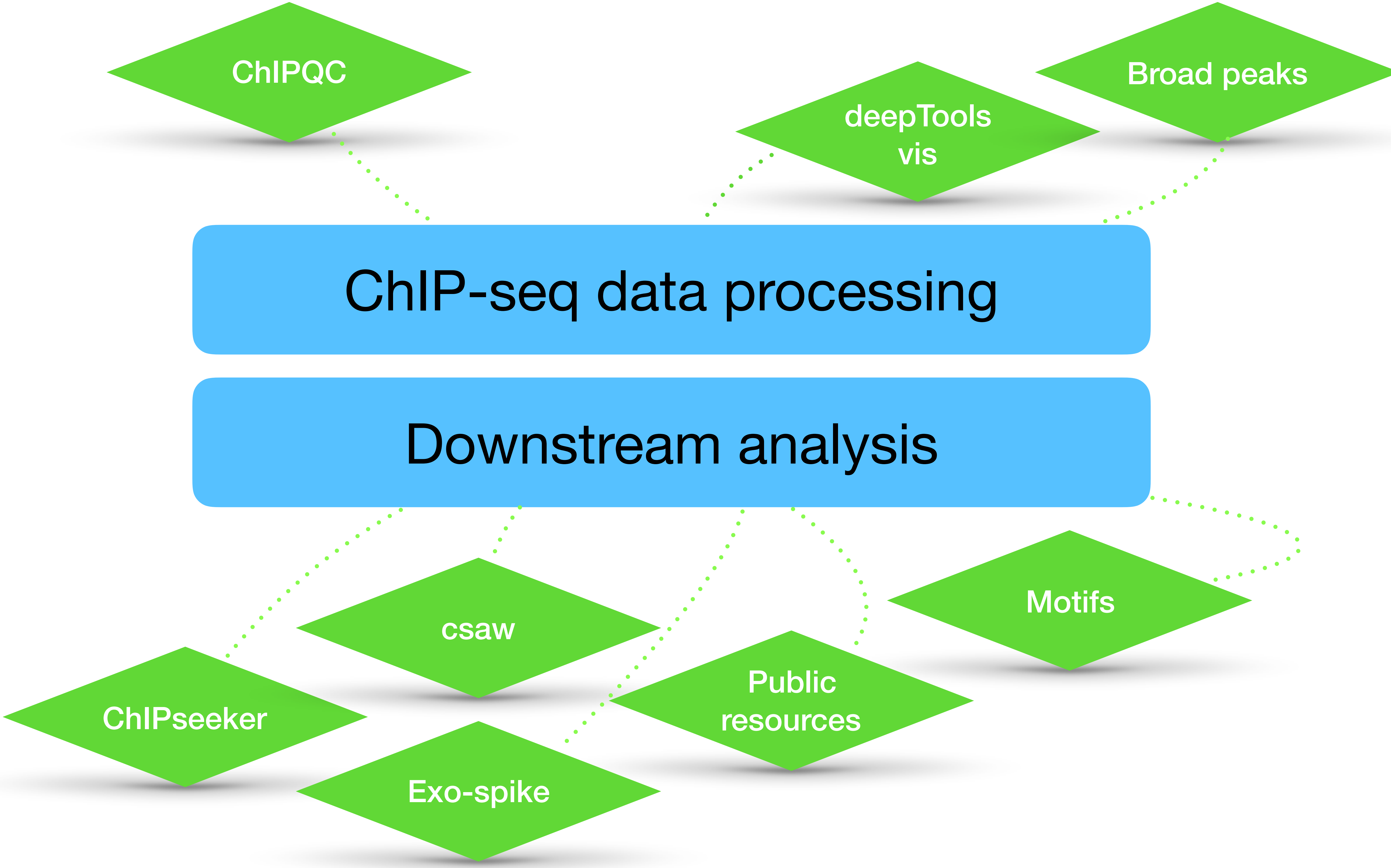
ChIP-seq data analysis

Introduction to practicals II: differential binding & functional annotations

Practicals



ChIPQC

deepTools vis

Broad peaks

ChIP-seq data processing

Downstream analysis

csaw

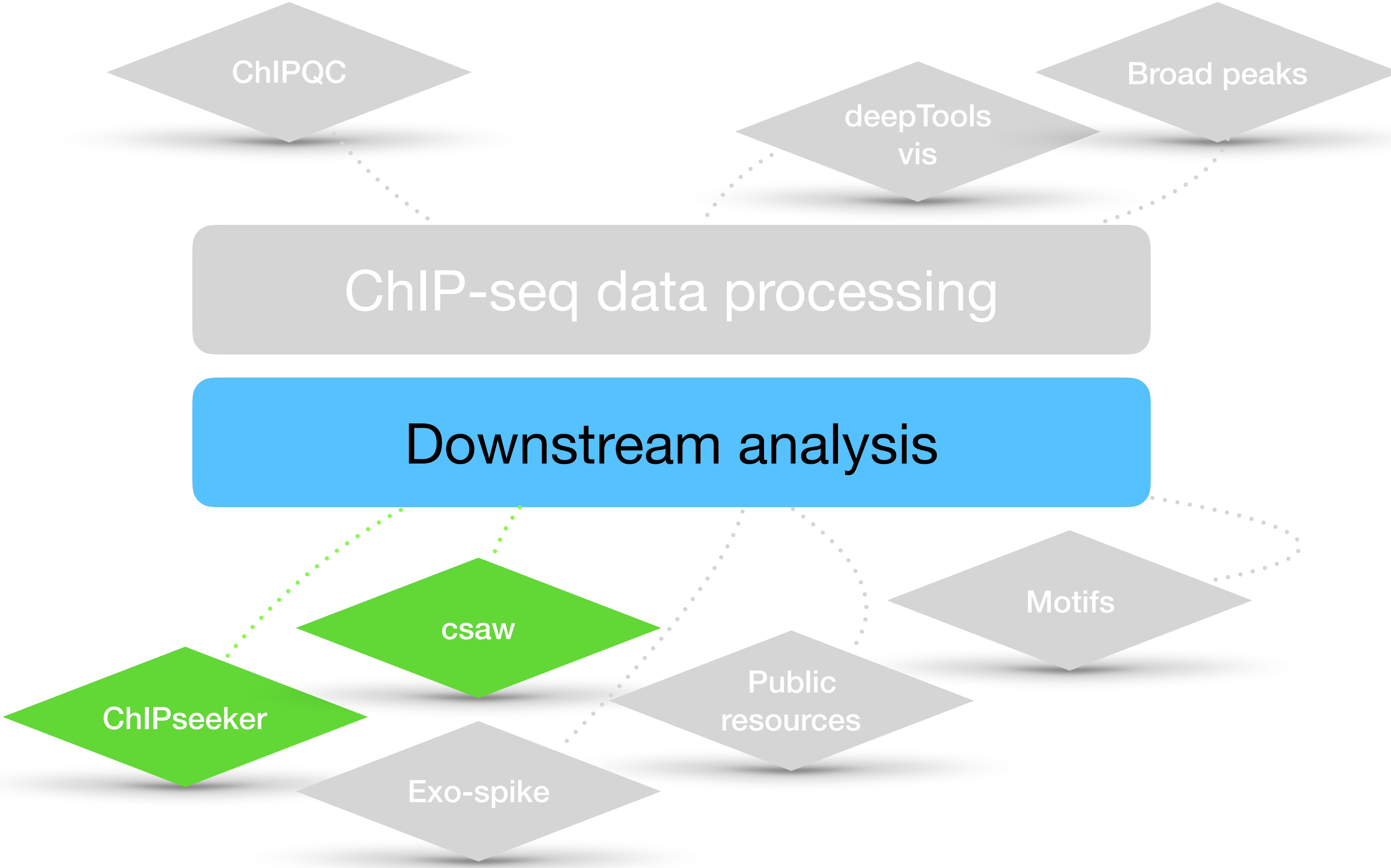
Motifs

Public resources

ChIPseeker

Exo-spike

Practicals



ChIPQC

deepTools vis

Broad peaks

ChIP-seq data processing

Downstream analysis

csaw

ChIPseeker

Motifs

Public resources

Exo-spike

Differential binding

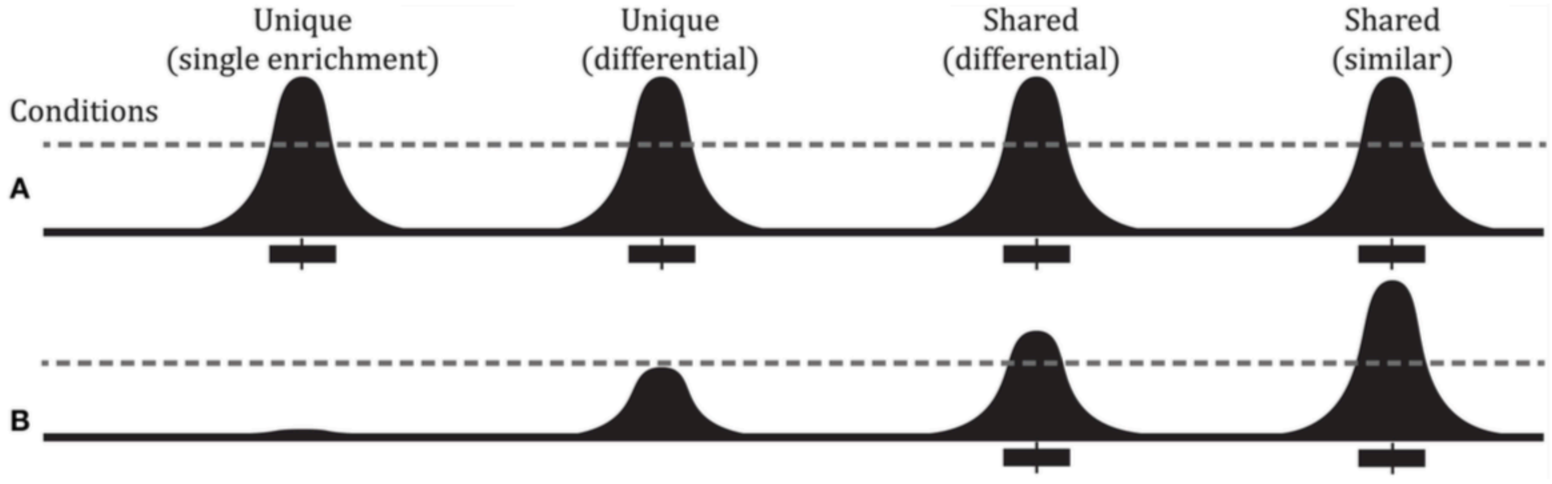


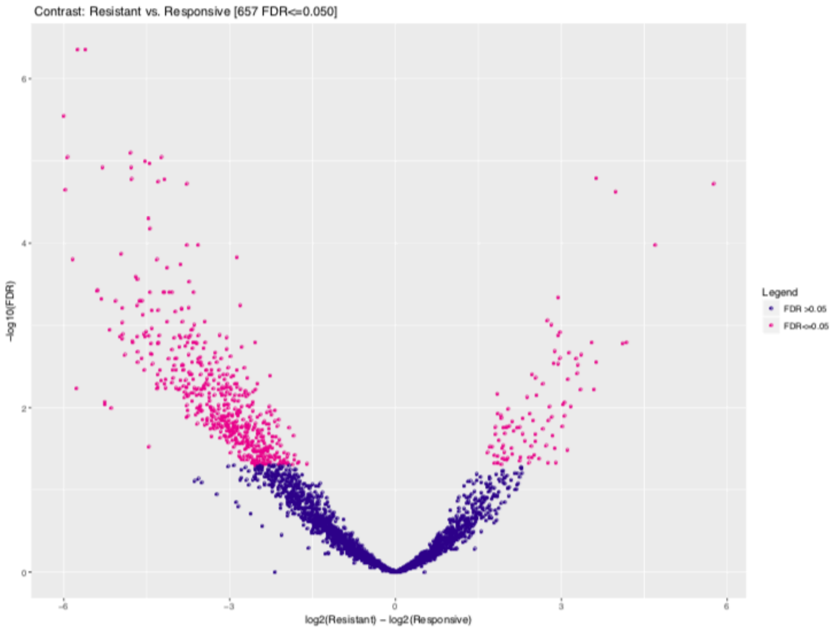
FIGURE 1 | Overview of peak types and defining reference binding regions. (A) Several different types of peak comparisons are shown. The black curve represents binding signal with the dotted line representing a hypothetical threshold for enrichment. The black boxes under each curve represent significant regions as defined by peak caller output with the vertical line in the box representing summit point. Comparing binding profiles in conditions **(A)** vs. **(B)** we find: binding in condition **(A)** but not **(B)** (Unique—single enrichment), varying degrees of binding between the two conditions (Unique and Shared peak—differential), and both conditions having a peak of about comparable signal intensity (Shared peak—similar).

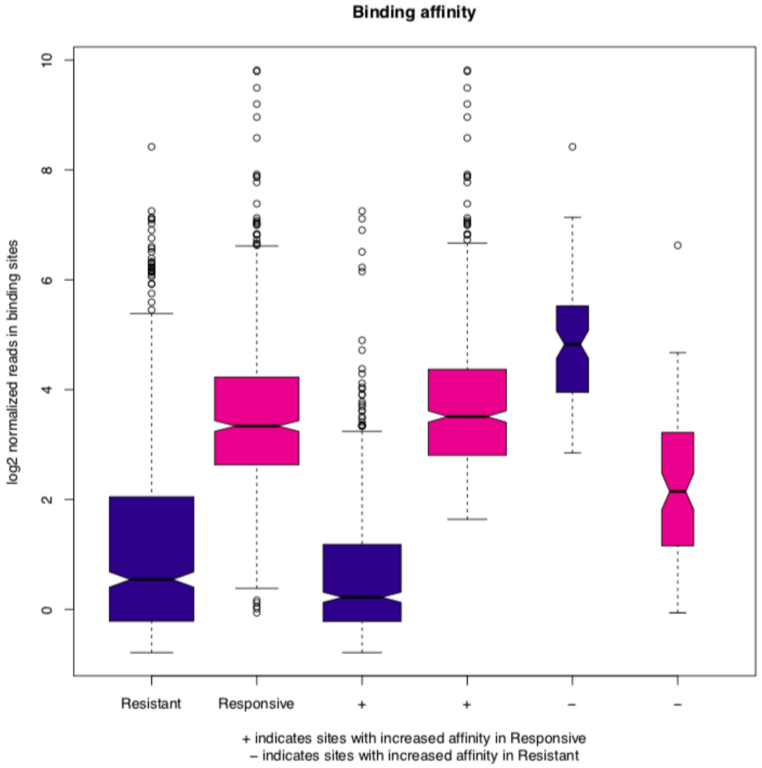
image source: Dai-Ying Wu et al. 2015, frontiers in Genetics

- Quantifying binding signal, e.g. in peaks regions
- Performing statistical analysis to discover quantitative changes between experimental groups
- i.e. to decide whether for a given region, an observed difference is significant, greater than would be expected just due to natural random variation

DiffBind

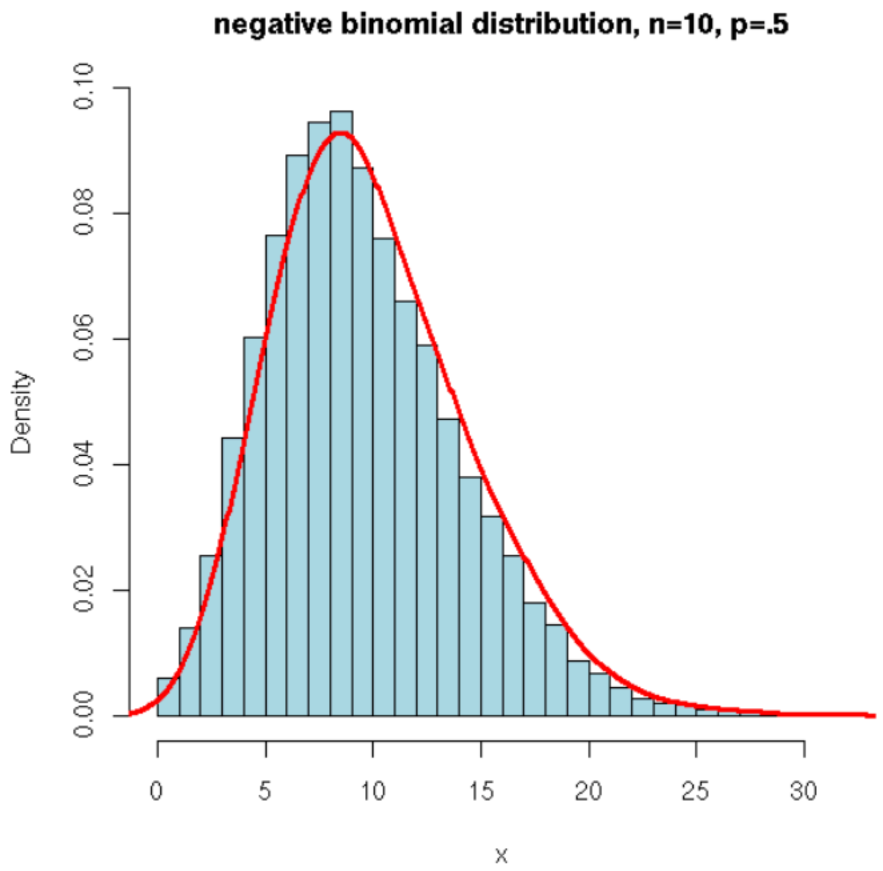
- helps define consensus peak set for analyses
- counts reads in the peaks regions
- calculate a binding matrix with scores based on read counts for every sample (normalised affinity scores)
- allows to set-up different contrasts for comparisons
- uses gene expression methods (edgeR or DESeq2) to compare regions





DiffBind: DESeq2

- matrix of raw counts is constructed for the contrast
- the raw number of reads in the control sample is subtracted
- library size is computed for use in subsequent normalisation, by default as in total number of reads in peaks
- dispersion is estimated
- nbinomWaldTest function is used to test for significance of coefficients in negative bionomial GLM model



Different flavours

Differential transcription factor binding

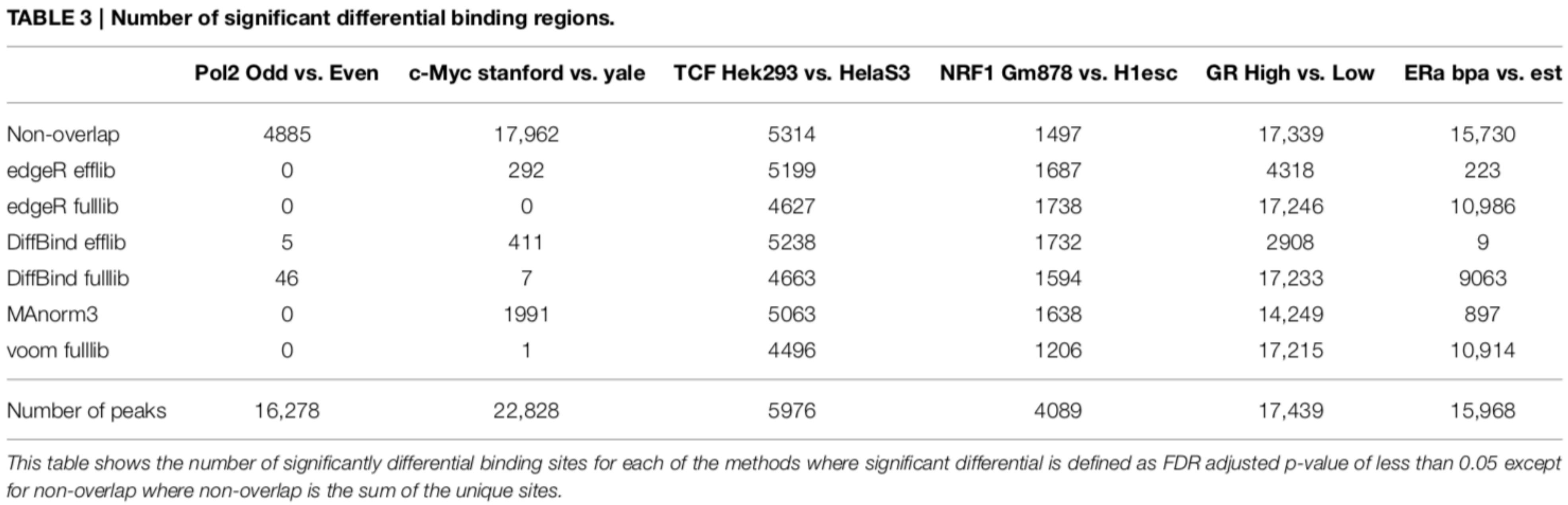


TABLE 3 | Number of significant differential binding regions.

	Pol2 Odd vs. Even	c-Myc stanford vs. yale	TCF Hek293 vs. HelaS3	NRF1 Gm878 vs. H1esc	GR High vs. Low	ERa bpa vs. est
Non-overlap	4885	17,962	5314	1497	17,339	15,730
edgeR efflib	0	292	5199	1687	4318	223
edgeR fulllib	0	0	4627	1738	17,246	10,986
DiffBind efflib	5	411	5238	1732	2908	9
DiffBind fulllib	46	7	4663	1594	17,233	9063
MAnorm3	0	1991	5063	1638	14,249	897
voom fulllib	0	1	4496	1206	17,215	10,914
Number of peaks	16,278	22,828	5976	4089	17,439	15,968

This table shows the number of significantly differential binding sites for each of the methods where significant differential is defined as FDR adjusted p-value of less than 0.05 except for non-overlap where non-overlap is the sum of the unique sites.

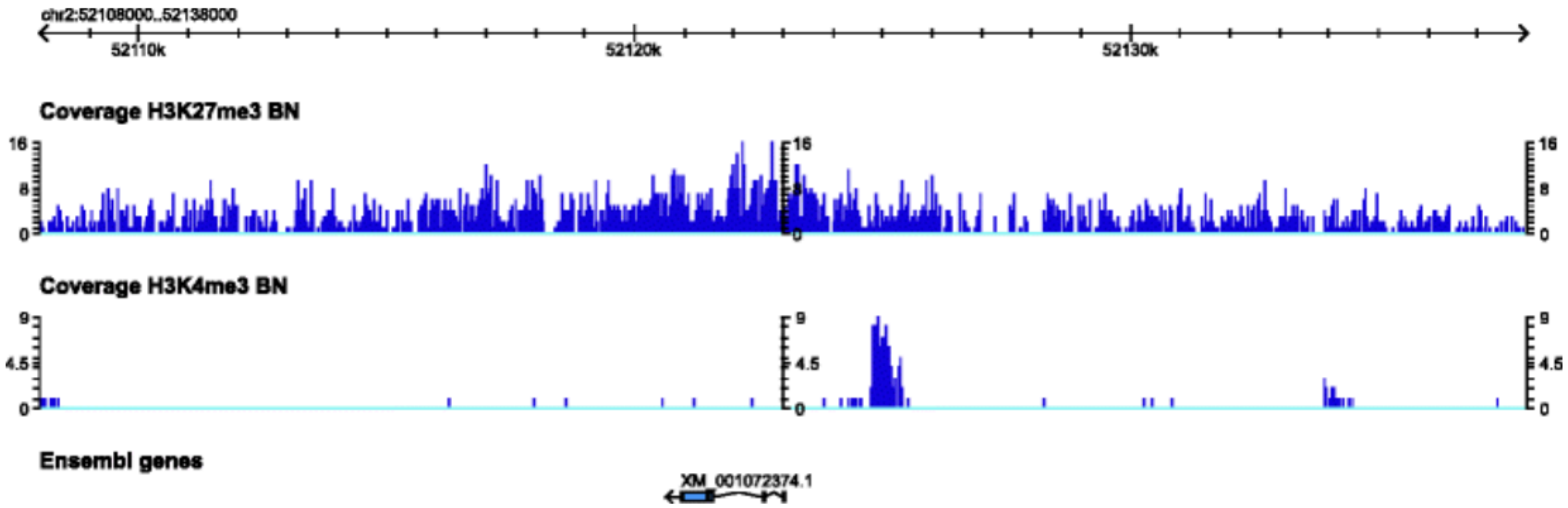
image source: Dai-Ying Wu et al. 2015, frontiers in Genetics
Identifying differential transcription factor binding in ChIP-seq

- Compared 6 ENCODE dataset to illustrate the impact of data processing under different study design
- The performance of normalisation methods depends strongly on the variation in total amount of protein bound between conditions, with total read count outperforming effective library size, when a large variation in binding was studied
- Use of input subtraction to correct for non-specific binding showed a relatively modest impact on the number of differentially peaks found and fold change accuracy
- Validation using fold-change estimates from qRT-PCR suggests there is still room for methods improvement…

csaw

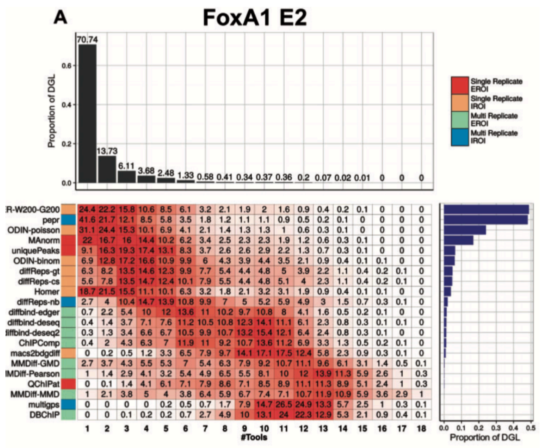
Different flavours

sliding windows: de novo detection



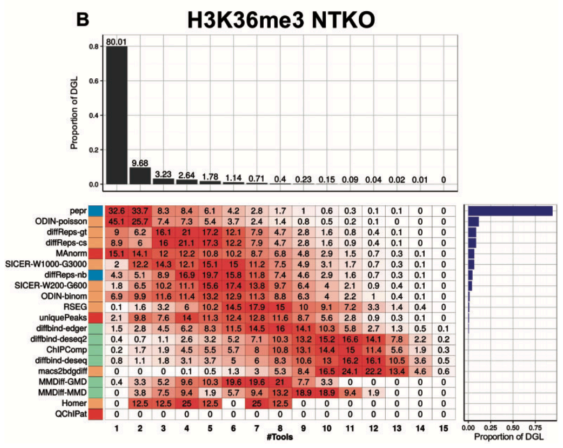
Example of ChIP-seq read coverage of H3K27me3 occurring in broad domains across the genome compared to other histone marks like H3K4me3 occurring in precisely defined peaks. Source: Heining et al., 2015, BMC Bioinformatics

- Region-derived or peaks-based differential binding may be problematic:
- if regions derived are not independent of the DB status fo these regions
- if regions are called with imprecise boundaries
- for protein-targets with broad enrichment, when histone marks shift or spread between conditions
- Example methods: csaw, histoneHMM



Different flavours

universe of methods



A comprehensive comparison of tools for differential ChIP-seq analysis

Sebastian Steinhauser, Nils Kurzawa, Roland Eils and Carl Herrmann

Corresponding author: Carl Herrmann, IPMB Universitðt Heidelberg and Department of Theoretical Bioinformatics, DKFZ, Im Neuenheimer Feld 364, D-69120 Heidelberg, Tel.: (+49) 6221 423612; E-mail: carl.herrmann@uni-heidelberg.de

Briefings in Bioinformatics, 17(6), 2016, 953–966

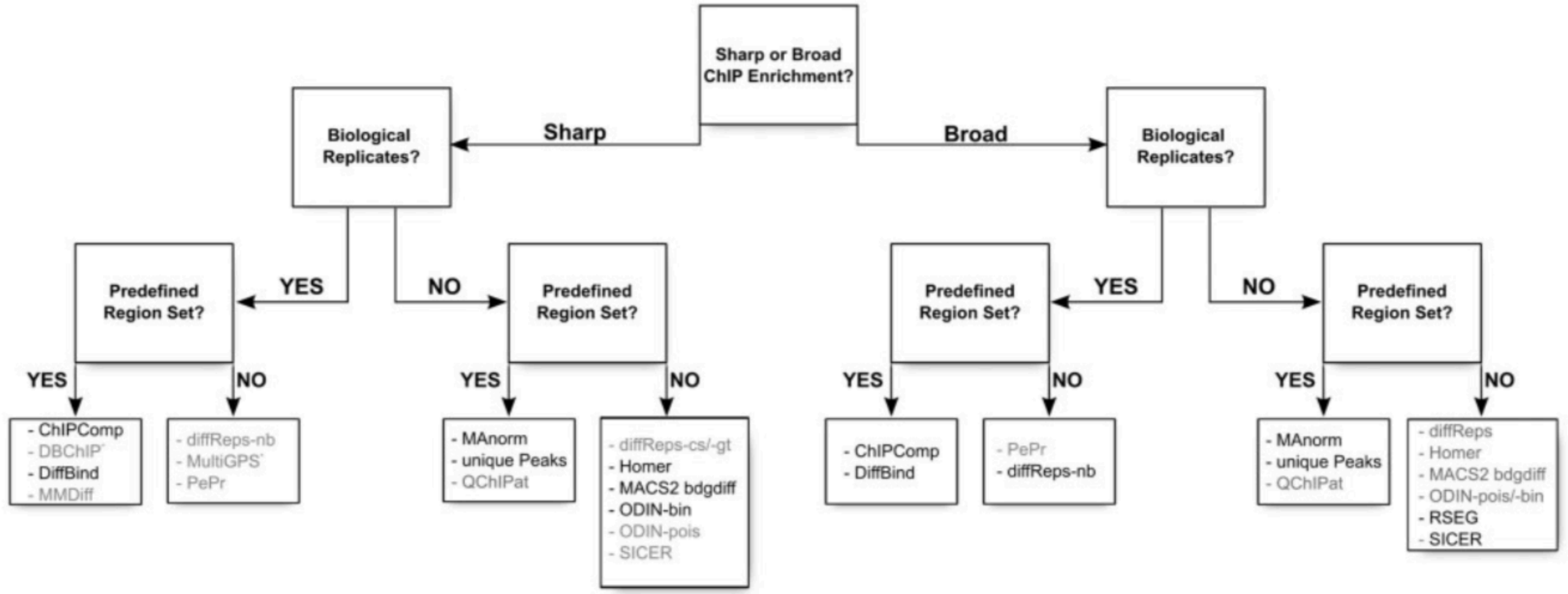


Figure 7. Decision tree indicating the proper choice of tool depending on the data set: shape of the signal (sharp peaks or broad enrichments), presence of replicates and presence of an external set of regions of interest. We have indicated in dark the name of the tools that give good results using default settings, and in gray the tools that would require parameter tuning to achieve optimal results: some tools suffer from an excessive number of DR (PePr, ODIN-pois), an insufficient number of DR (QChIPat, MMDiff, DBChIP) or from an imprecise definition of the DR for sharp signal (SICER, diffReps-nb). *MultiGPS has been explicitly developed for transcription factor ChIP-seq.

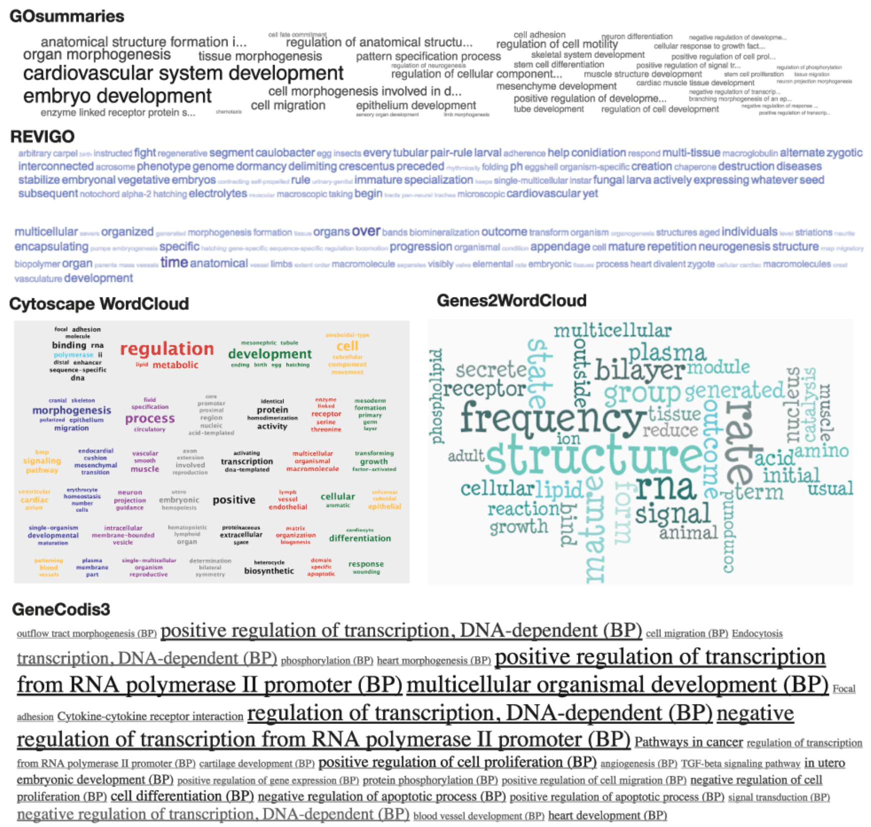
Functional annotations

“Functional annotations is defined as the process of collecting information about and describing a gene’s biological identity: its various aliases, molecular function, biological role(s), sub-cellular location etc.”

genes

peaks

genomic regions



biological knowledge

Functional annotations Over-representation analysis

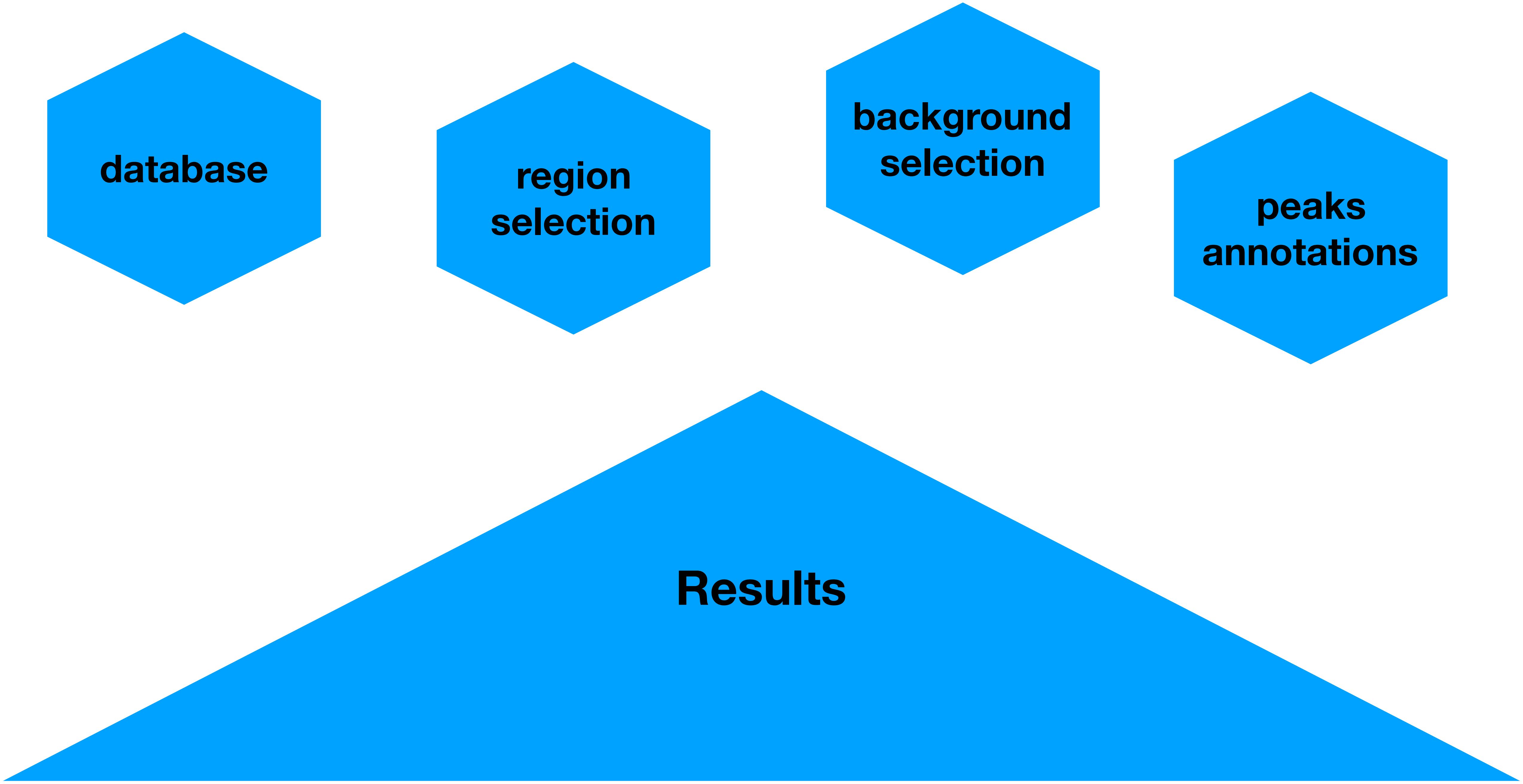
$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

- Widely used approach to identify biological themes is based on hypergeometric model to assess whether the number of selected genes is larger than expected
- To determine whether any terms annotate a specified list genes at frequency greater than that would be expected by chance, calculates p-value using the hypergeometric distribution
- N, total number of genes in the **background** distribution
- M, number of genes within that distribution that are annotated to the node of interest
- n, size of the list of genes of interest
- k, number of genes within that list are annotated to the node

Functional annotations
Gene Set Enrichment Analysis
GSEA

- Over-representation analysis will not detect a situation where the difference is small but demonstrated in a coordinated way in a set of related genes
- GSEA aims to address this limitation, all genes can be used
- GSEA aggregates the per gene statistics across genes within a gene set
- Genes are ranked based on the statistics
- Given a priori defined set of genes *S* (e.g. genes sharing the same GO category), the goal of GSEA is to determine whether the member of *S* are randomly distributed throughout the ranked gene list (*L)* or primarily found at the top or bottom

Functional annotations
it all depends on



database

region selection

background selection

peaks annotations

Results

Functional annotations
in the practicals

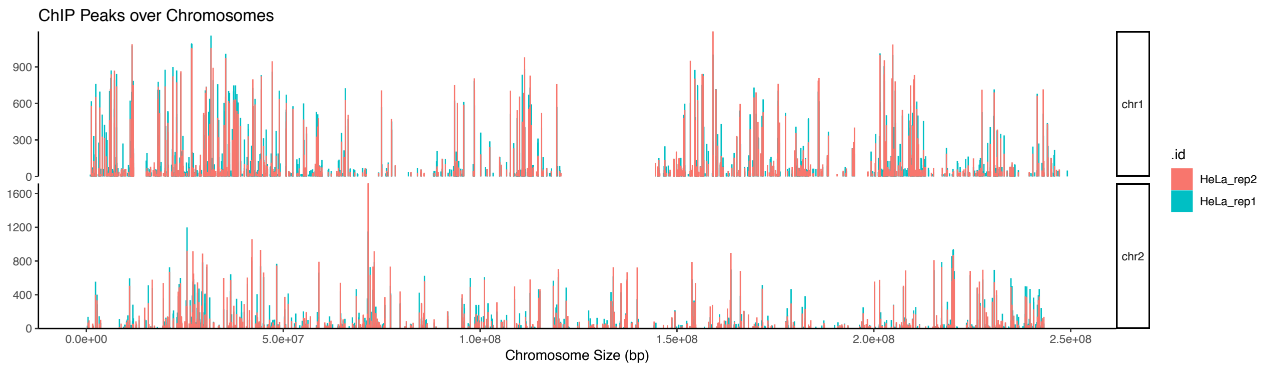
Downstream analysis

ChIPpeakAnno

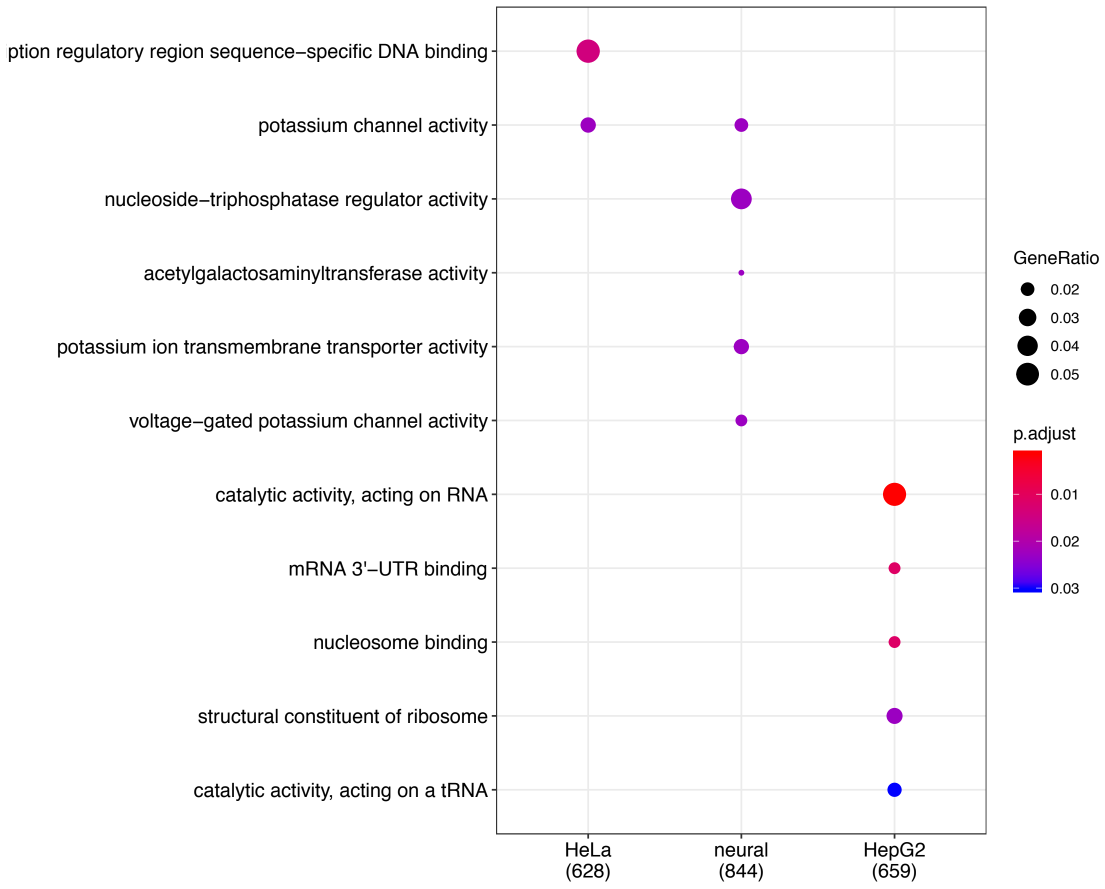
- **annotatePeakInBatch()** to annotate peaks to nearest TSS using TSS.human.GRCh37 precompiled BiomaRt data
- assigning chromosome regions with **assignChromosomeRegion()** function: peaks distributions over genomic features
- over-representation of GO terms with **getEnrichedGO()** function
- over-representation of REACTOME pathways with **getEnrichedPATH()** function

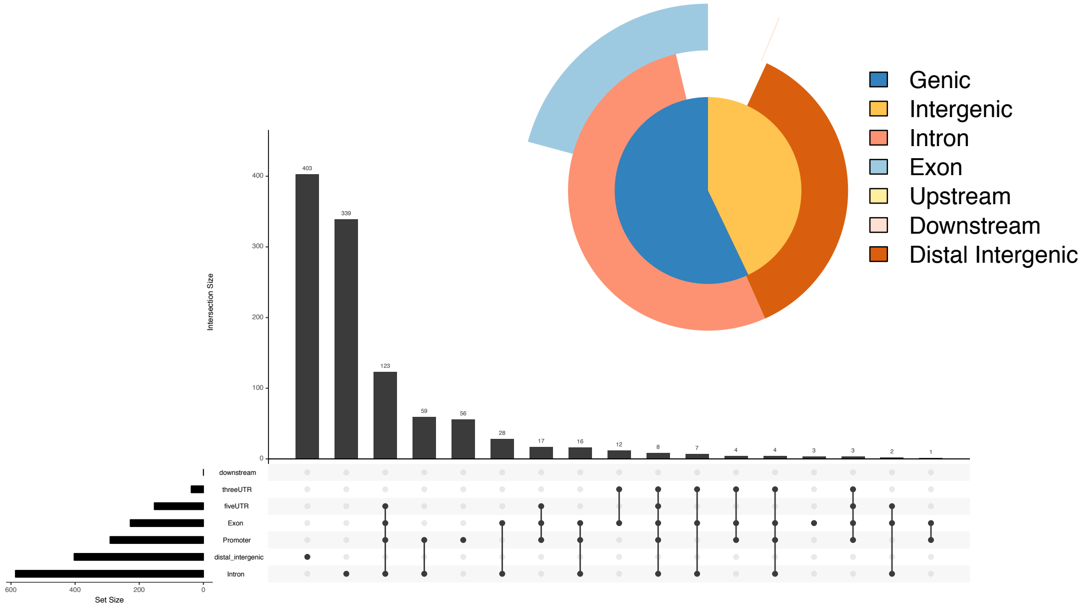
Functional annotations in the practicals

ChIPseeker



Coverage plots





Peaks annotations and visualisations

comparing & reducing GO terms

seq2gene: many-to-many mapping

defining background universe