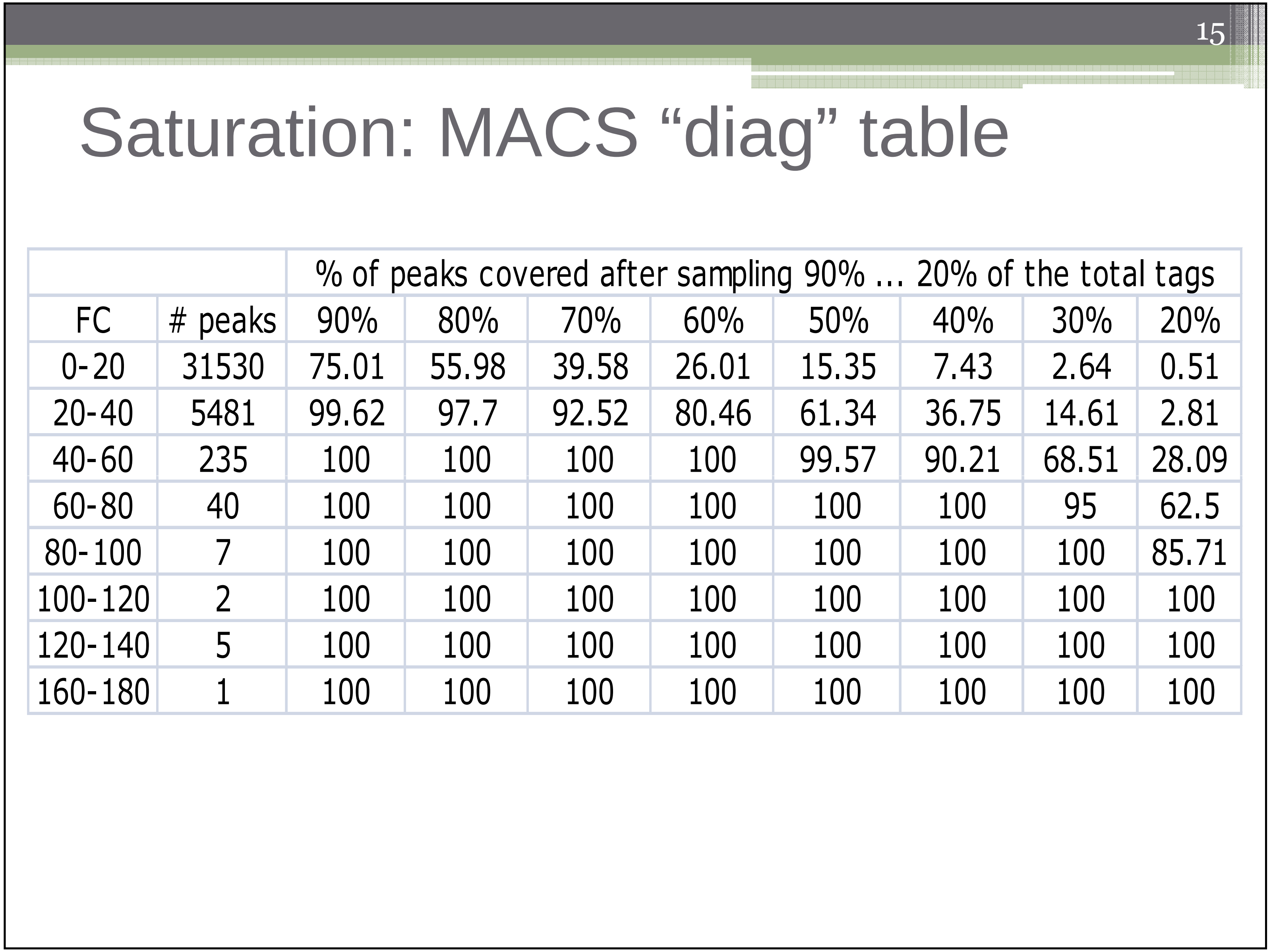
Depth of sequencing

- 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



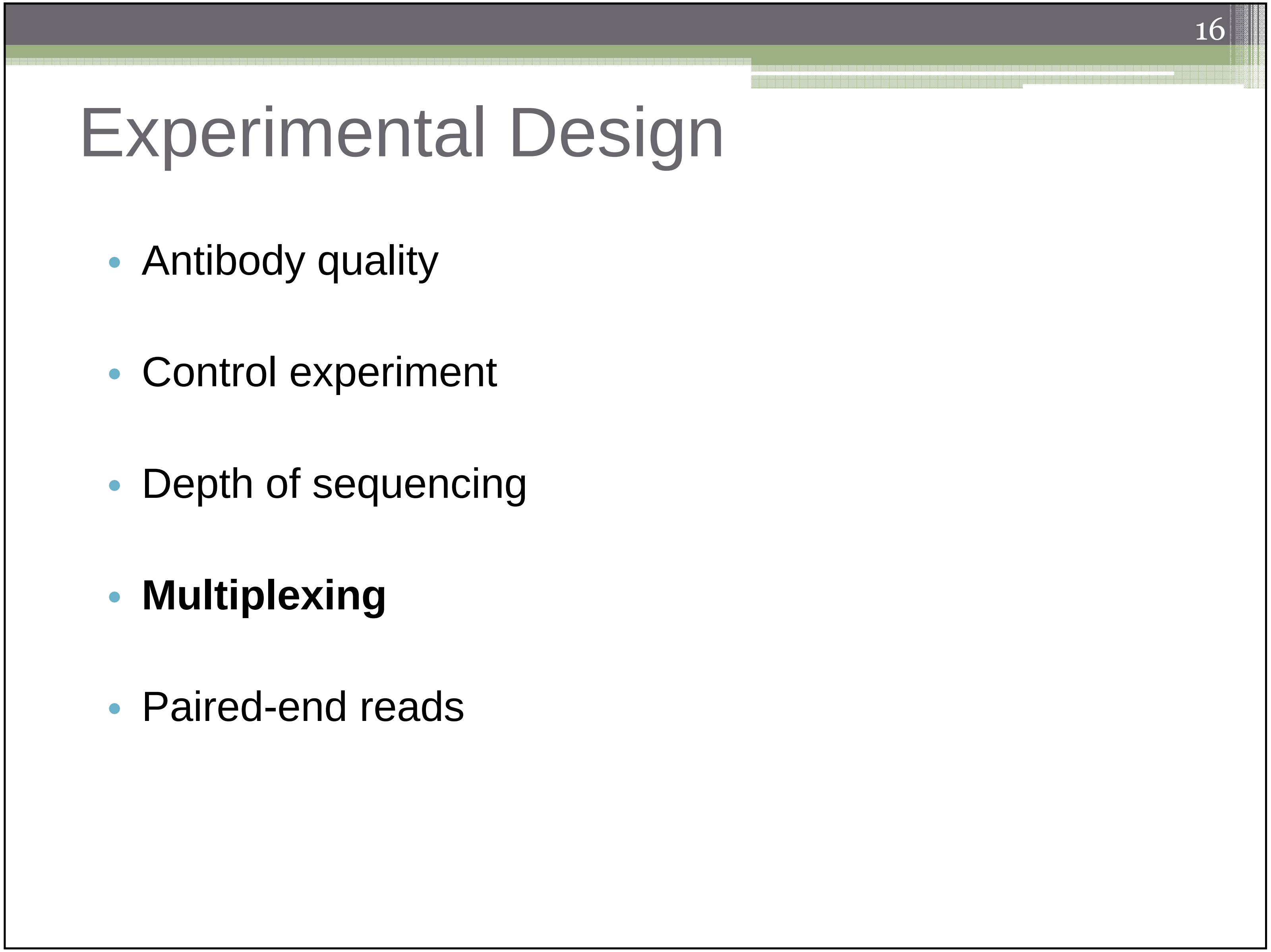
Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

Sequencing Depth and Coverage: key considerations in genomic analyses
Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP.
Nat Rev Genet. 2014 Feb;15(2):121-32



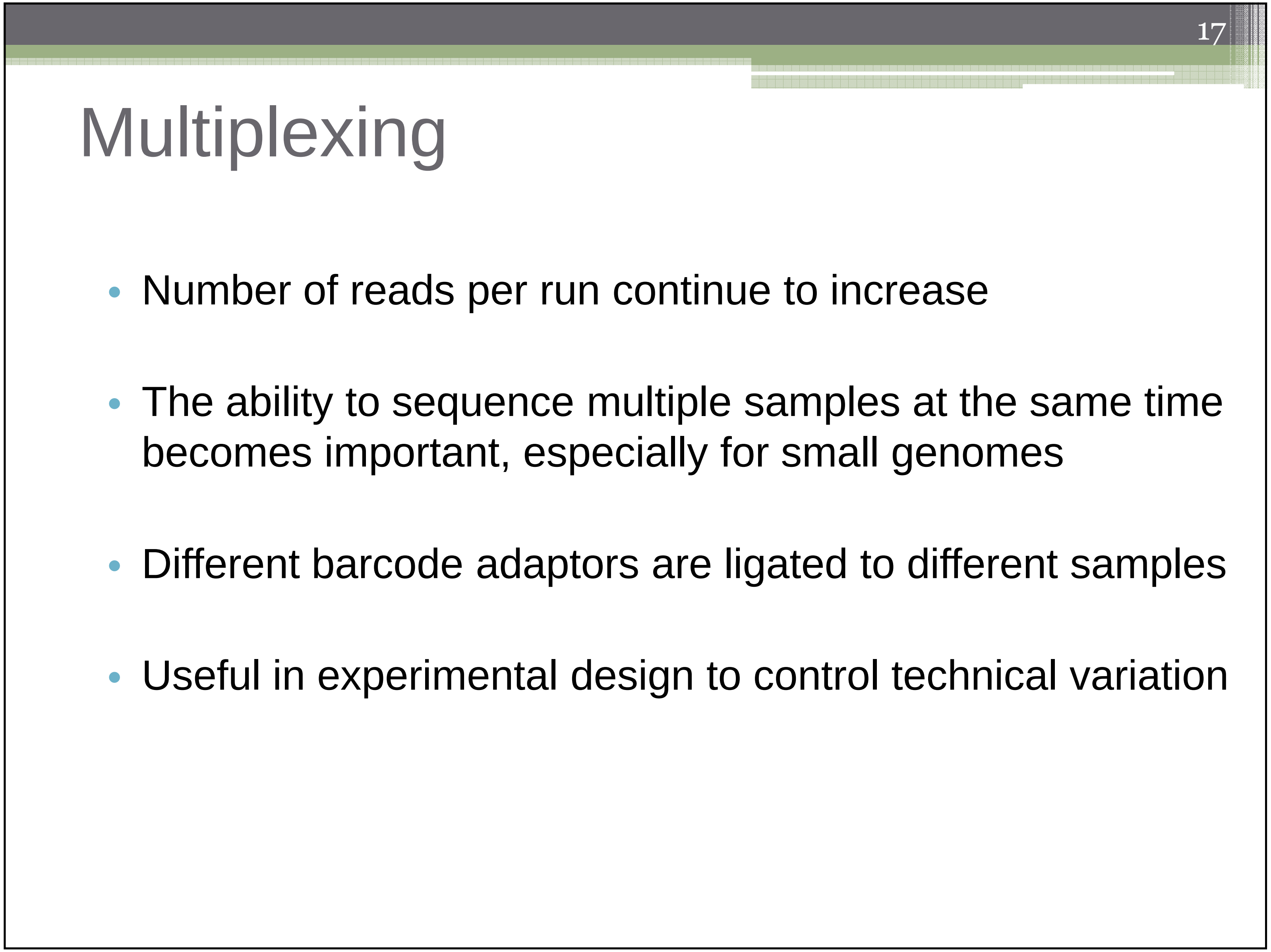
Saturation: MACS “diag” table

		% of peaks covered after sampling 90% … 20% of the total tags							
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



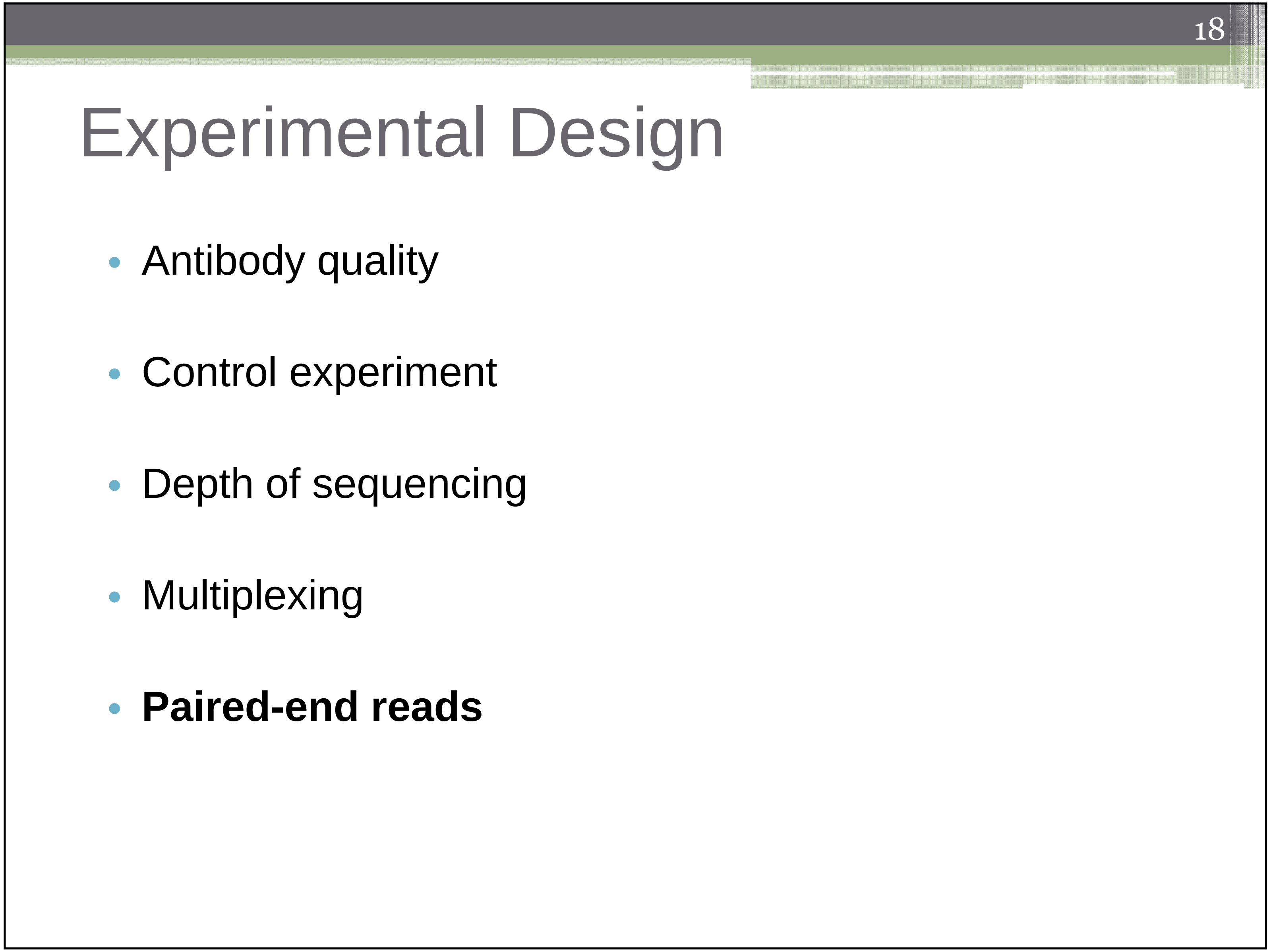
Experimental Design

- 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
- Useful in experimental design to control technical variation

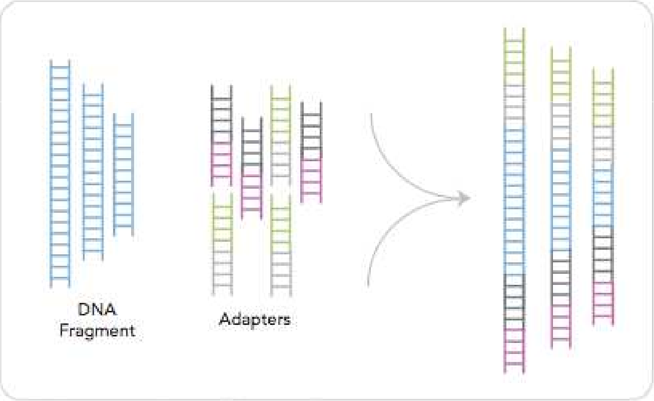


Experimental Design

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

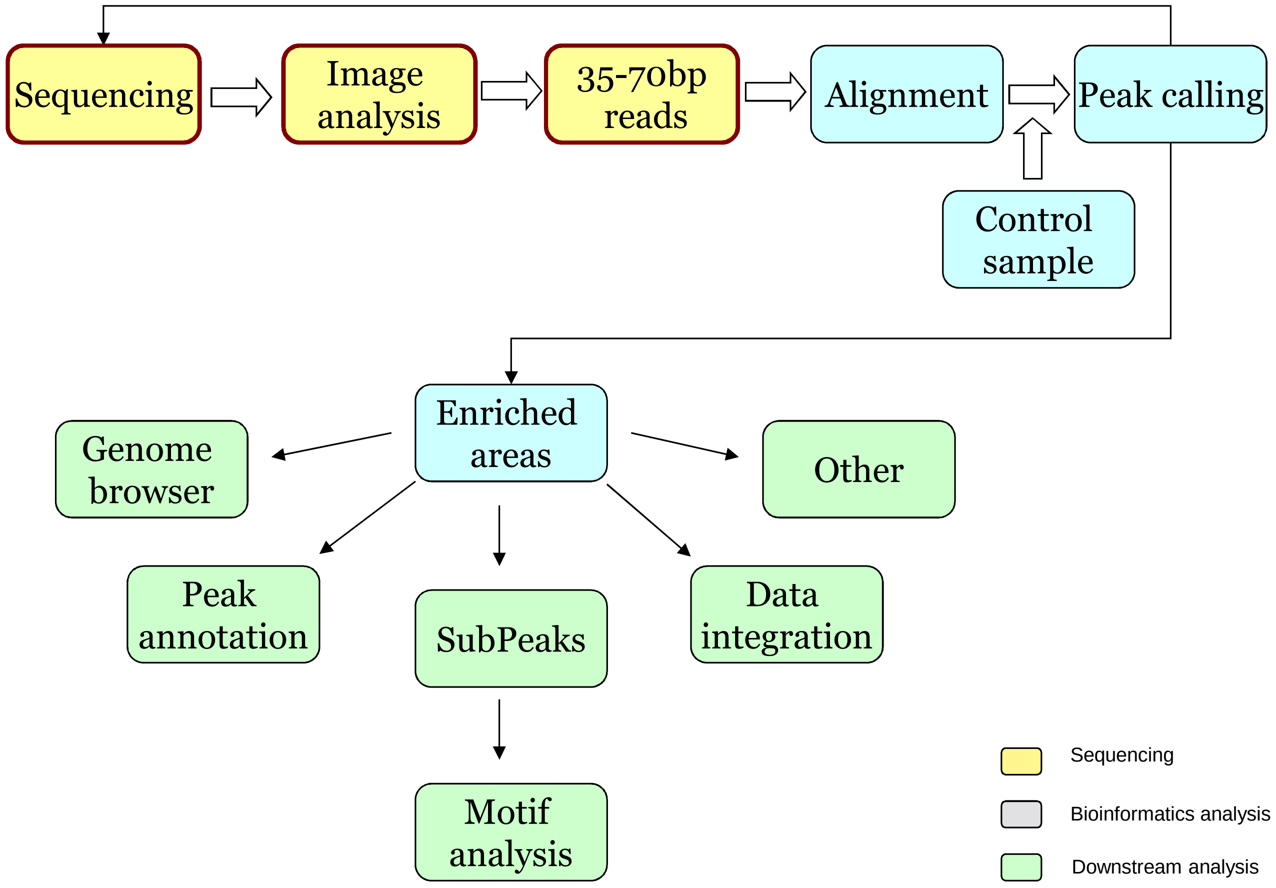
Paired-end sequencing

- 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
- Can assist in identifying duplicates

Analysis - Overview



Sequencing

Image analysis

35-70bp reads

Alignment

Peak calling

Control sample

Enriched areas

Genome browser

Other

Peak annotation

SubPeaks

Data integration

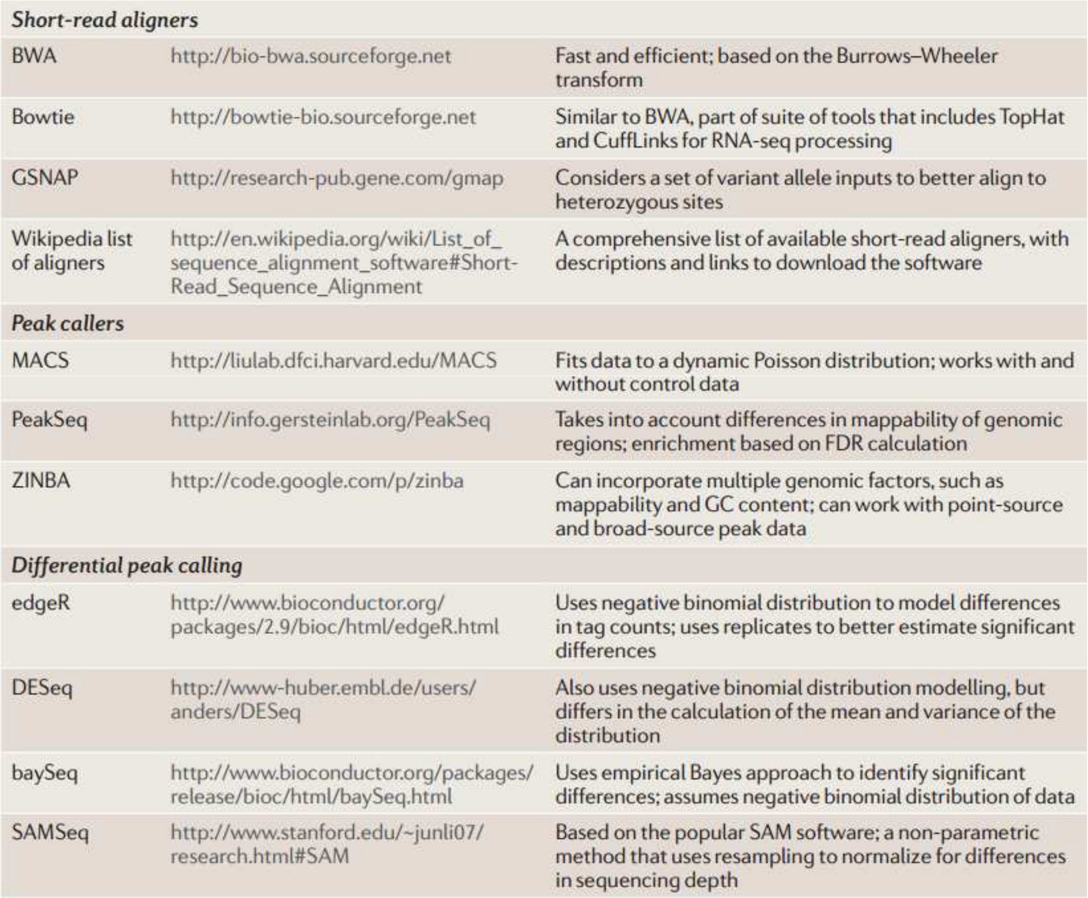
Motif analysis

🟨 Sequencing

⬜ Bioinformatics analysis

🟩 Downstream analysis

Analysis - Overview

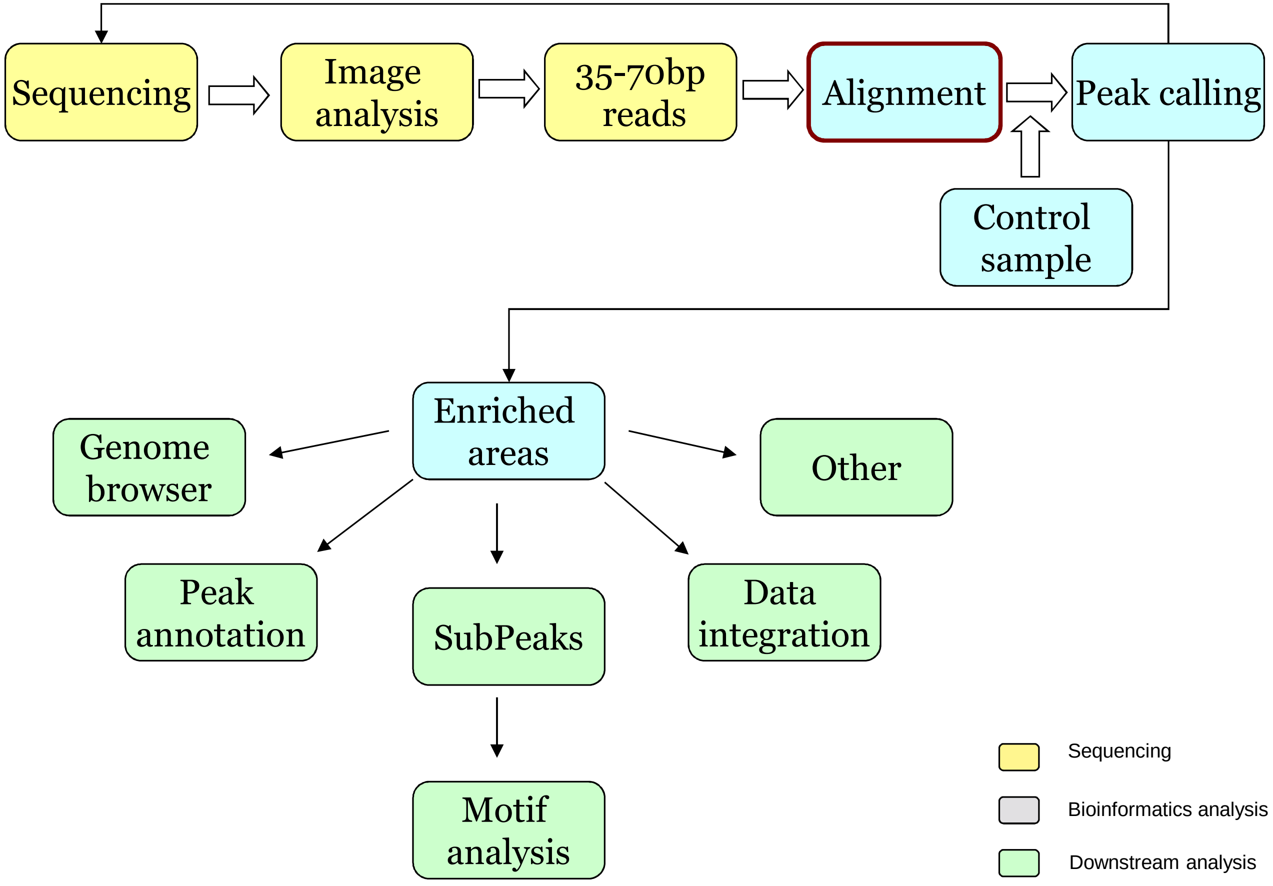


Short-read aligners		
BWA	http://bio-bwa.sourceforge.net	Fast and efficient; based on the Burrows–Wheeler transform
Bowtie	http://bowtie-bio.sourceforge.net	Similar to BWA, part of suite of tools that includes TopHat and CuffLinks for RNA-seq processing
GSNAP	http://research-pub.gene.com/gmap	Considers a set of variant allele inputs to better align to heterozygous sites
Wikipedia list of aligners	http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment	A comprehensive list of available short-read aligners, with descriptions and links to download the software
Peak callers		
MACS	http://liulab.dfci.harvard.edu/MACS	Fits data to a dynamic Poisson distribution; works with and without control data
PeakSeq	http://info.gersteinlab.org/PeakSeq	Takes into account differences in mappability of genomic regions; enrichment based on FDR calculation
ZINBA	http://code.google.com/p/zinba	Can incorporate multiple genomic factors, such as mappability and GC content; can work with point-source and broad-source peak data
Differential peak calling		
edgeR	http://www.bioconductor.org/packages/2.9/bioc/html/edgeR.html	Uses negative binomial distribution to model differences in tag counts; uses replicates to better estimate significant differences
DESeq	http://www-huber.embl.de/users/anders/DESeq	Also uses negative binomial distribution modelling, but differs in the calculation of the mean and variance of the distribution
baySeq	http://www.bioconductor.org/packages/release/bioc/html/baySeq.html	Uses empirical Bayes approach to identify significant differences; assumes negative binomial distribution of data
SAMSeq	http://www.stanford.edu/~junli07/research.html#SAM	Based on the popular SAM software; a non-parametric method that uses resampling to normalize for differences in sequencing depth

Furey, TS, 2012

Nature Reviews | Genetics

Analysis - Overview



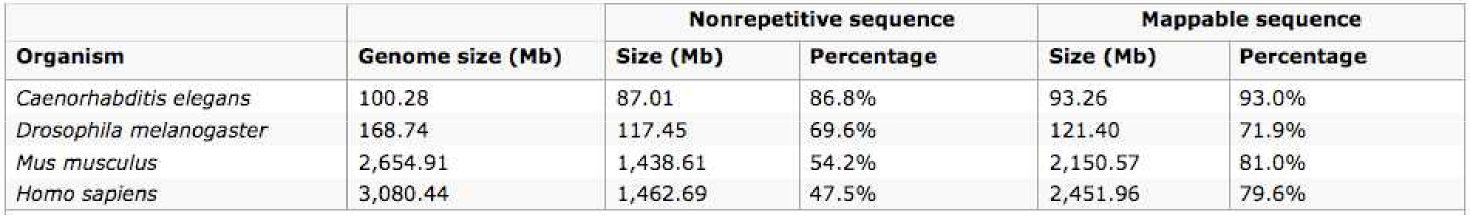
🟨 Sequencing

⬜ Bioinformatics analysis

🟩 Downstream analysis

Mappability

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



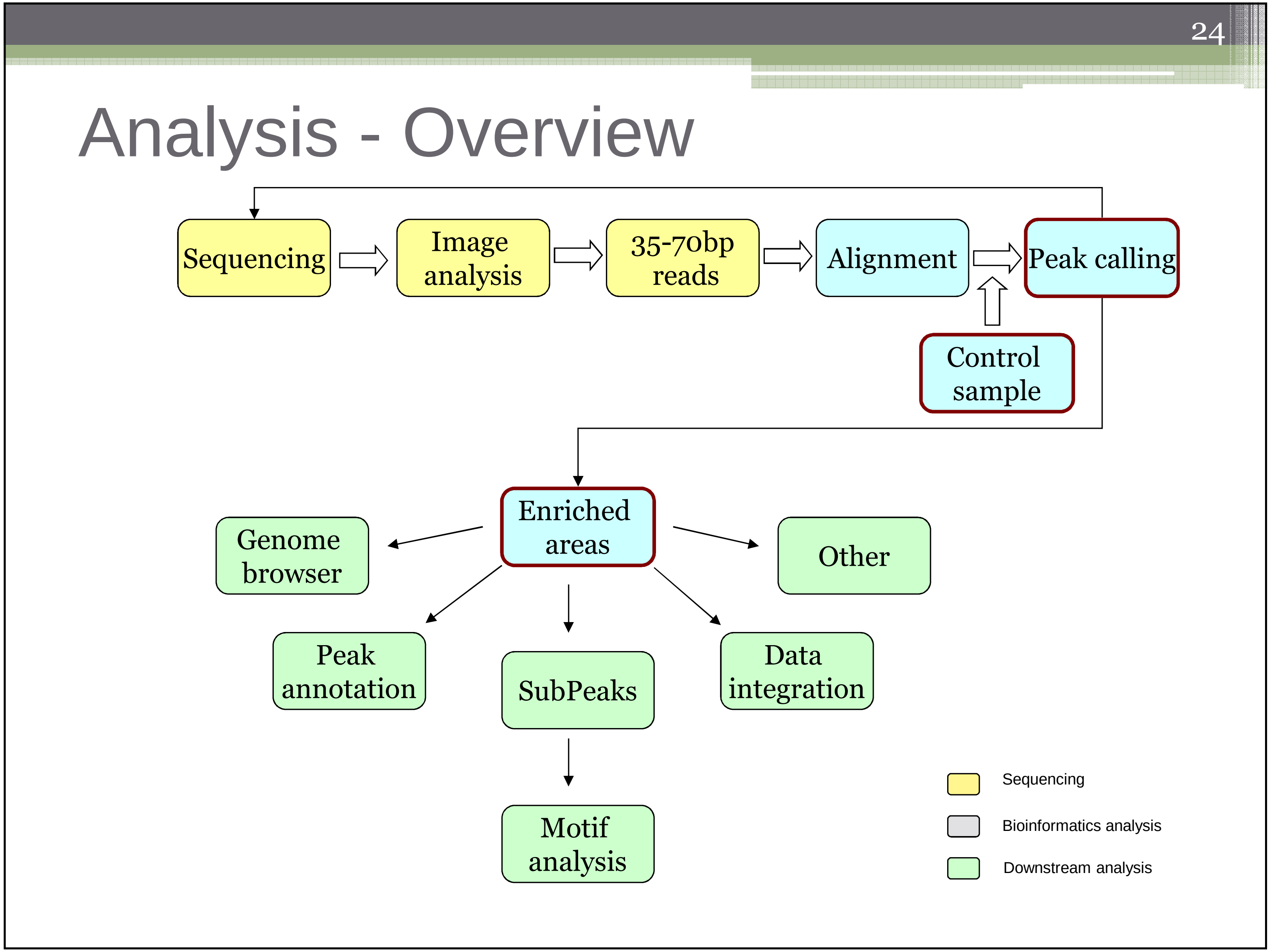
		Nonrepetitive sequence		Mappable sequence	
Organism	**Genome size (Mb)**	**Size (Mb)**	**Percentage**	**Size (Mb)**	**Percentage**
Caenorhabditis elegans	100.28	87.01	86.8%	93.26	93.0%
Drosophila melanogaster	168.74	117.45	69.6%	121.40	71.9%
Mus musculus	2,654.91	1,438.61	54.2%	2,150.57	81.0%
Homo sapiens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



Analysis - Overview

Sequencing

Image analysis

35-70bp reads

Alignment

Peak calling

Control sample

Enriched areas

Genome browser

Other

Peak annotation

SubPeaks

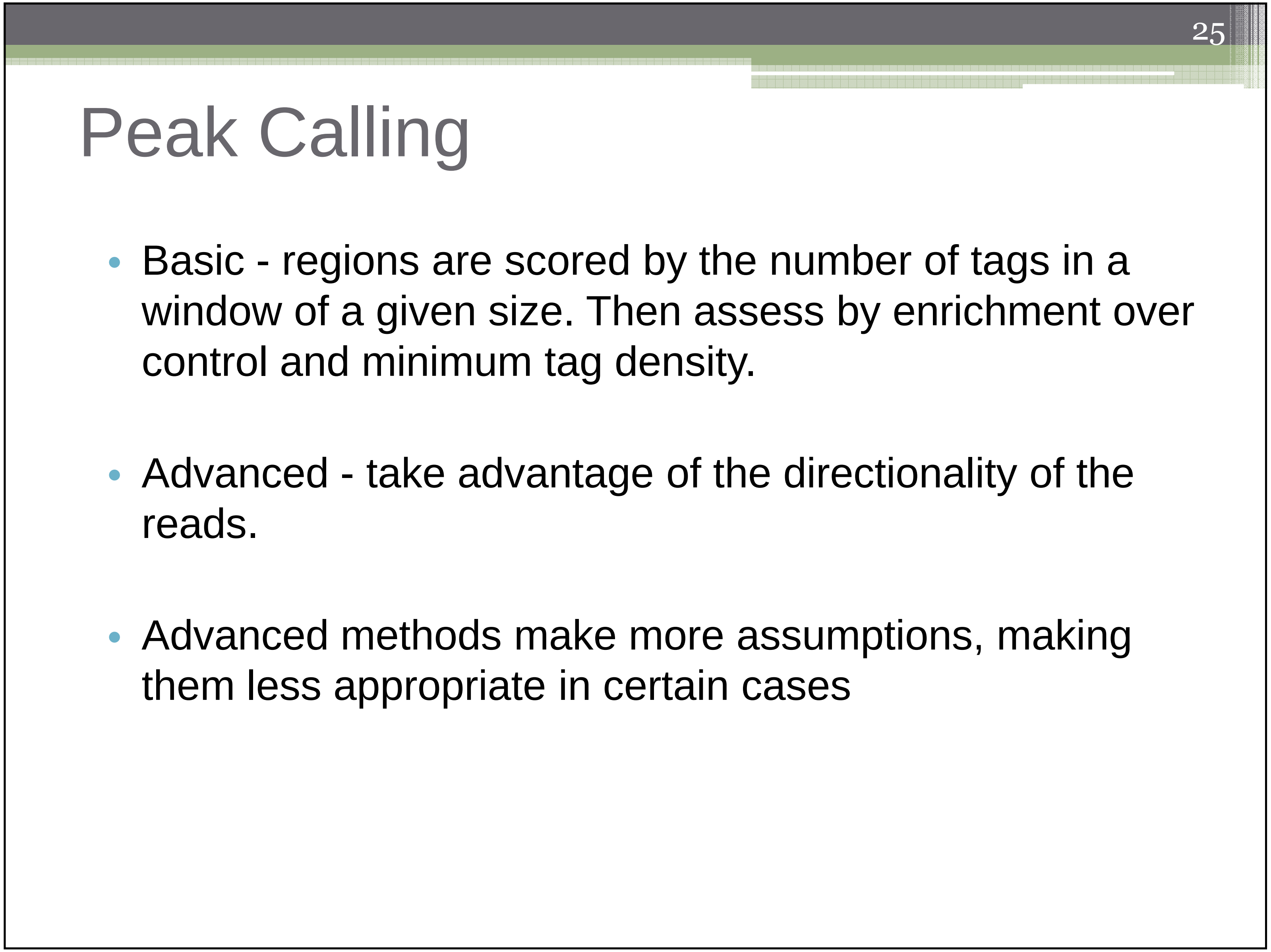
Data integration

Motif analysis

🟨 Sequencing

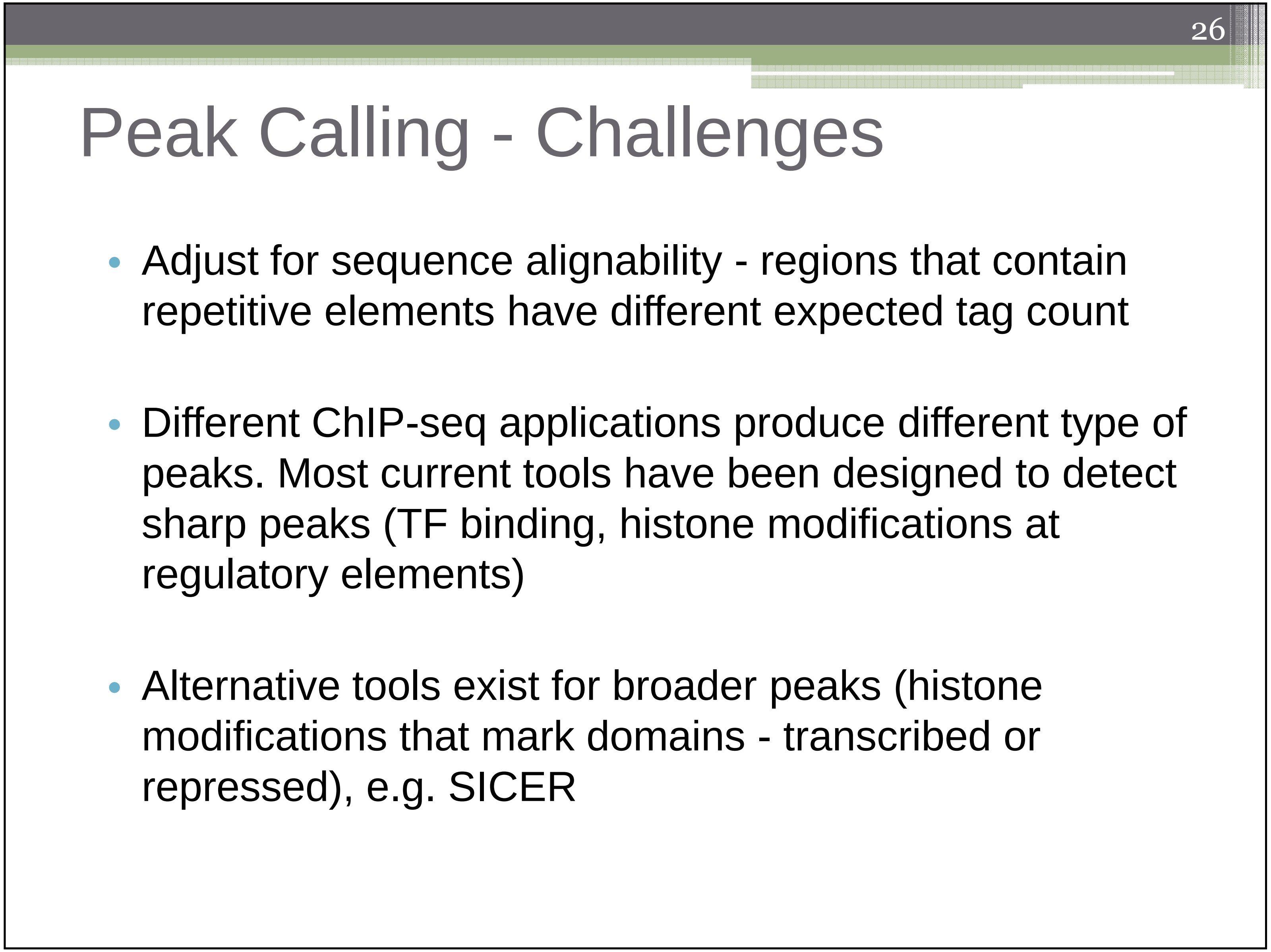
⬜ Bioinformatics analysis

🟩 Downstream analysis



Peak Calling

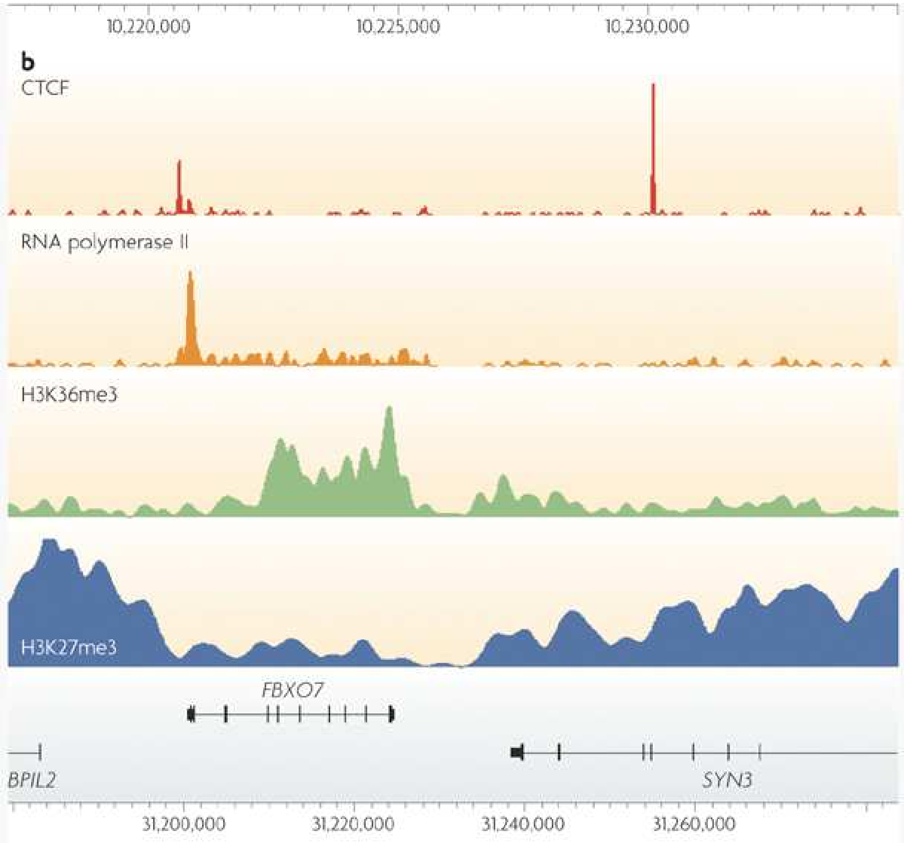
- Basic - regions are scored by the number of tags in a window of a given size. Then assess by enrichment over control and minimum tag density.
- Advanced - take advantage of the directionality of the reads.
- Advanced methods make more assumptions, making them less appropriate in certain cases



Peak Calling - Challenges

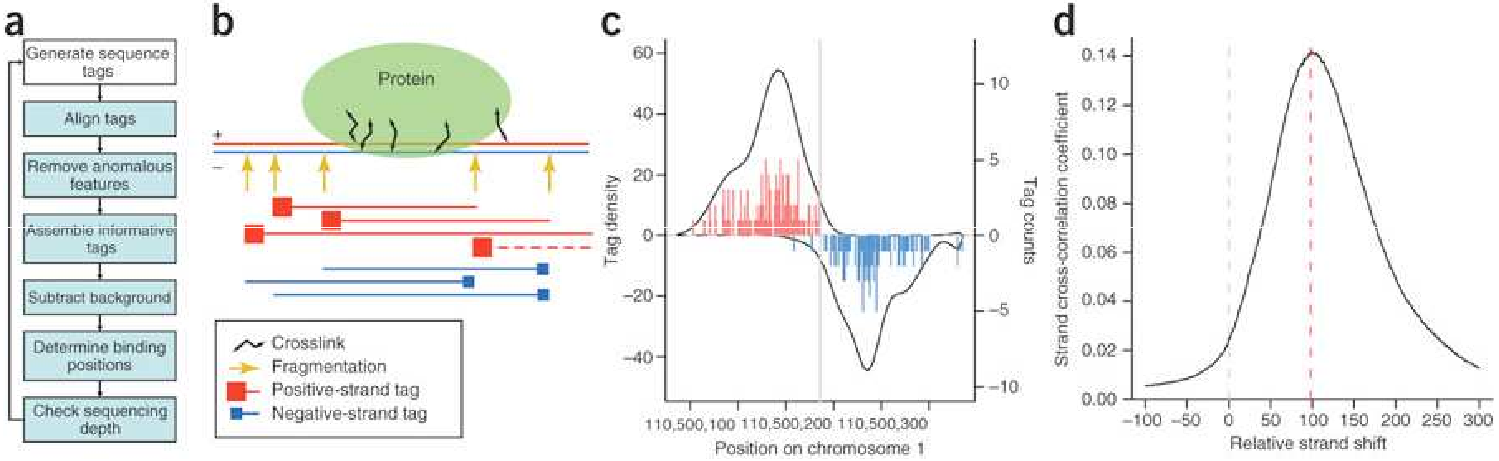
- Adjust for sequence alignability - regions that contain repetitive elements have different expected tag count
- Different ChIP-seq applications produce different type of peaks. Most current tools have been designed to detect sharp peaks (TF binding, histone modifications at regulatory elements)
- Alternative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

Peak Profiles



Park J, Nature Reviews Genetics, 2009

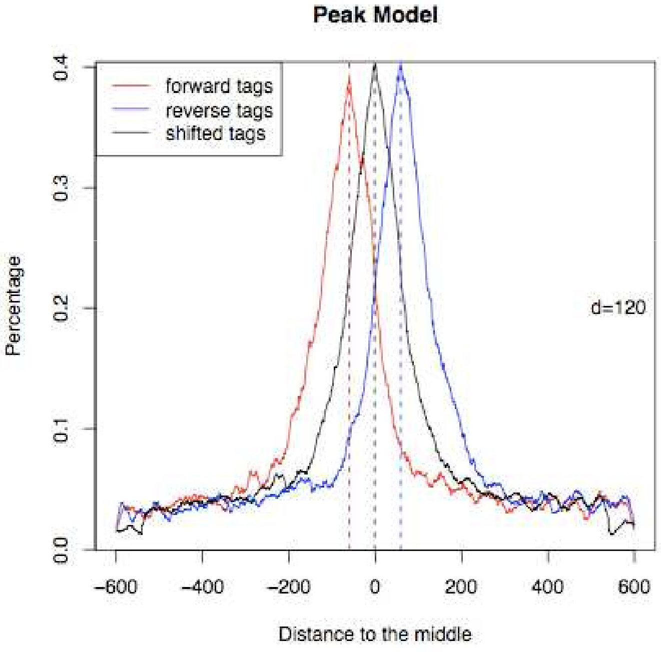
Strand Specific Profile

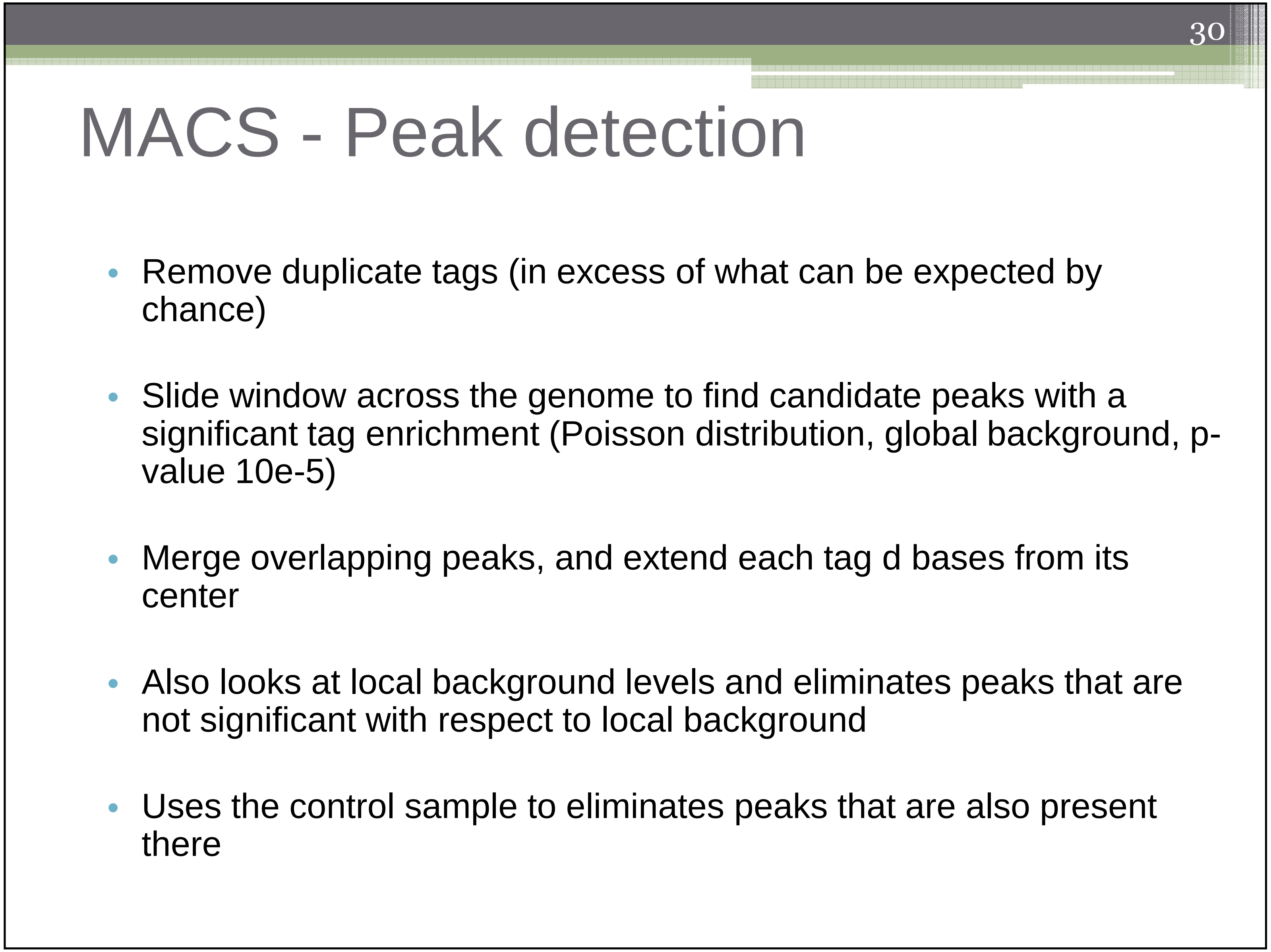


Kharchenko, Nature Biotechnology, 2008

MACS tool

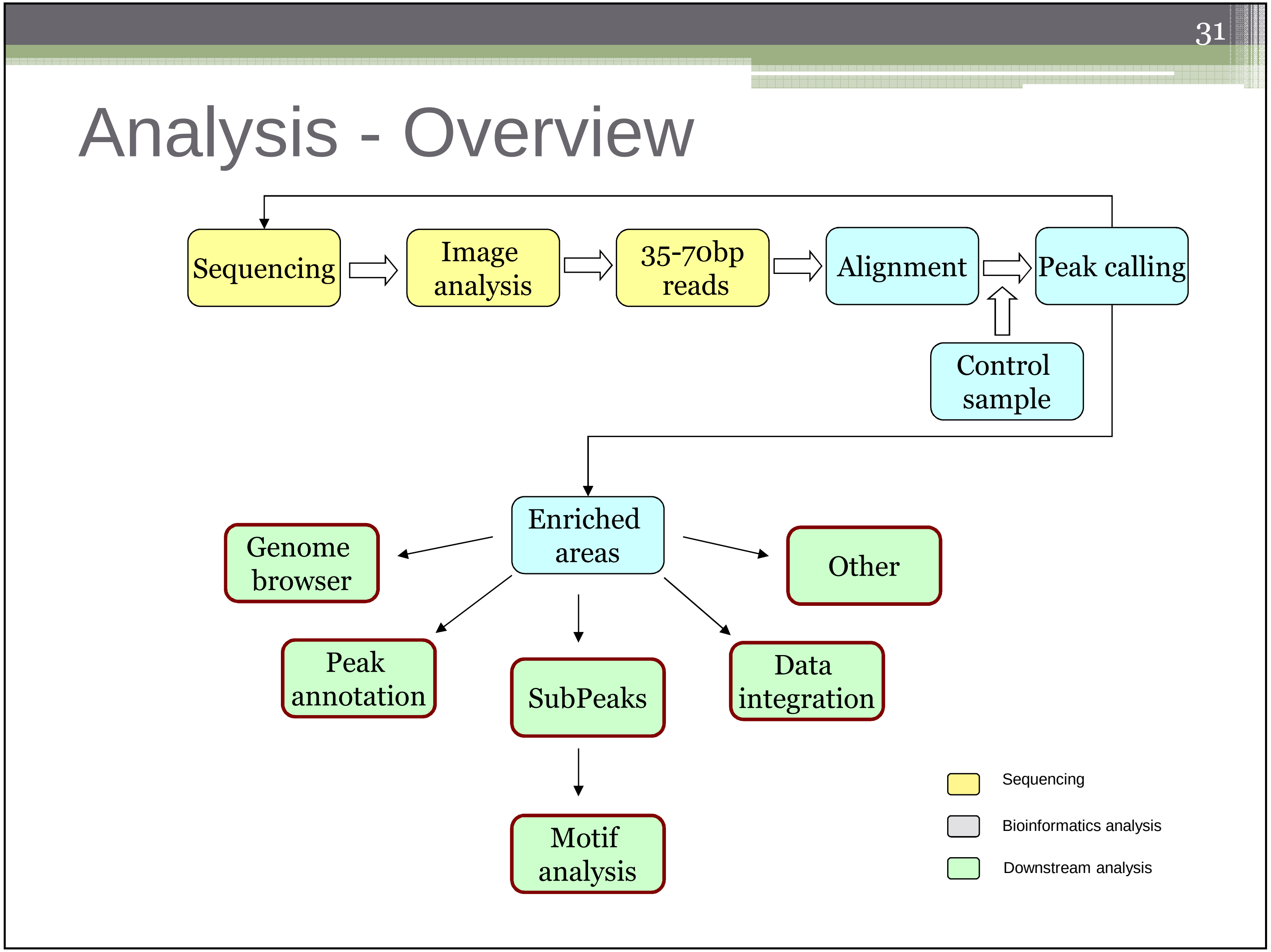
- Model the shift size between +/- strand tags
  - Scan the genome to find regions with tags more than m-fold enriched relative to random tag distribution
  - Randomly sample 1000 of these (high quality peaks) and calculate the distance between the modes of their +/- peaks
  - Shift all the tags by d/2 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



Analysis - Overview

Sequencing

Image analysis

35-70bp reads

Alignment

Peak calling

Control sample

Enriched areas

Genome browser

Other

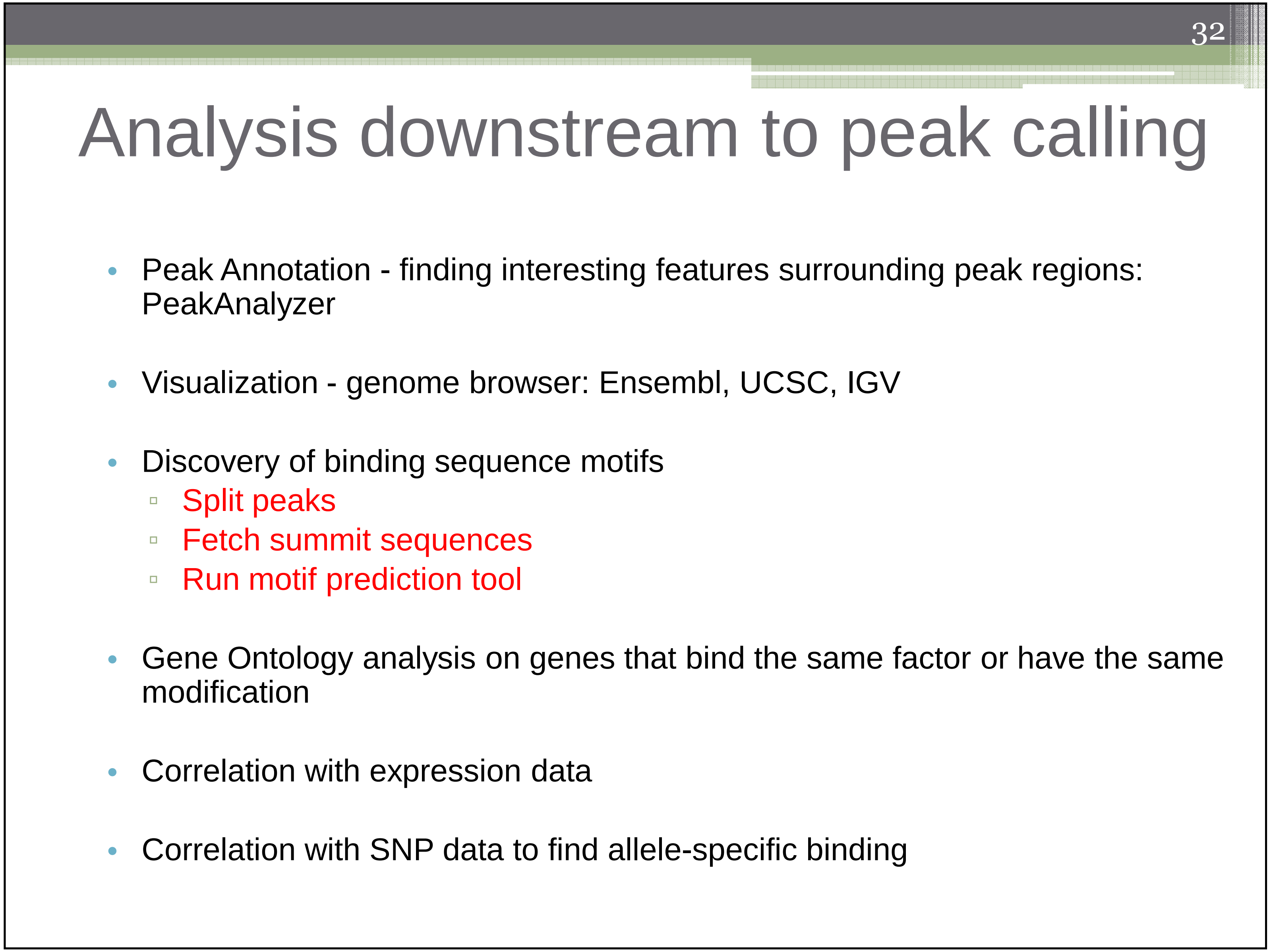
Peak annotation

SubPeaks

Data integration

Motif analysis

🟨	Sequencing
⬜	Bioinformatics analysis
🟩	Downstream analysis



Analysis downstream to peak calling

- Peak Annotation - finding interesting features surrounding peak regions: PeakAnalyzer

- Visualization - genome browser: Ensembl, UCSC, IGV

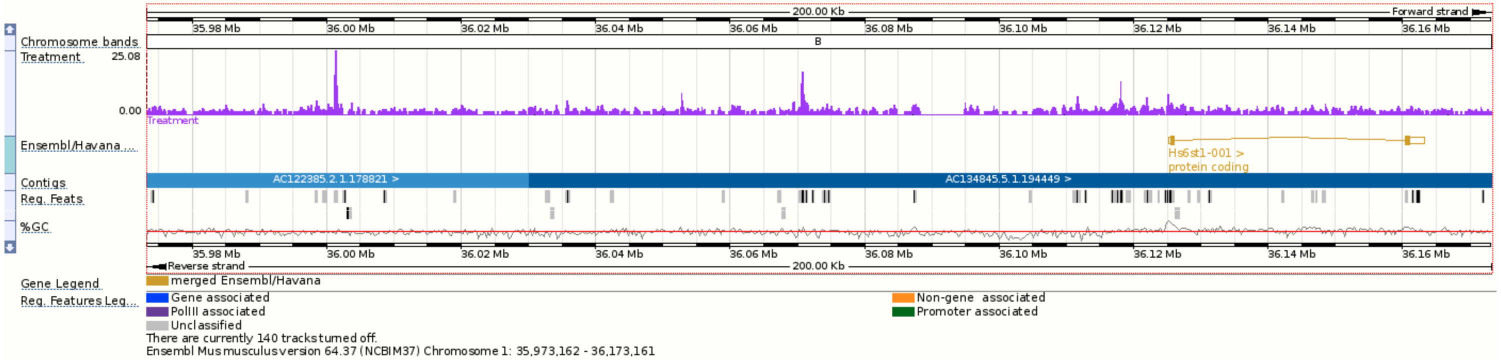
- Discovery of binding sequence motifs
  - Split peaks
  - Fetch summit sequences
  - Run motif prediction tool

- Gene Ontology analysis on genes that bind the same factor or have the same modification

- Correlation with expression data

- Correlation with SNP data to find allele-specific binding

Visualization in a genome browser

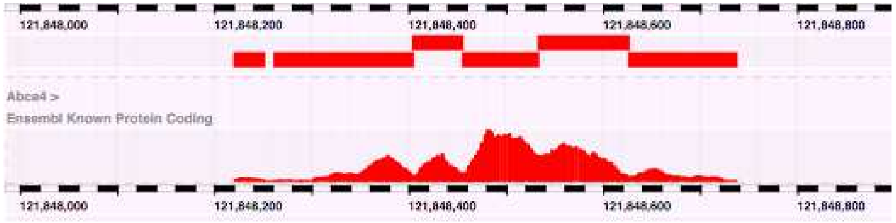


🟧 merged Ensembl/Havana	
🟦 Gene associated	🟧 Non-gene associated
🟪 PolIII associated	🟩 Promoter associated
⬜ Unclassified	

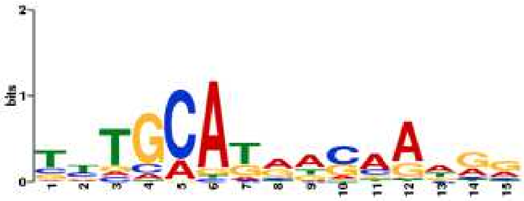
There are currently 140 tracksturned off.

Ensembl Mus musculusversion 64.37 (NCBIM37) Chromosome 1: 35,973,162 - 36,173,161

Motif Analysis



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



Bioplatforms Australia

CSIRO

Thank you

EMBL-EBI