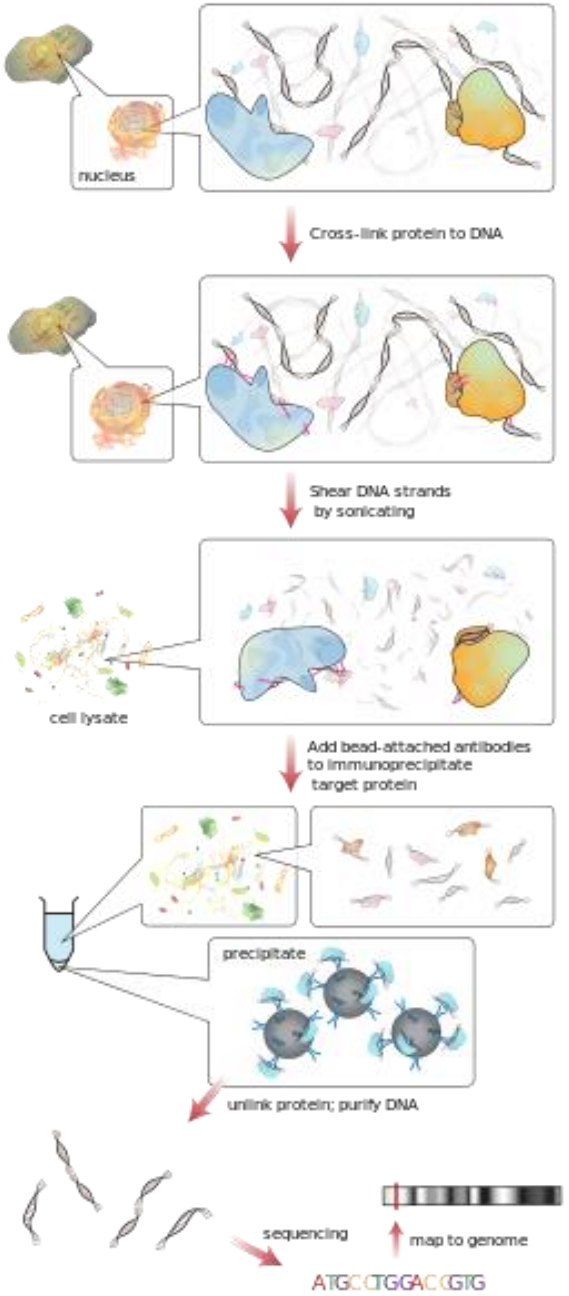
ChIP-Seq Workflow

Oxford Biomedical Data Science Training Programme

ChIP-seq

- **Ch**romatin **i**mmuno**p**recipitation combined with high throughput **seq**uencing
- Method used to study DNA-protein interactions
  - Crosslink proteins and DNA
  - Lyse cells and shear DNA
  - Enrich for protein / epitope of interest using antibody
  - Reverse crosslinks and purify DNA
  - Add adaptors and sequence DNA
- Used for:
  - Transcription factors
  - Chromatin marks



ChIP-seq Analyses

- Identify transcription factor binding sites
  - Genomic context – promotors, enhancers
  - Target genes
  - New sequence motifs
  - Differential binding (e.g. during development, different conditions)
- Chromatin state
  - Combination of chromatin marks to infer functional state
  - Active, poised, repressed

Sources of Bias

- Starting material
  - Do you have enough material? ChIP-seq is a population average – need many cells for robust results
  - Is this consistent between replicates/conditions you are testing?
- Fragmentation of the library
  - Requires optimization – tissue / target specific
  - Sonication
    - open chromatin shears more easily
  - Enzymatic
    - Mnase, Tagmentation – sequence bias
- Immunoprecipitation
  - Antibody - is it specific? Is it efficient?
  - Use a cocktail of antibodies
    - Target different epitopes of the same protein
    - Avoid epitope masking
- Batch effects
  - starting material, fragmentation, antibody...

ChIP-seq Controls

- Mock
  - No chromatin just the antibody of interest
- No Antibody
  - Chromatin but IP with IgG
- Input control
  - Chromatin not immuno-precipitated
- Spike in
  - External control e.g. Drosophila DNA

Very little material
Worth sequencing?

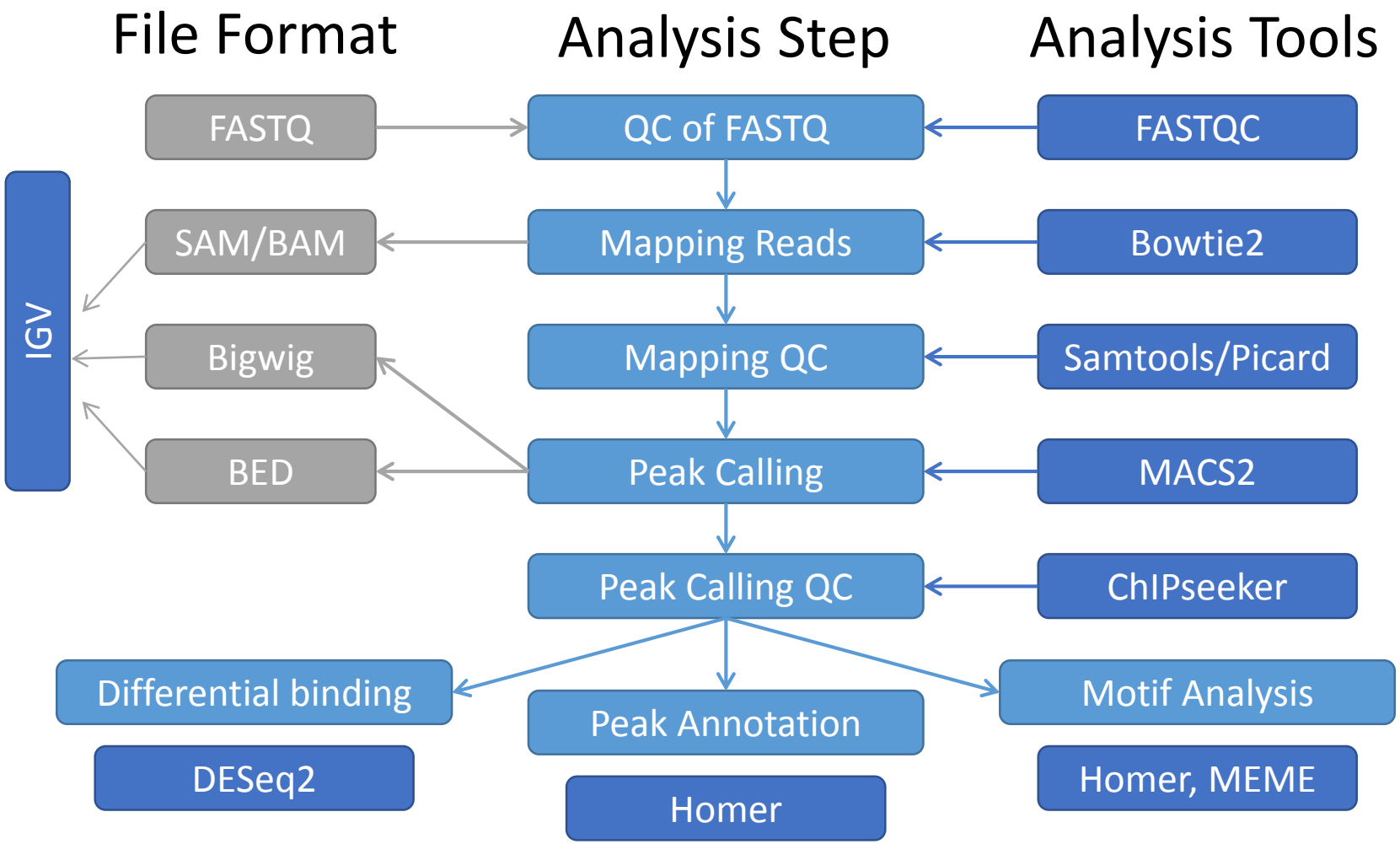
Fragmentation bias
A lot of material – deeper sequencing
One input per sample

Normalise for large changes across conditions

Sequencing Considerations

- Single-end or paired-end
  - Paired end better for removal of duplicates
- Read length
  - Long enough for unique mapping
- Sequencing depth
  - Greater for input (2x IP)
  - Greater for broad peaks e.g. H3K27me3
  - Encode guidelines
    - Narrow 20m mapped reads
    - Broad 45m mapped reads

ChIP-seq Analysis Workflow



Mapping QC

- Read mapping statistics
  - > 90% mapping
  - Mapping context – intergenic / promoter regions
- Read filtering
  - Mitochondrial reads
  - Duplicates
  - Multi-mapping reads
  - Reads that are not properly paired (map too far apart)

Check numbers lost at each filter

Peak Calling

- Identify regions of the genome with more reads than control or local background regions
- Pool samples for peak calling
  - Better coverage
  - Increase peak calling ability
  - Look at overall properties of dataset
  - Especially useful for inputs
- Caveats
  - Samples should be of similar depth
  - If not one sample might influence peak calling more than the others
  - Down-sampling may be required

Peak Calling Tools

- MACS2
  - most widely used
  - Broad & Narrow marks
- SICER – good for broad peaks
- LanceOtron – machine-learning approach
- Homer, SPP, PeakRanger…

Peak Calling QC

- Reproducibility
  - Overlap replicates (Bedtools) & keep peaks 2+ samples
- Irreproducible Discovery Rate
  - Peter Bickle/ Anshul Kundaje (Encode Project)
  - Rank peaks on p-value
  - Identify reproducible peaks based on correlation of ranks
- Filter out black-listed regions

Peak Annotation

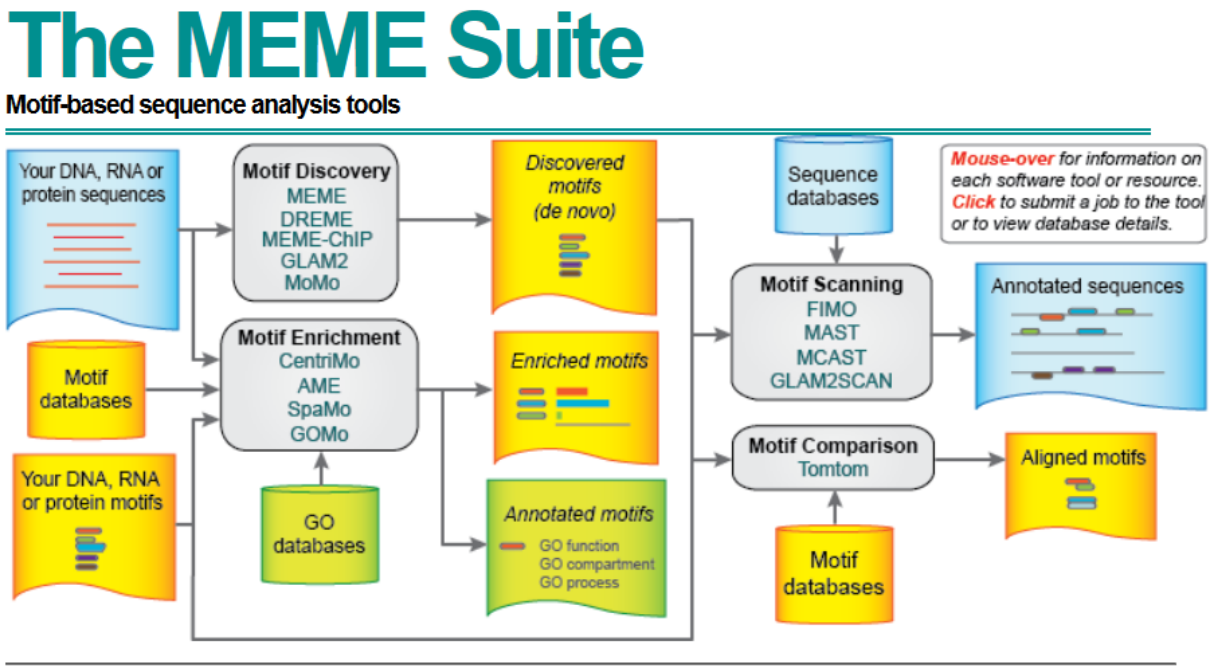
- Genomic context
  - Promoter, enhancer, CpG island, repeat, exon, intron
  - Enrichment in different feature classes
  - Homer
  - GREAT - http://great.stanford.edu/public/html/
- Target genes
  - Nearest neighbour (GREAT, Homer)

Differential Binding

- featureCounts
  - Count reads under peaks in each condition
  - Can also test all regions using a window approach
- DESeq2 / edgeR
  - Treat like RNAseq data
  - Must have >=3 replicates
- R-package Diffbind provides a wrapper for both counting and statistical analysis
  - Black box
  - Limited flexibility

Motif Analysis

- Look for enrichment in short sequences within peaks
- Compare with known transcription factor motifs
- Location of motif within peaks
- Proportion of peaks containing motif
- Co-occurring motifs



http://meme-suite.org/

HOMER Motif Analysis

http://homer.ucsd.edu/homer/motif/

Viewing Peaks

- Bam
  - Checkout the raw reads
  - Very memory intensive
- Bedcoverage
  - Breaks genome into regions and gives coverage score
  - Can make using Bedtools
  - Plain text
  - Less memory intensive then bam
- Bigwig
  - Like bedcoverage but compressed & indexed for random access
  - Faster, less memory intensive

Exercise

- Download data from:
- https://www.cell.com/cell-reports/pdf/S2211-1247(17)31705-9.pdf
- https://www.ebi.ac.uk/ena/data/view/PRJNA262583
- Experimental design:
  - 2 x Dexamethasone treated
  - 2 x untreated
  - All ChIP samples have an input control
  - Samples: SRX716927 - SRX716934
- Build a pipeline to:
  - Call peaks (MACS2)
  - Filter blacklisted peaks (Bedtools)
  - Count reads under peaks (Bedtools)
  - Compare replicate overlap (Bedtools)
  - Annotate peaks (Homer)
  - Motif analysis (Homer)
  - View outputs (IGV)