Aix*Marseille université
Initiative d'excellence

Décryptage de dynamiques épigénomiques au cours de la thymopoïèse et de la spermatogénèse en appliquant une méthodologie de recherche reproductible à des données de séquençage à haut débit

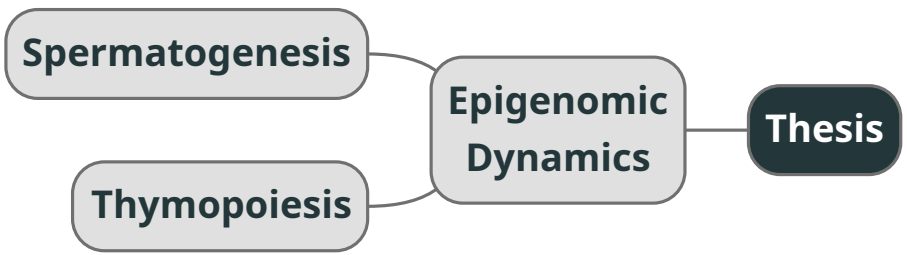
Guillaume Charbonnier

4 Octobre 2019

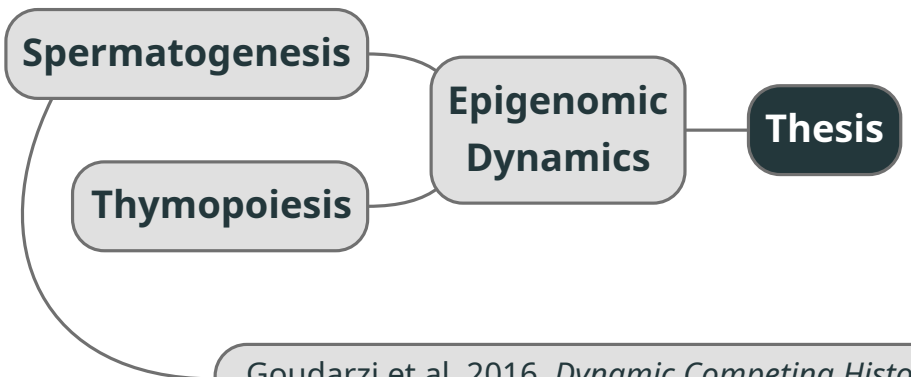
Jury

Dr Carl HERRMANN (GL)	DKFZ (Heidelberg)	Rapporteur
Dr Marco Antonio MENDOZA (CR2)	ISSB (Paris)	Rapporteur
Dr Sophie ROUSSEAUX (DR2)	IAB (Grenoble)	Examinateur
Dr Catherine NGUYEN (DR2)	TAGC (Marseille)	Examinateur
Dr Denis PUTHIER (MCU)	TAGC (Marseille)	Co-directeur de thèse
Dr Salvatore SPICUGLIA (DR2)	TAGC (Marseille)	Directeur de thèse

Thesis summary



Thesis summary

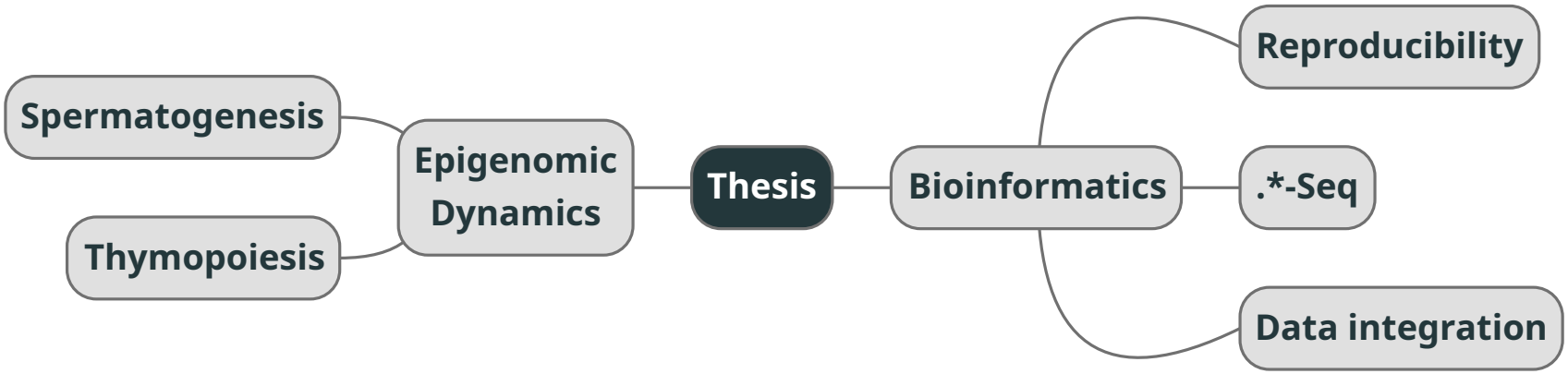


Goudarzi et al. 2016. *Dynamic Competing Histone H4 K5K8 Acetylation and Butyrylation Are Hallmarks of Highly Active Gene Promoters*. Molecular Cell

Barral et al. 2017. *Histone Variant H2A.L.2 Guides Transition Protein-Dependent Protamine Assembly in Male Germ Cells*. Molecular Cell

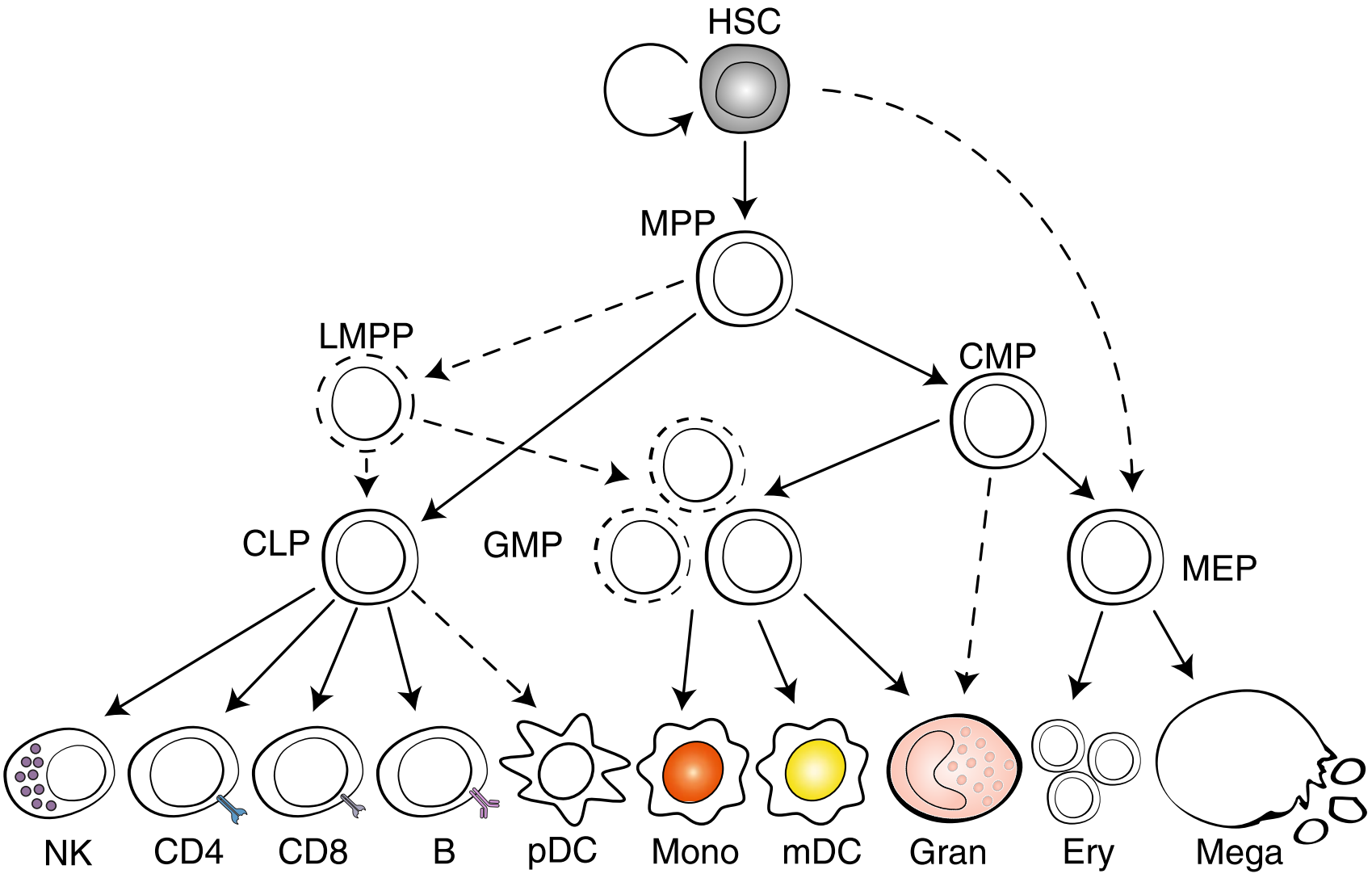
Shiota et al. 2018. *Nut Directs p300-Dependent, Genome-Wide H4 Hyperacetylation in Male Germ Cells*. Cell Reports

Thesis summary



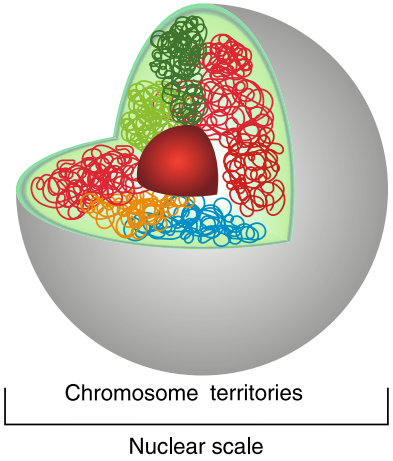
Epigenomic Dynamics

One genome, thousands phenotypes



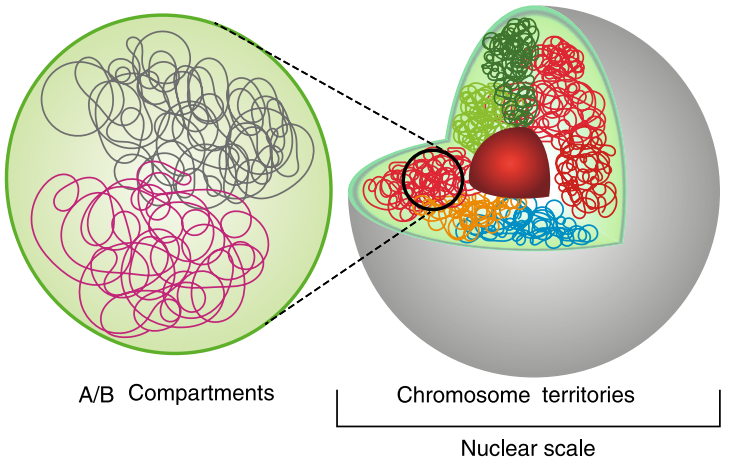
Adapted from *Ulirsch et al., 2019, Nature Genetics*

Genome Complexity: From structural organization...



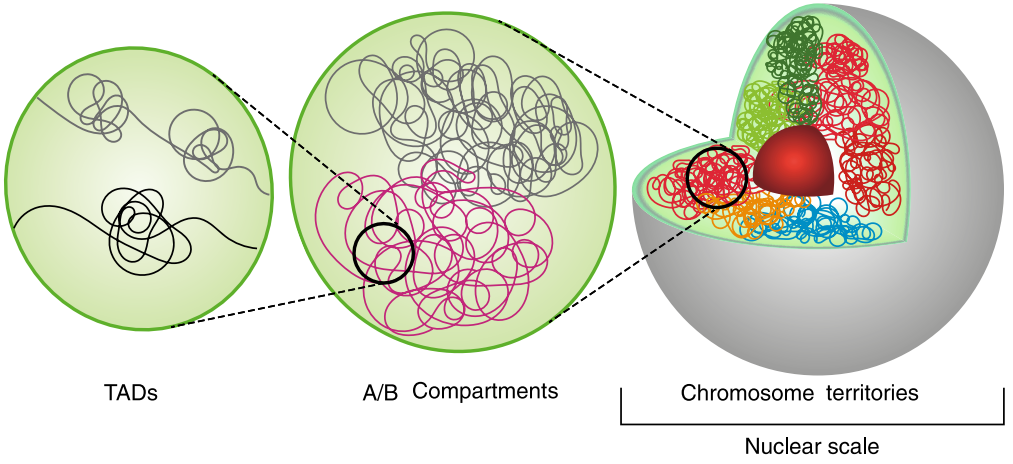
Adapted from *Dogan and Liu, 2018, Nature Plants*

Genome Complexity: From structural organization...



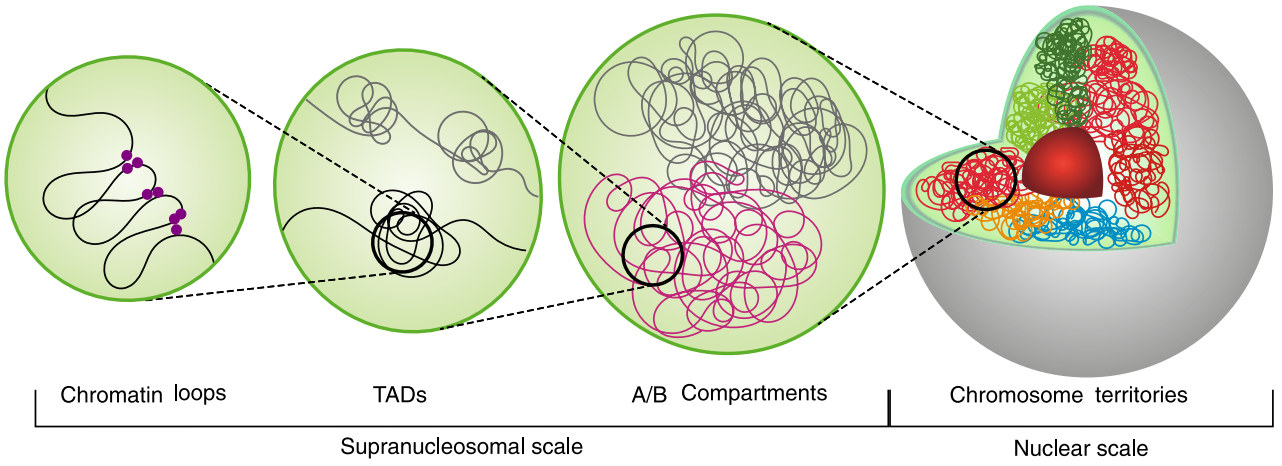
Adapted from *Dogan and Liu, 2018, Nature Plants*

Genome Complexity: From structural organization...



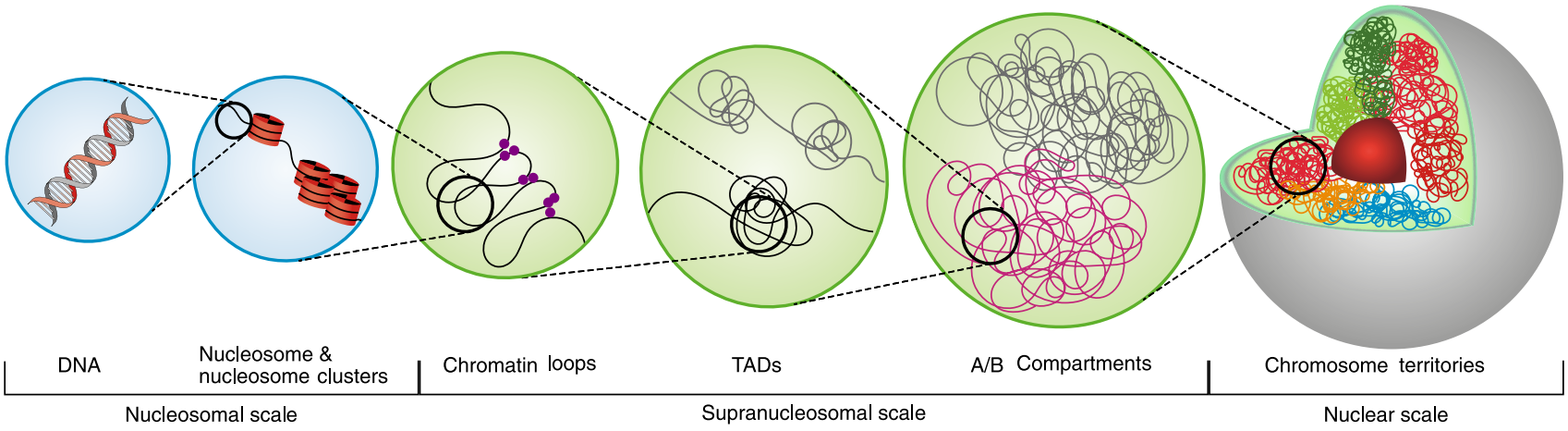
Adapted from *Dogan and Liu, 2018, Nature Plants*

Genome Complexity: From structural organization...



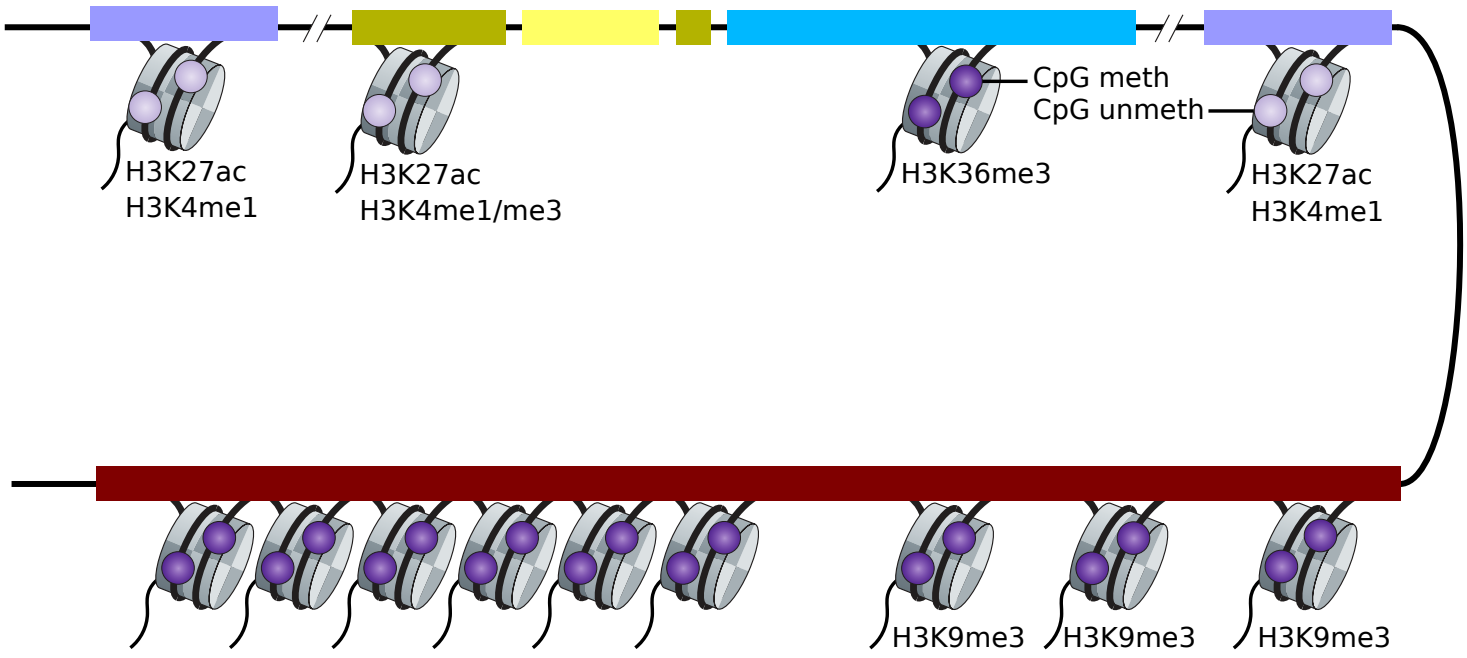
Adapted from *Dogan and Liu, 2018, Nature Plants*

Genome Complexity: From structural organization...

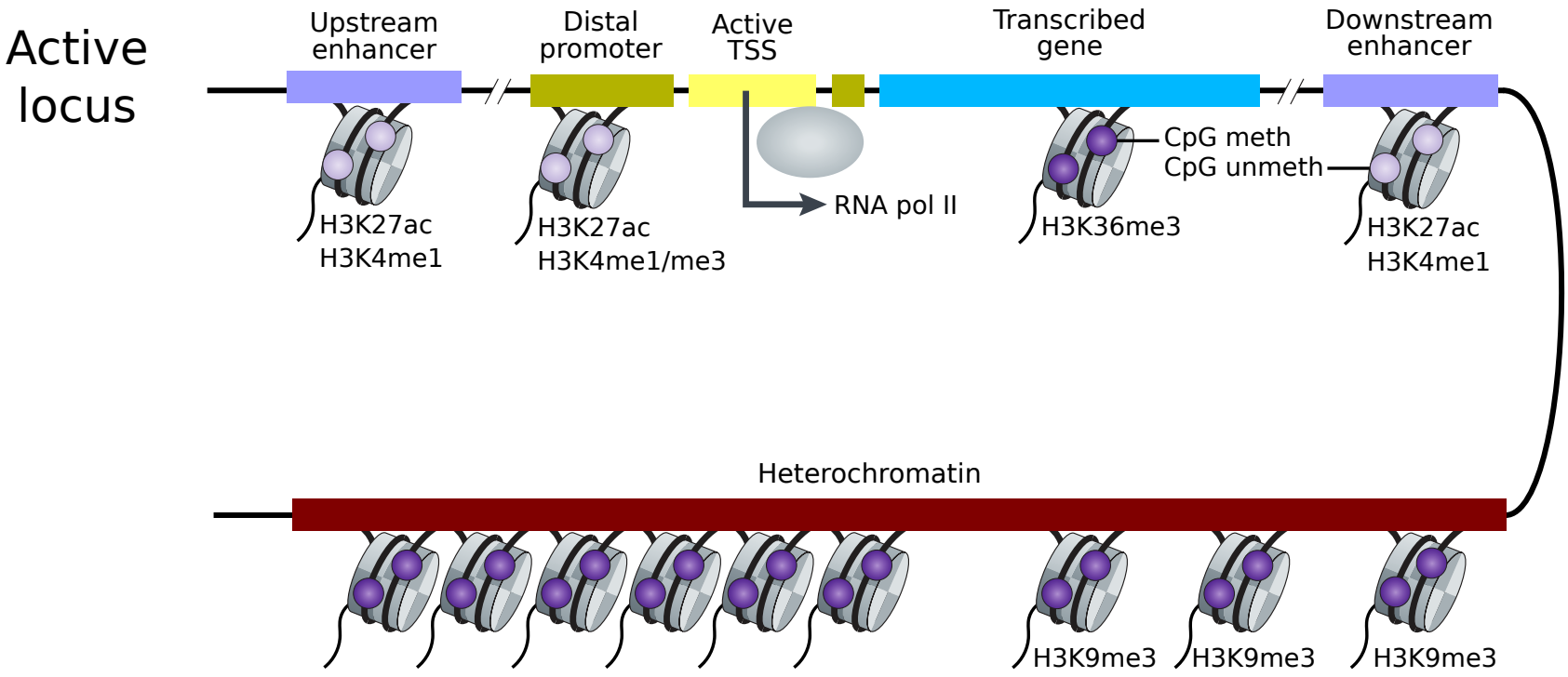


Adapted from *Dogan and Liu, 2018, Nature Plants*

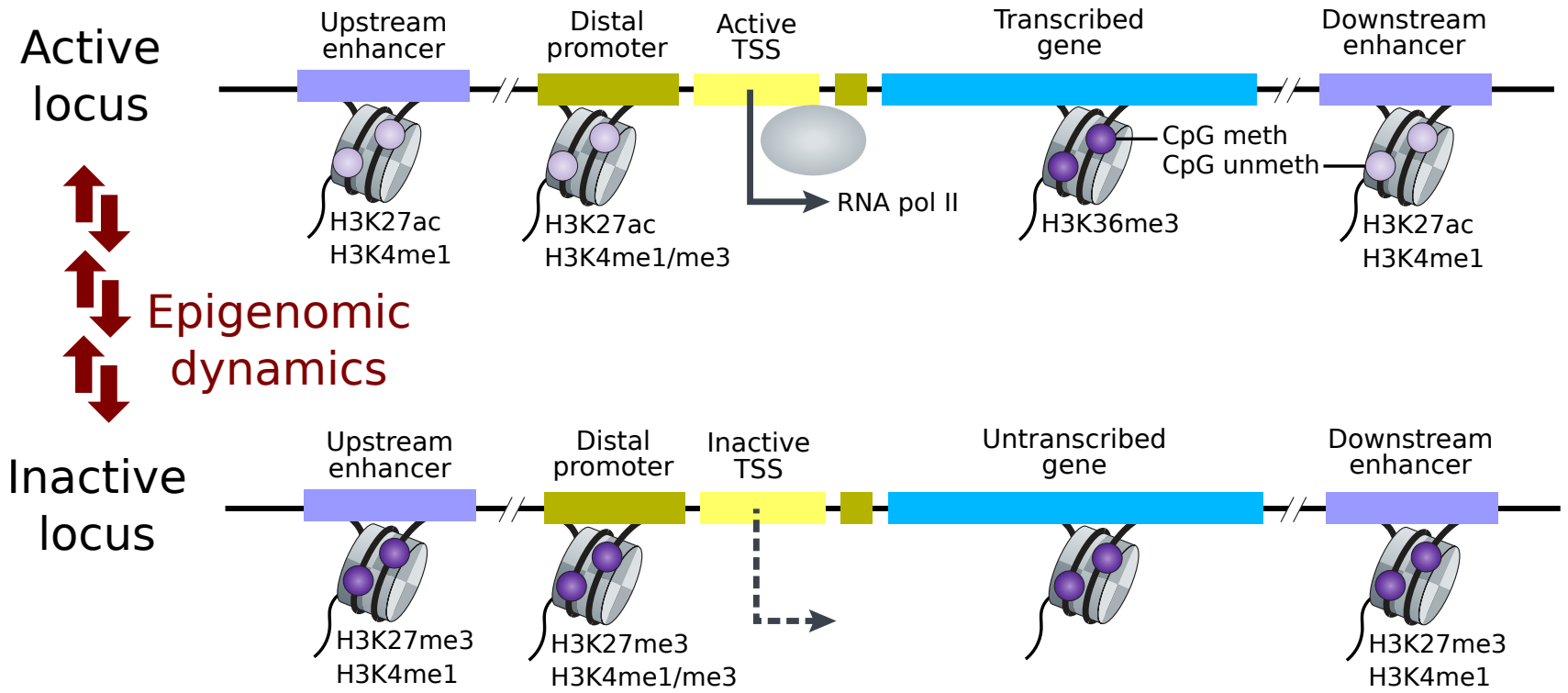
Genome Complexity: From structural organization to epigenetic regulation



Genome Complexity: From structural organization to epigenetic regulation



Genome Complexity: From structural organization to epigenetic regulation

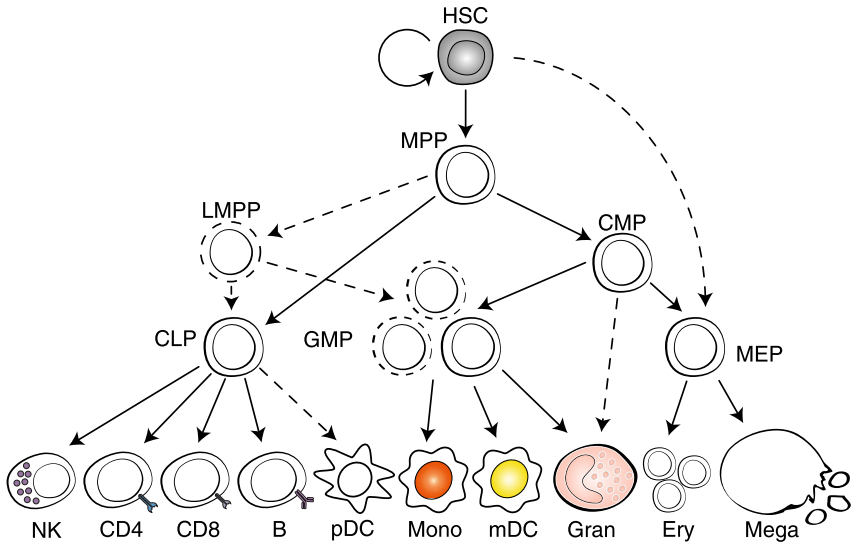


Seeing the big epigenomic picture by mapping blood cells

BLUEPRINT epigenome

Duration	October 2011 - April 2016
Budget	EUR 40 million
Collaboration	41 partners from 9 EU members, Switzerland and Israel
Coordinator	Prof. Dr. Henk Stunnenberg

High-throughput sequencing-based approaches	H3K4me1 H3K4me3 H3K27ac H3K27me3 H3K9me3 H3K36me3 Bisulfite RNA DNAseI / FAIRE / ATAC

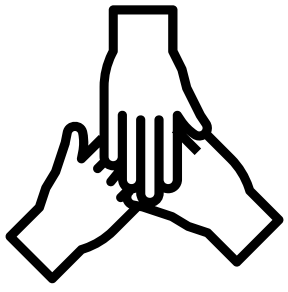


Outcome	112 complete epigenomes 1137 partial epigenomes

Necker and TAGC: Blueprint collaboration and beyond

BLUEPRINT epigenome

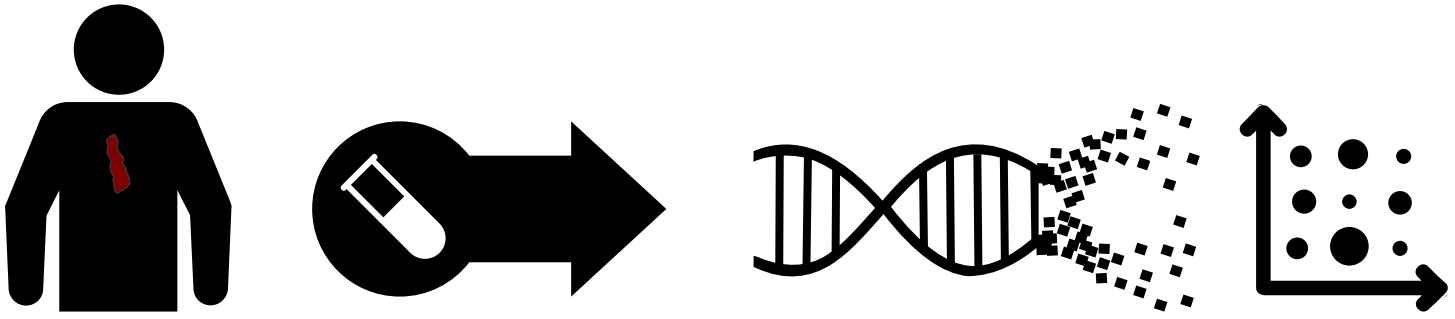
Institut Necker Enfants Malades
Centre de Recherche



umr U 1090
TAGC
theories and approaches of genomic complexity

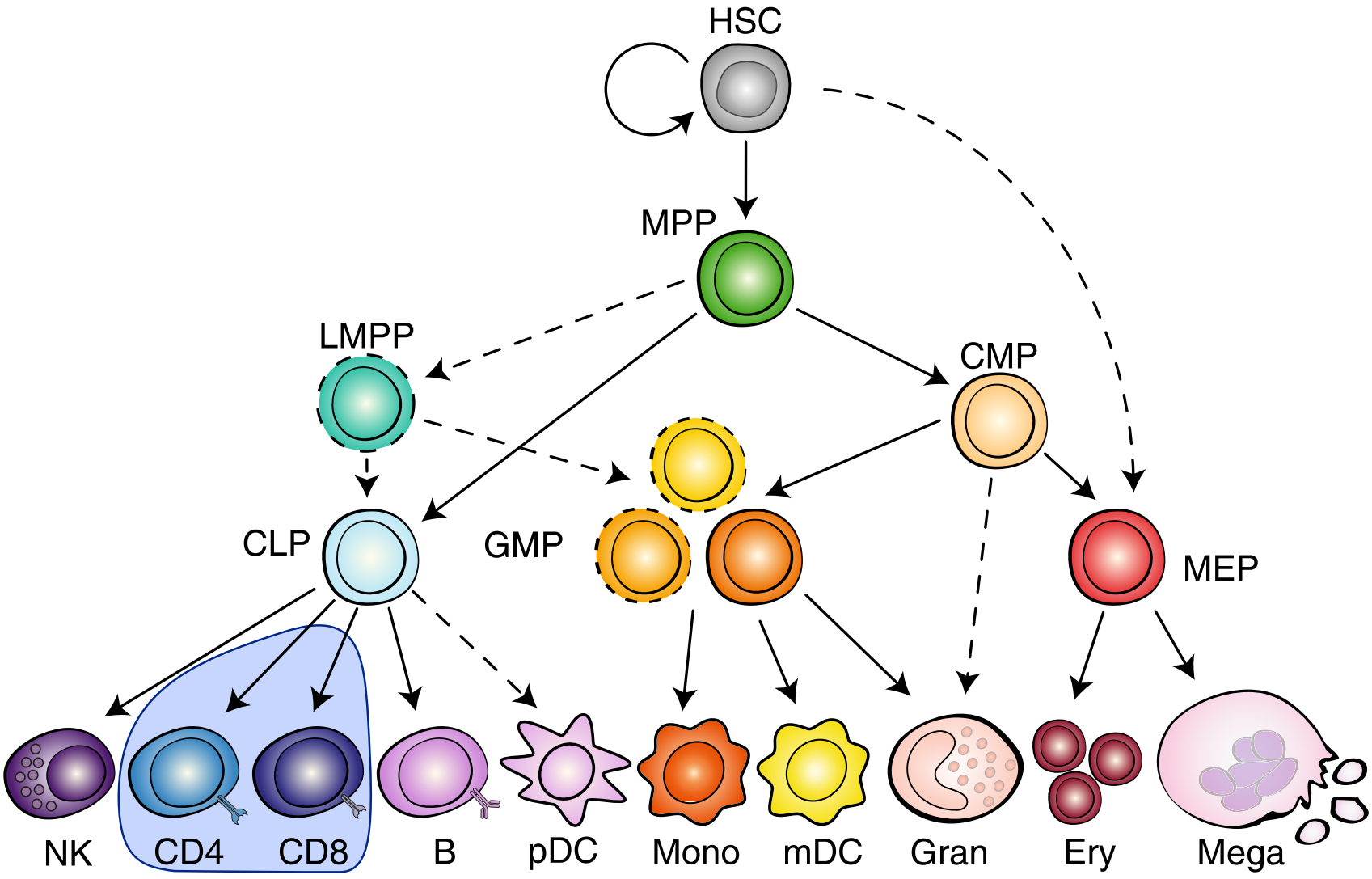
Vahid Asnafi
Agata Cieslak
et al.

Salvatore Spicuglia
Guillaume Charbonnier
Denis Puthier



Thymopoiesis

Thymopoiesis as a subset of hematopoiesis



The thymus, the organ of thymopoiesis

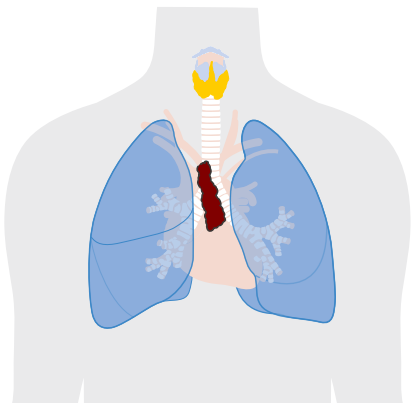
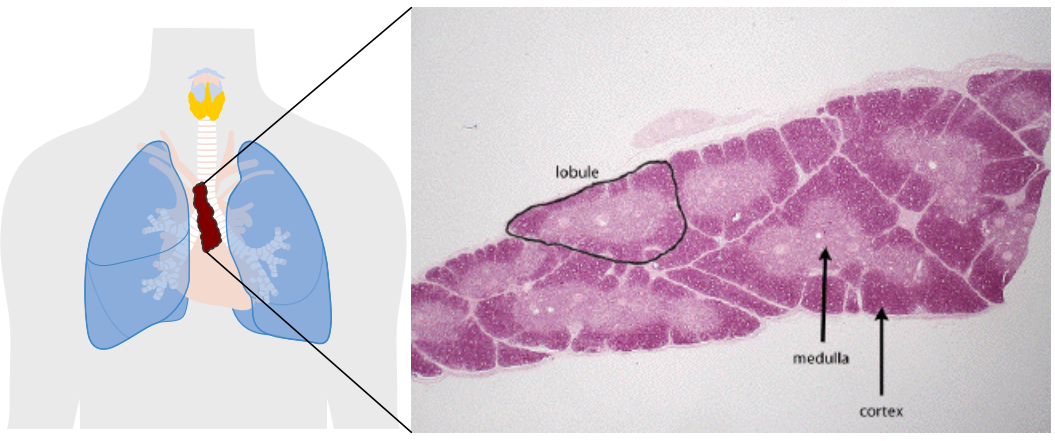


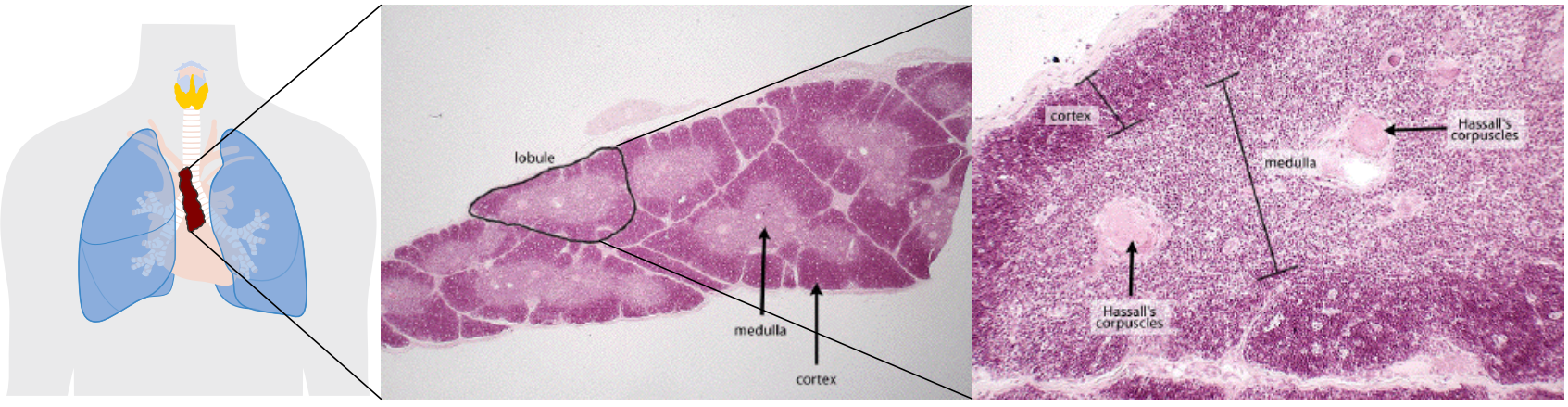
Diagram adapted from *Cancer UK Research / Wikimedia Commons*

The thymus, the organ of thymopoiesis



H&E stained section of human infant thymus taken from *Darthmouth College*

The thymus, the organ of thymopoiesis



H&E stained section of human infant thymus taken from *Darthmouth College*

The thymus, the organ of thymopoiesis

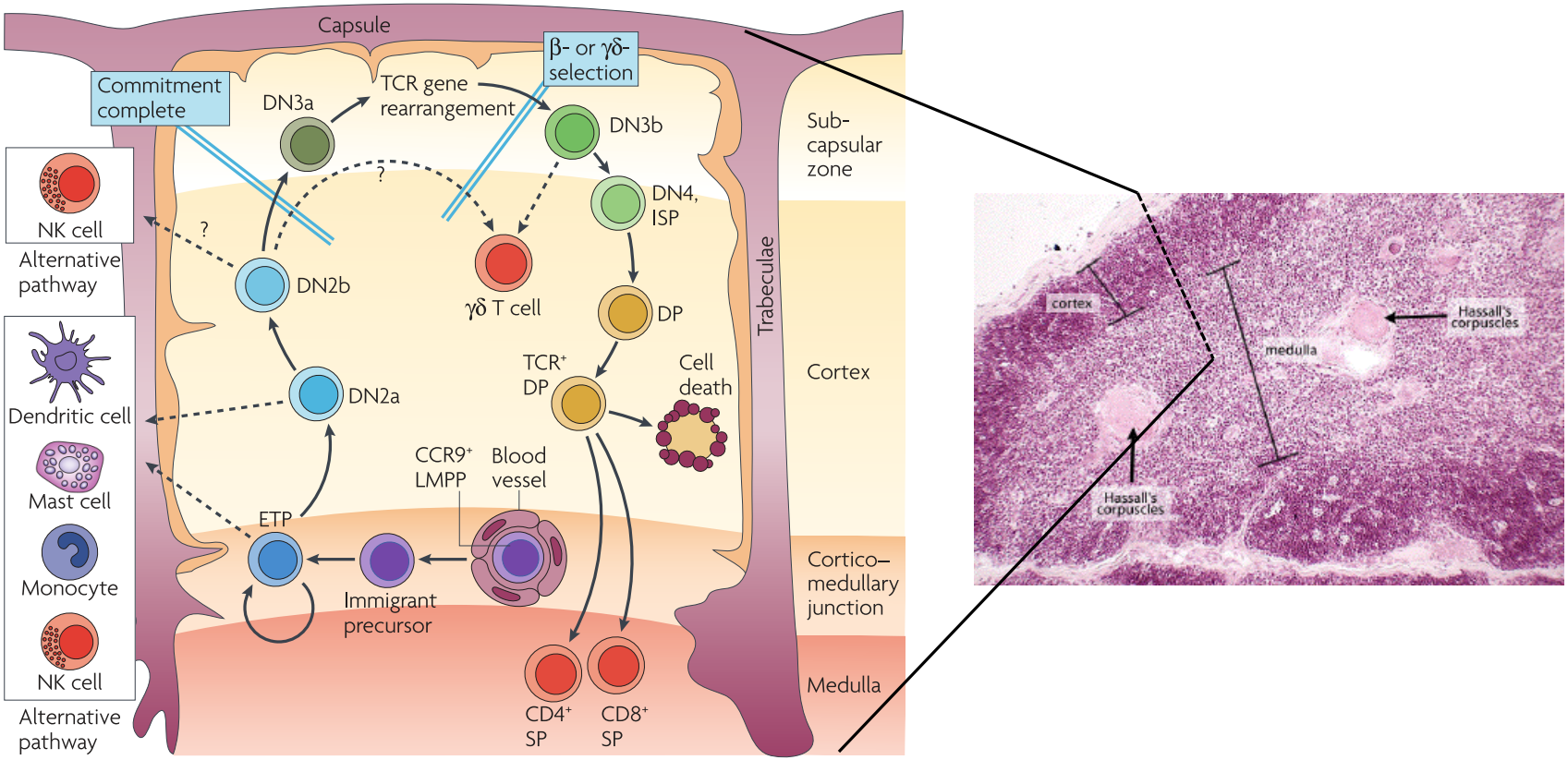
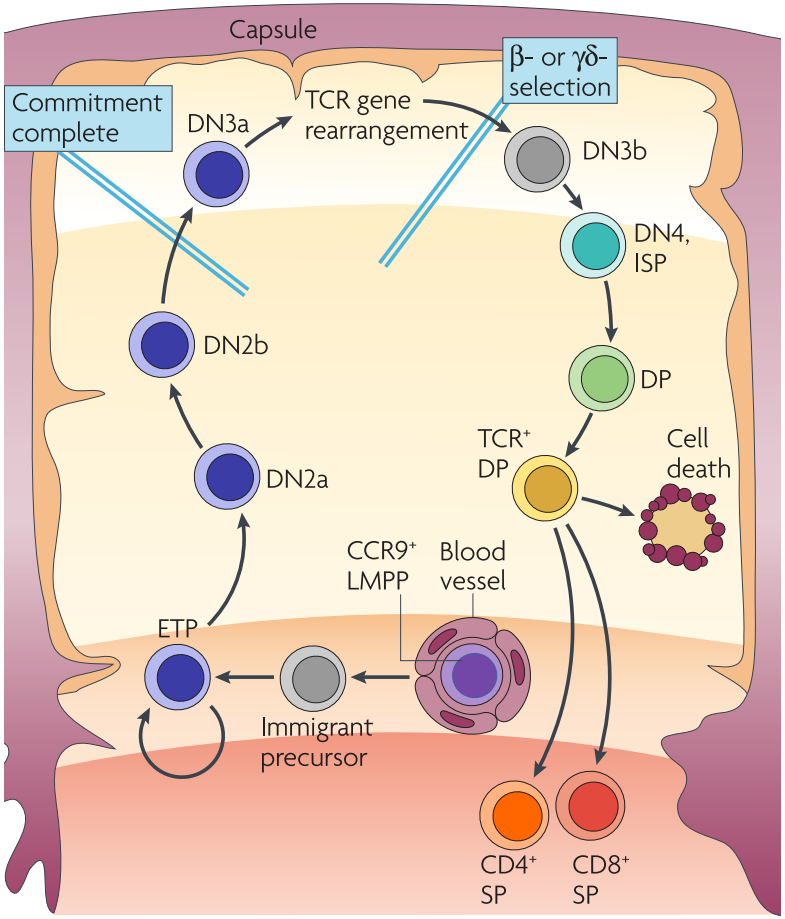
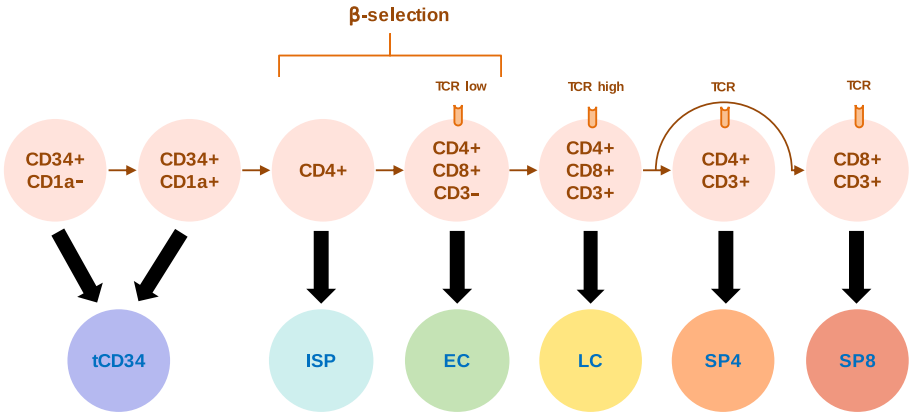


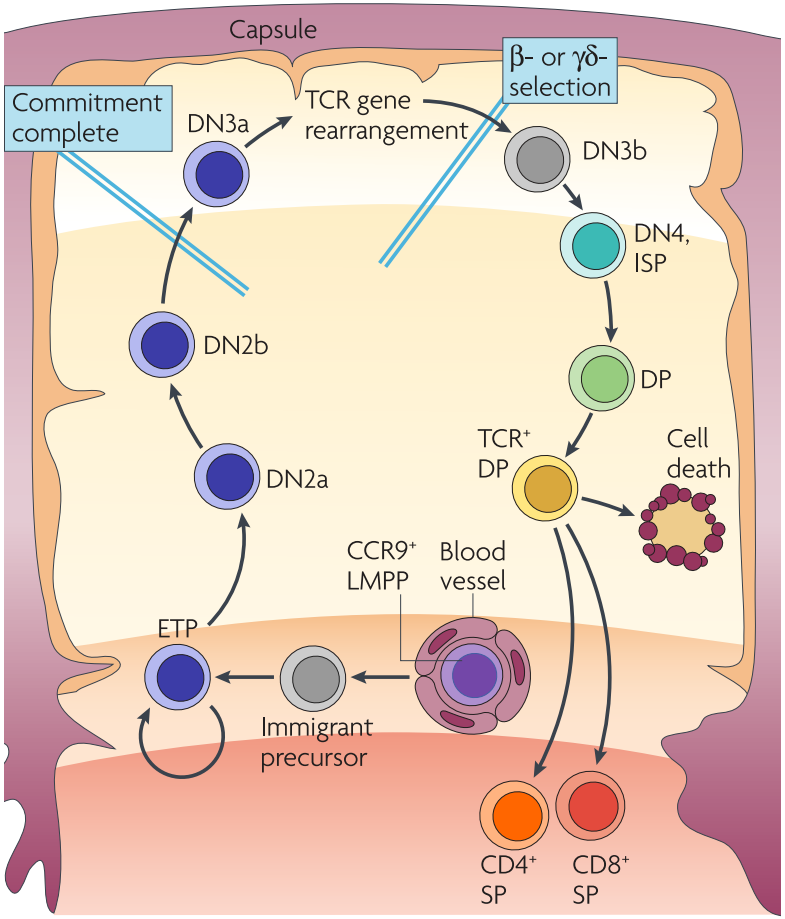
Diagram taken from *Rothenberg et al., 2008, Nature Reviews Immunology*

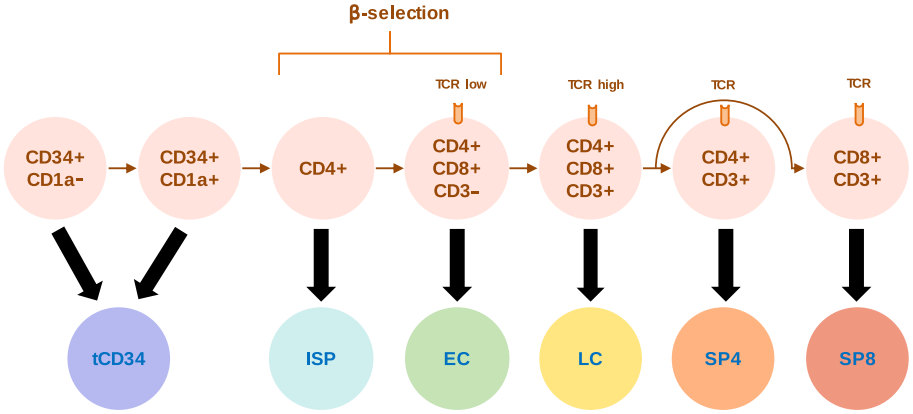
Nomenclature for major T cell differentiation stages





Available experimental approaches

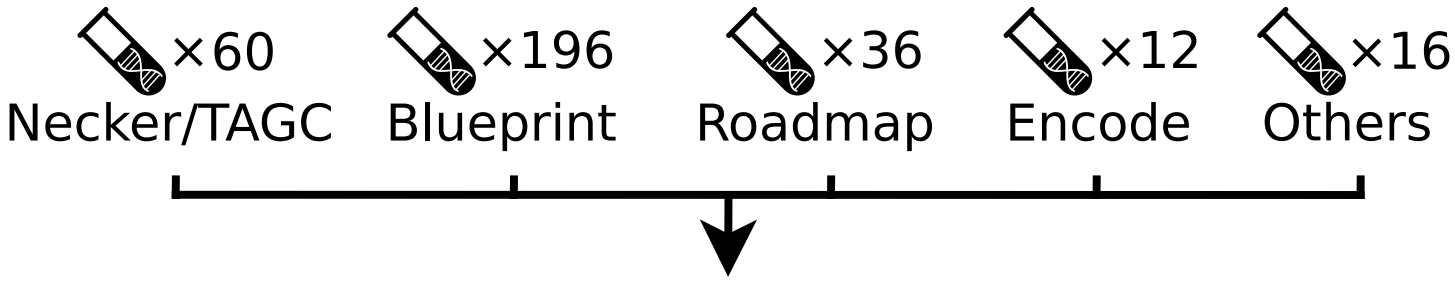




studied thymic subpopulations

Methods	tCD34	ISP	EC	LC	SP4	SP8
H3K4me1	✓	✗	✓	✓	✓	✓
H3K4me3	✓	✗	✓	✓	✓	✓
H3K27ac	✓	✗	✓	✓	✓	✓
H3K27me3	✓	✗	✓	✓	✓	✓
H3K9me3	✓	✗	✓	✓	✓	✓
H3K36me3	✓	✗	✓	✓	✓	✓
BS-Seq	✓	✗	✓	✓	✓	✓
RNA-Seq	✓	✓	✓	✓	✓	✓
ATAC-Seq	✓	✗	✓	✓	✗	✗

Biological objectives



Integrate and visualize data

Validate new thymic epigenomes

Characterize epigenomic dynamics of T cell differentiation

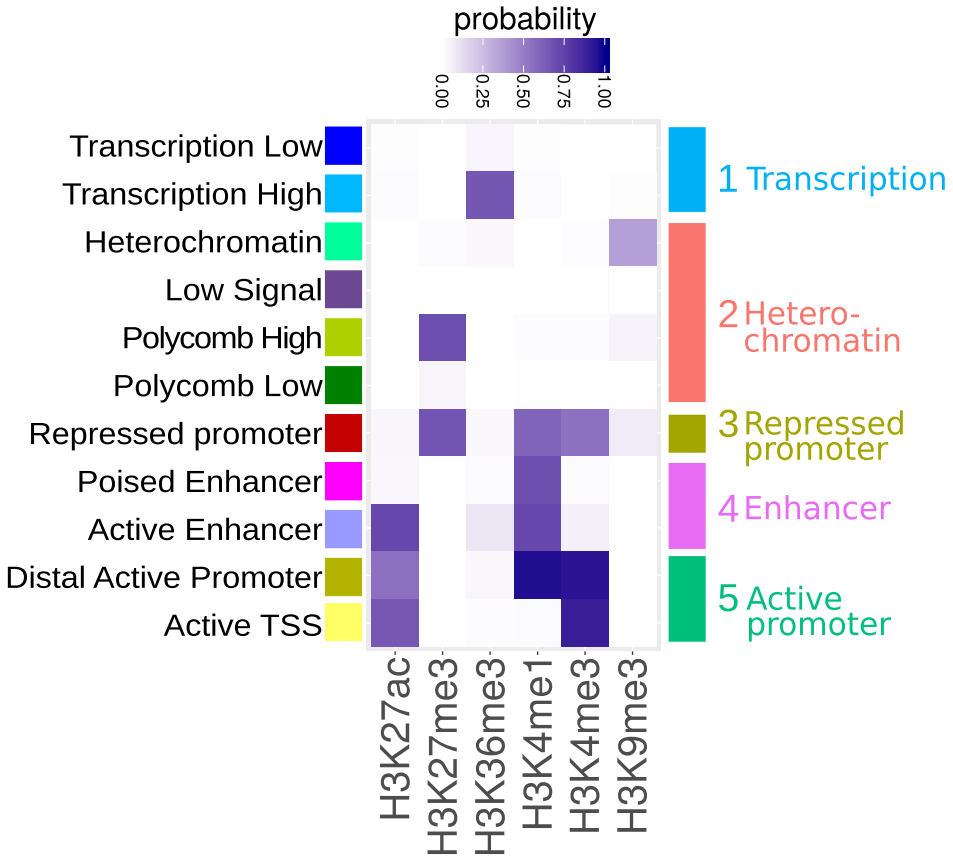
Identify new epigenetic regulation mechanism

Thymopoiesis

A new reference epigenome of human early T cell differentiation

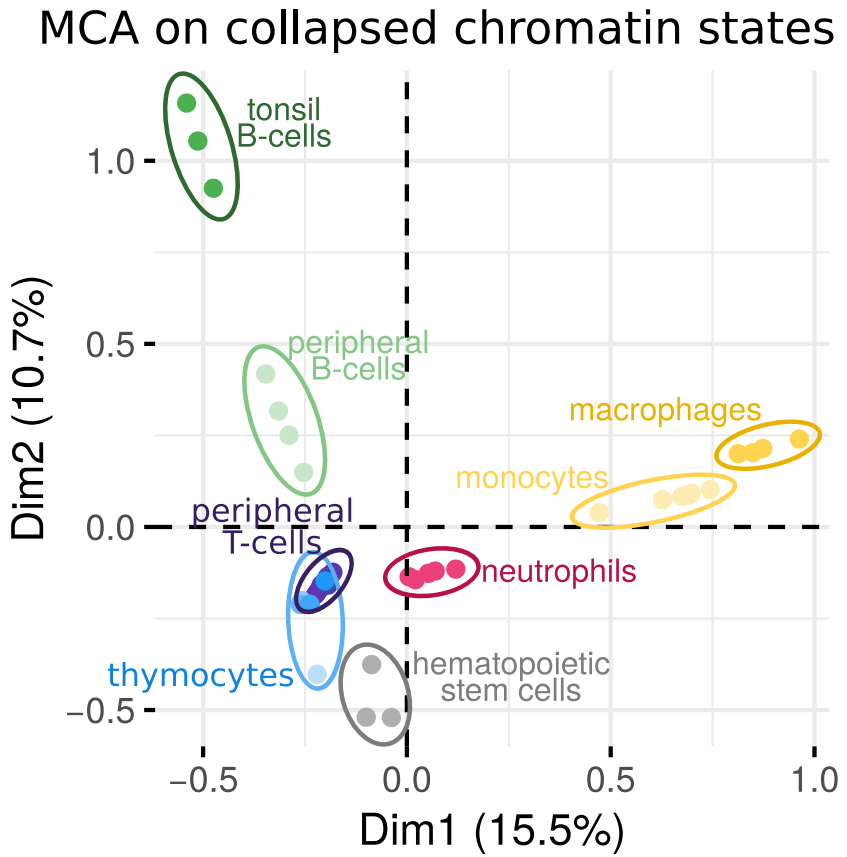
Epigenomic landscape summarized by chromatin segmentation

sample \ region index	1	2	3	...
HSC	1	1	2	...
tCD34	3	5	1	...
EC	3	2	4	...
...	...	...	...	...

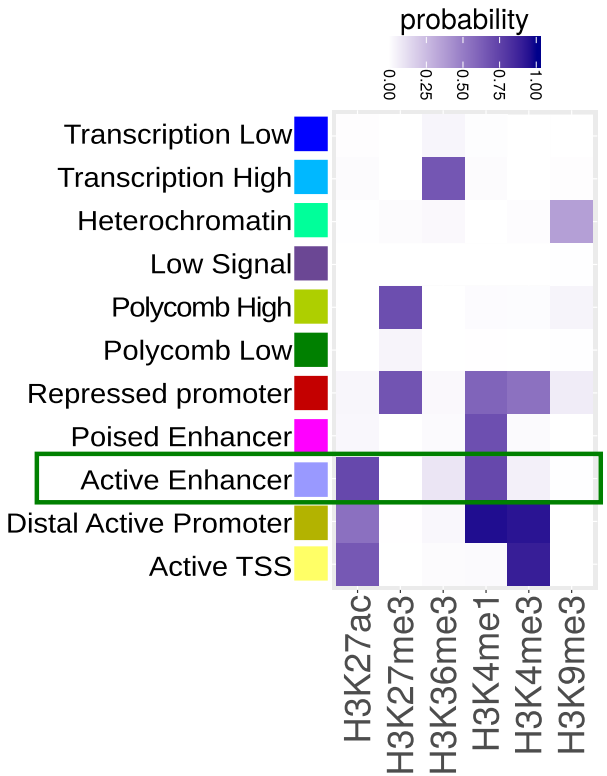


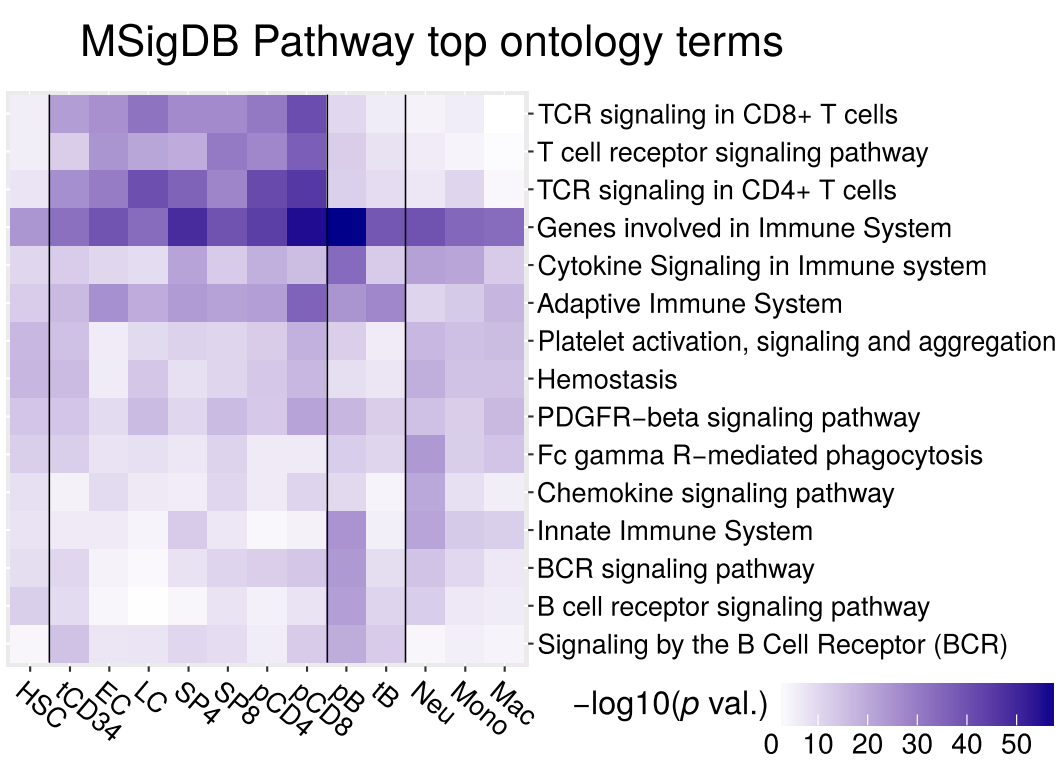
A consistent epigenomic landscape of hematopoiesis

sample \ region index	1	2	3	...
HSC	1	1	2	...
tCD34	3	5	1	...
EC	3	2	4	...
...	...	...	...	...

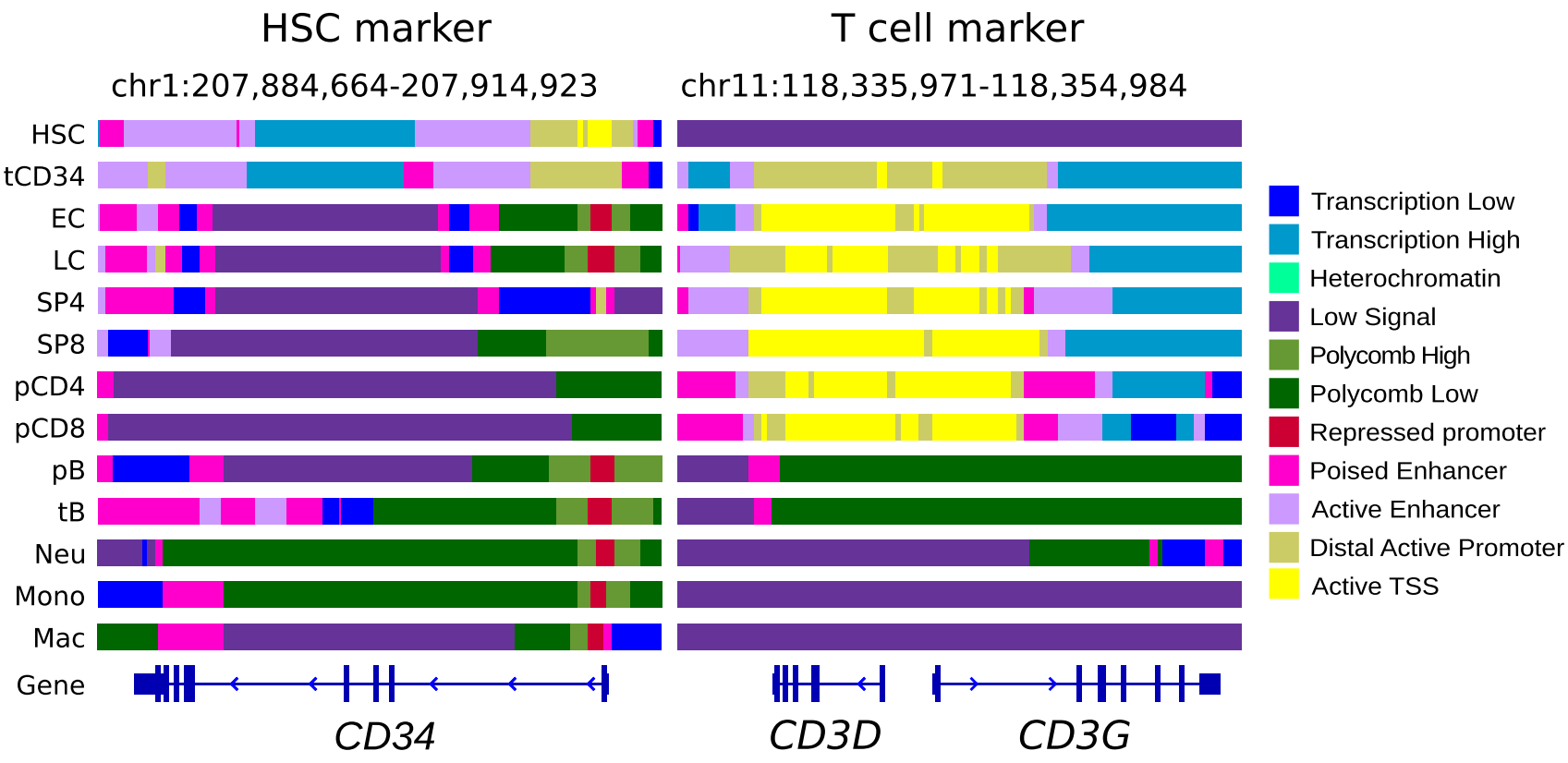


Enrichment analysis of active enhancers highlights lineage-specific signatures

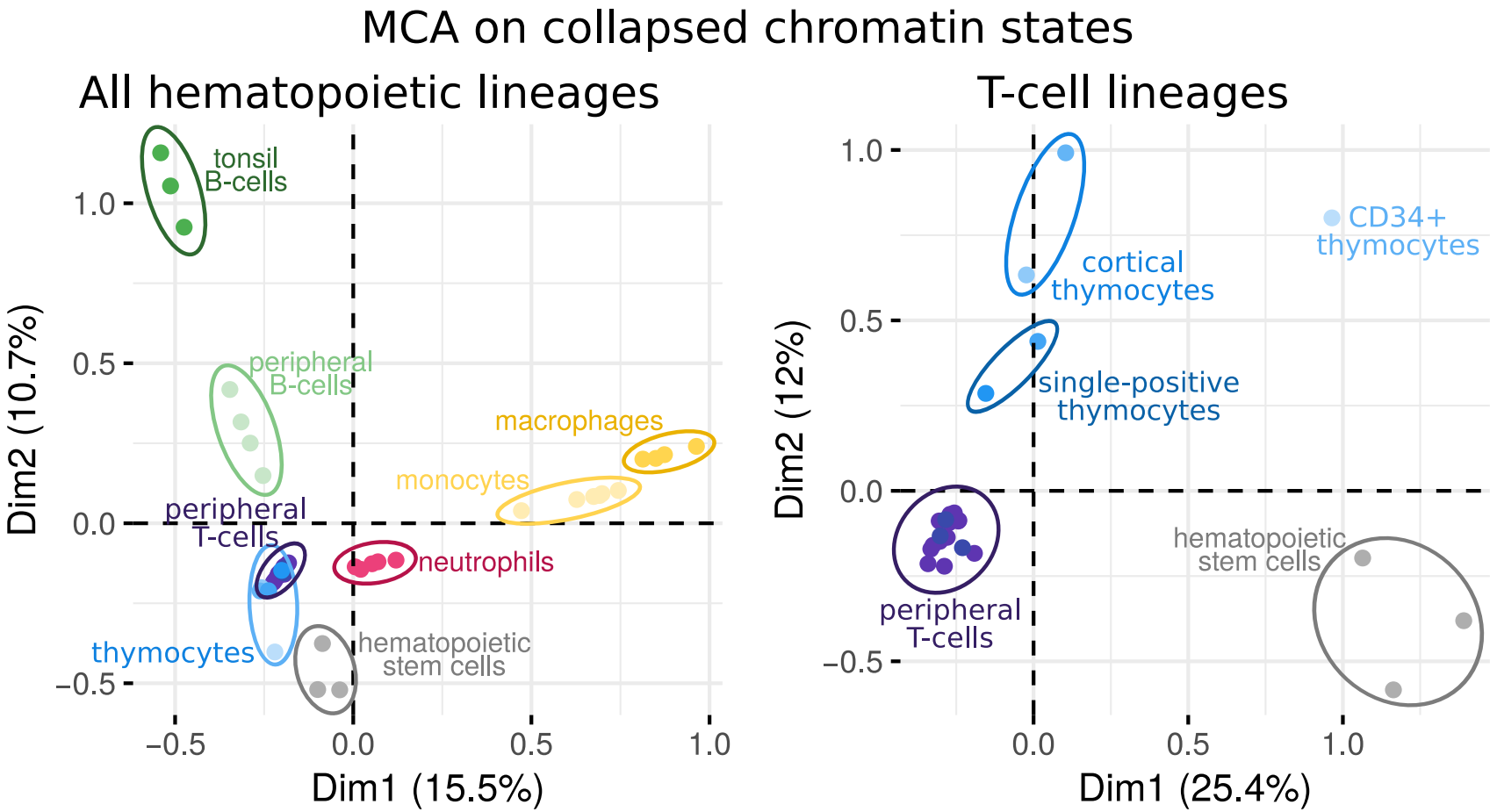




Chromatin segmentation meets expected epigenetics status for known cell-type specific genes

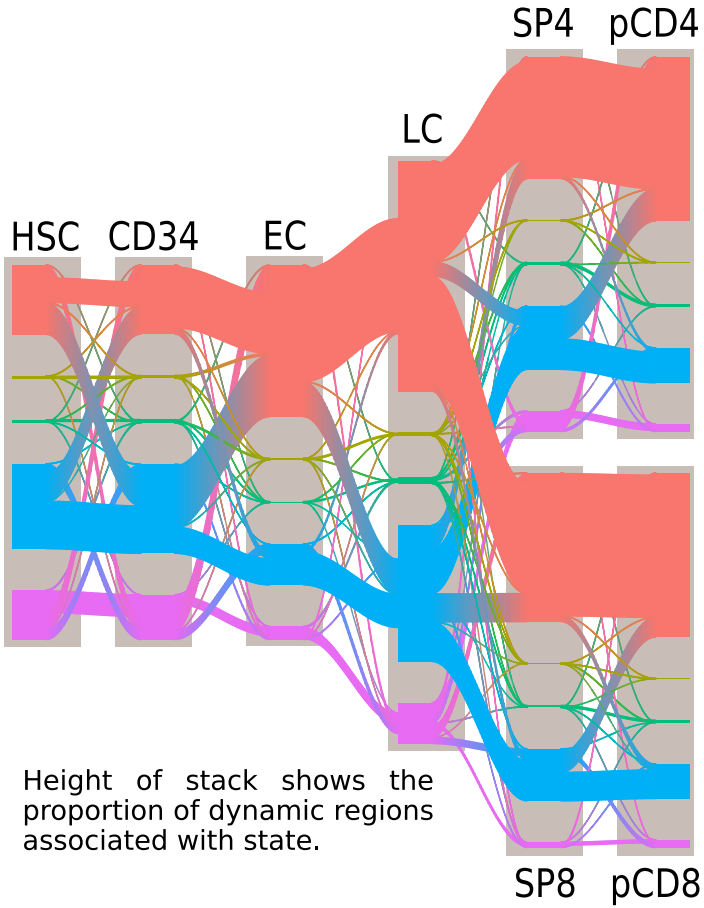


A consistent epigenomic landscape of thymopoiesis

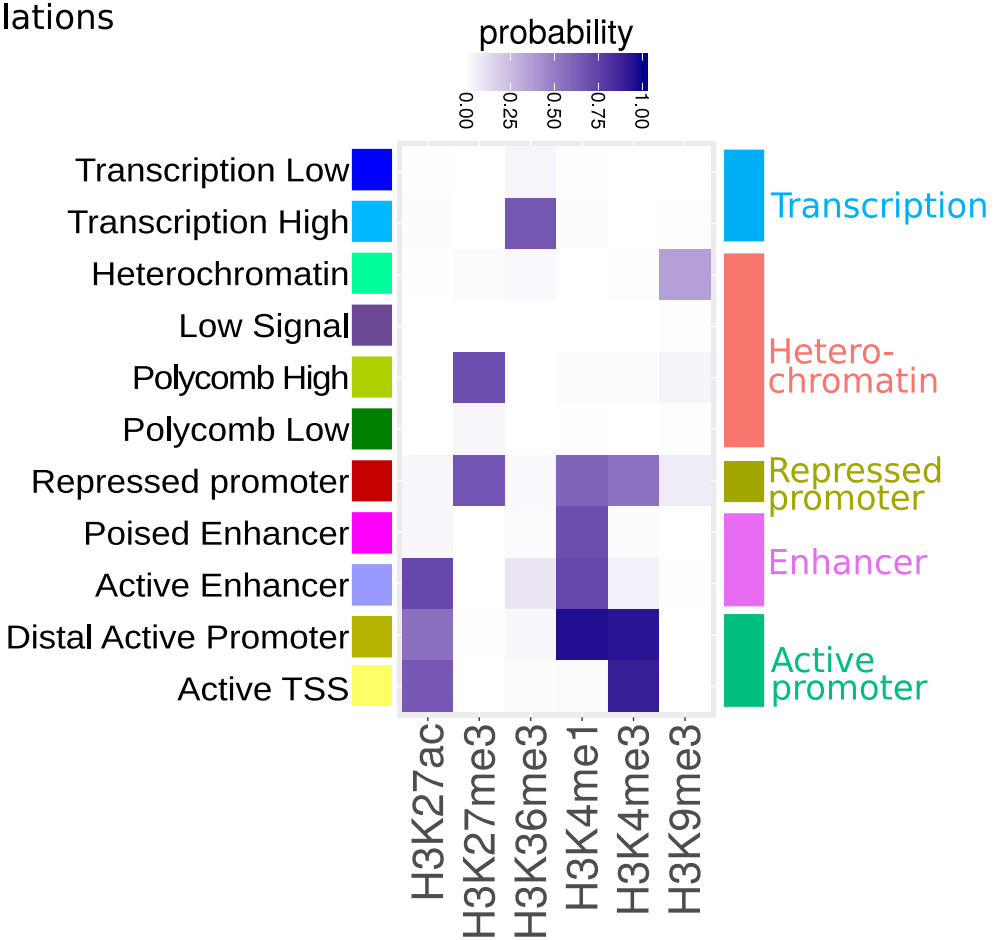


A progressive loss of plasticity during differentiation

Riverplot depicting the main epigenomic transitions between HSC and mature T cell populations

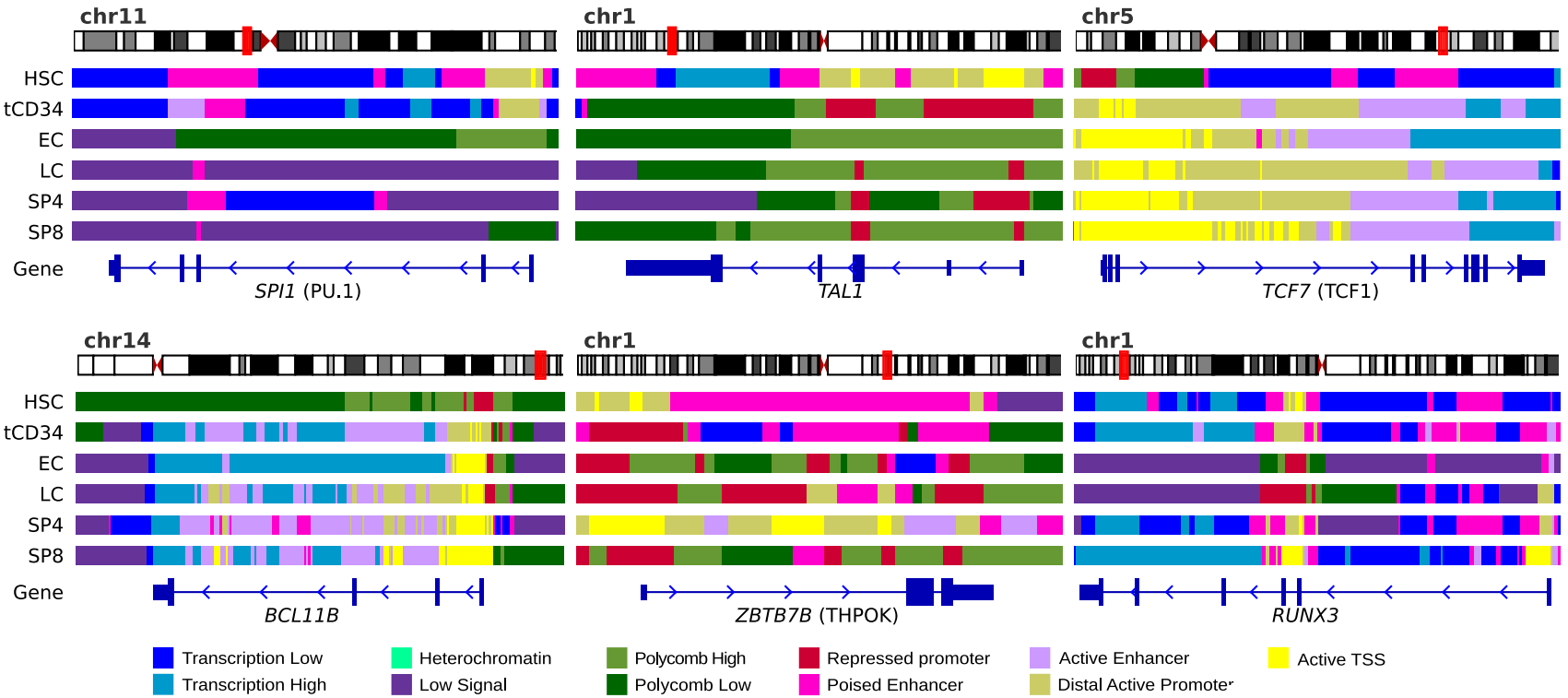


Height of stack shows the proportion of dynamic regions associated with state.



Chromatin segmentation meets expected epigenetics status for known cell-type specific genes

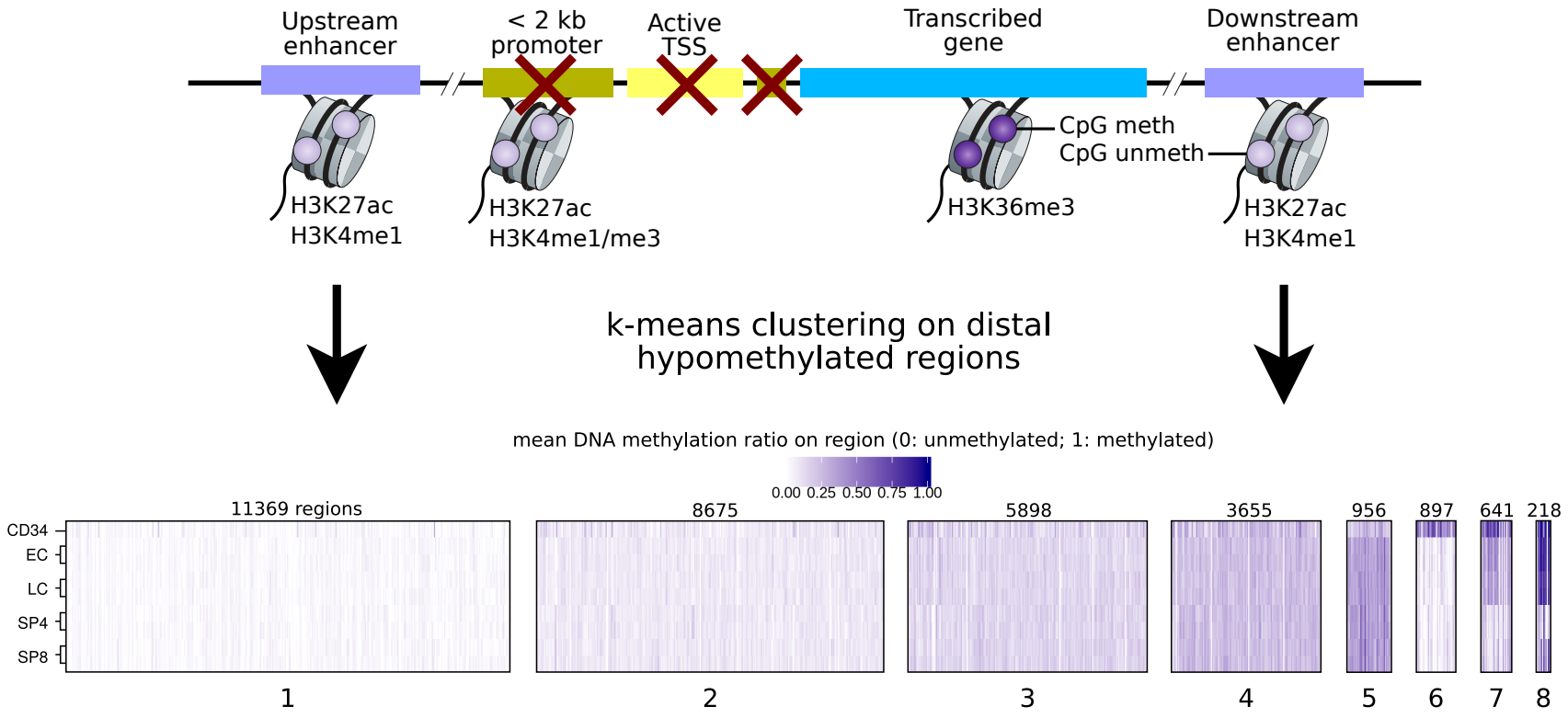
Chromatin state profiles of representative genes specifically expressed at different stages of T cell differentiation



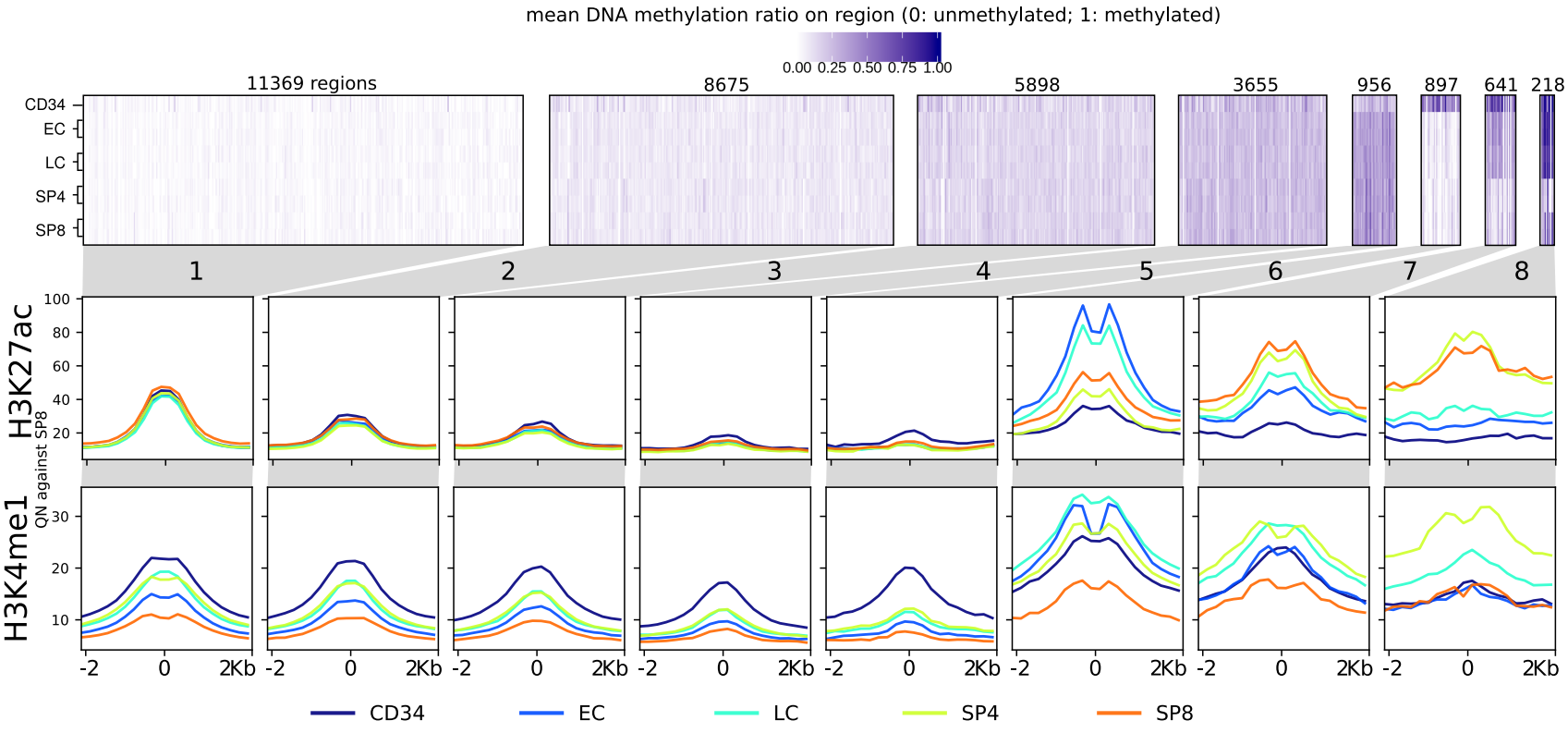
Thymopoiesis

Epigenomic dynamics of distal regulatory regions

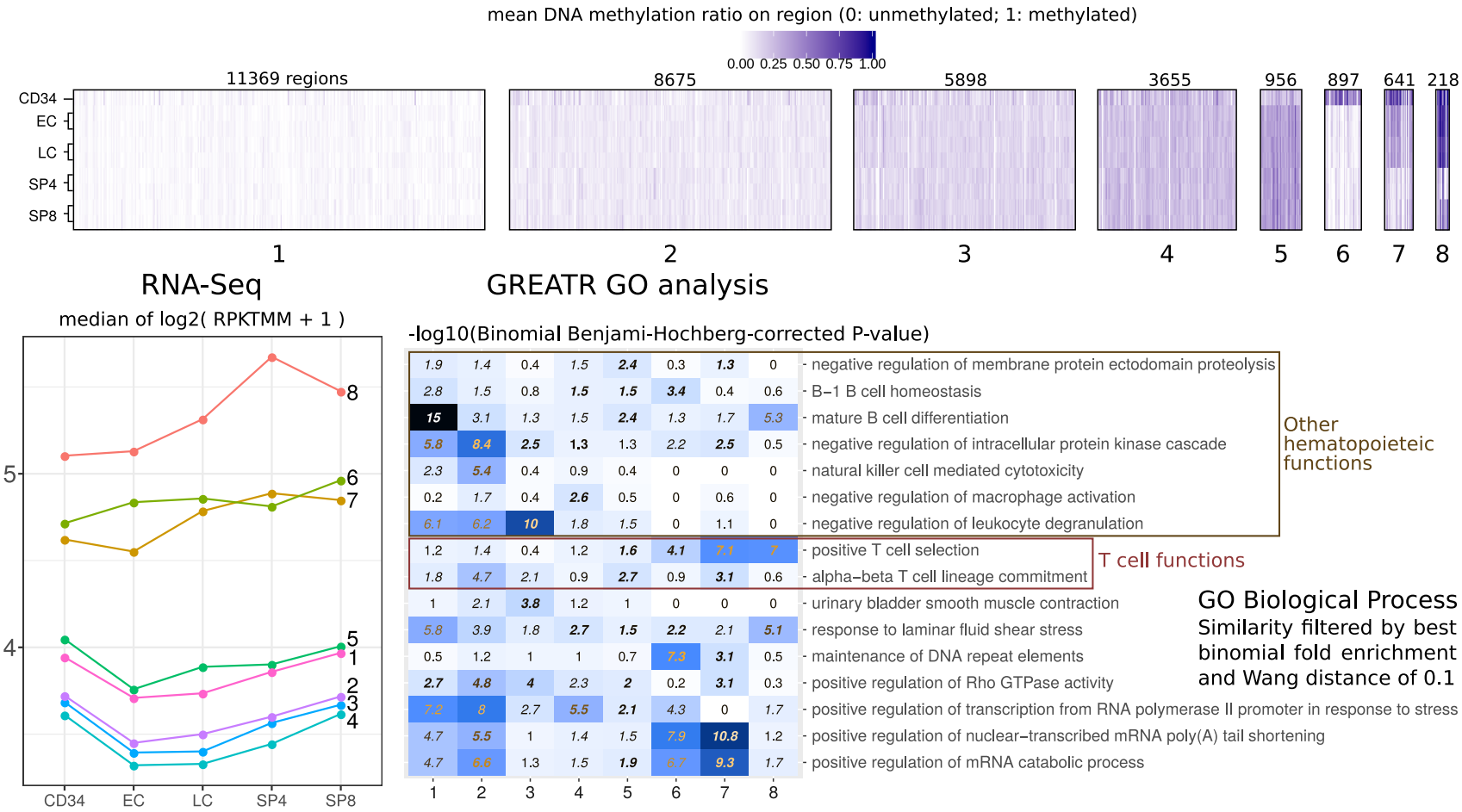
DNA methylation dynamics of distal regions



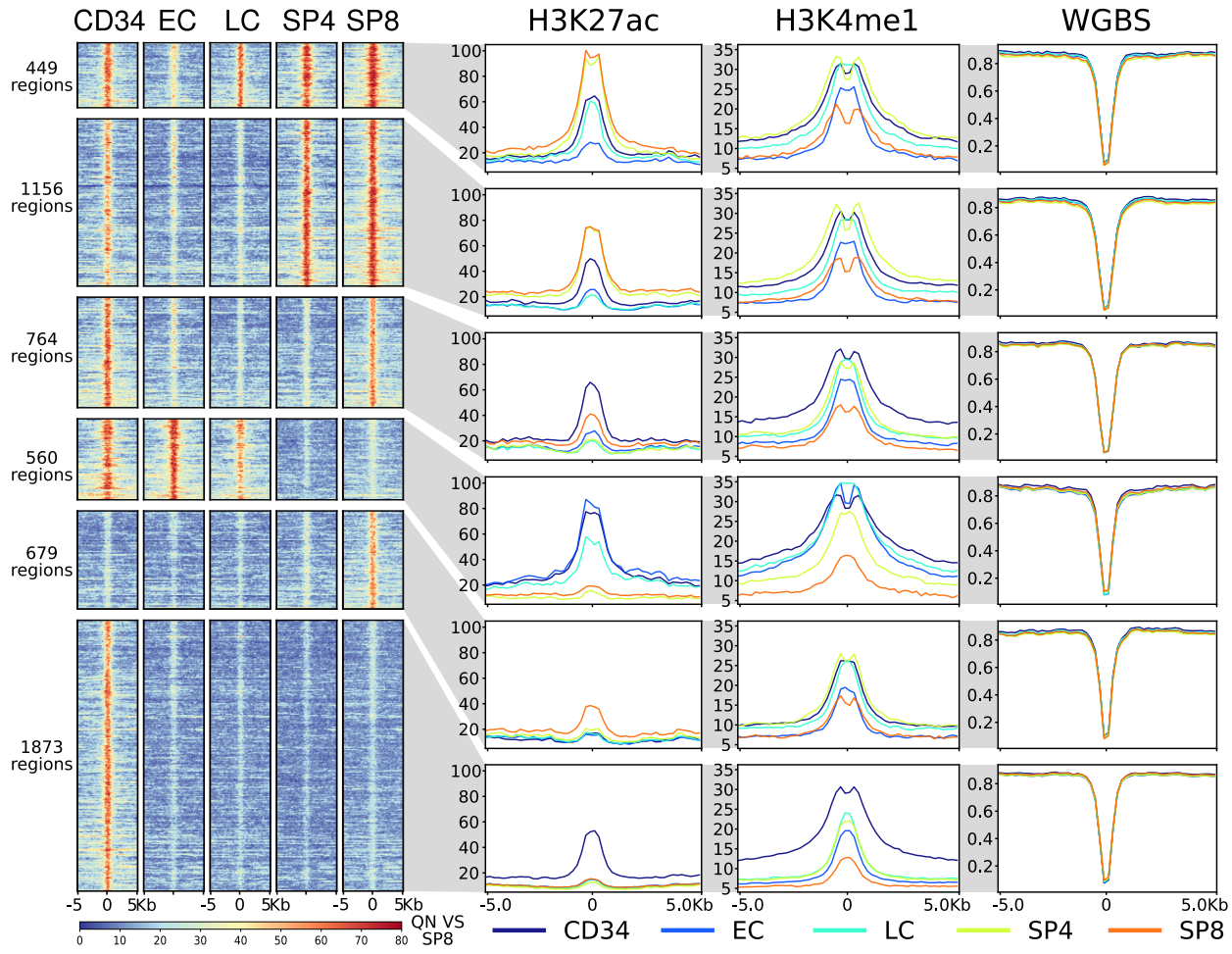
DNA hypomethylation is a hallmark of distal regulatory regions...



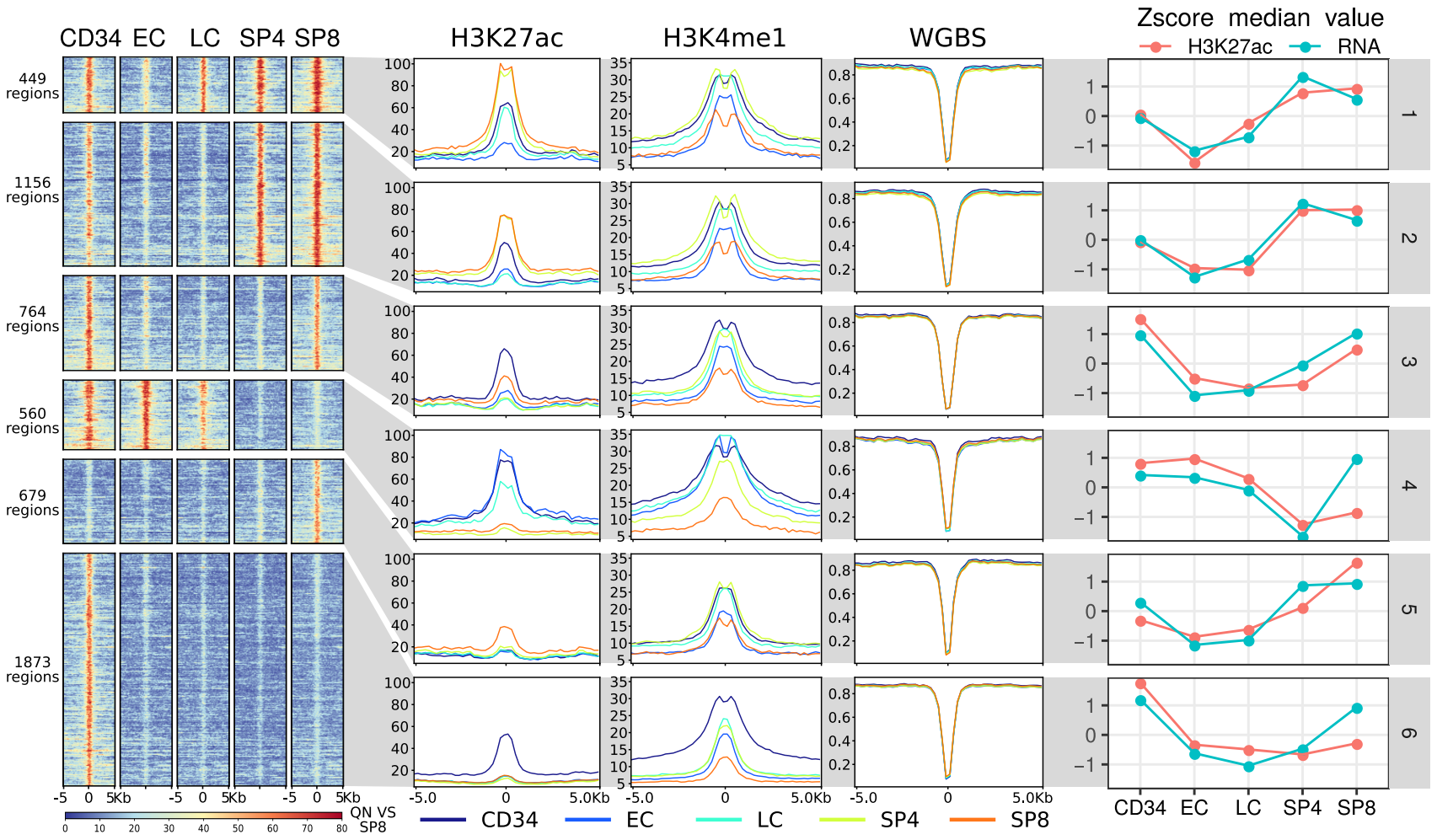
DNA hypomethylation is a hallmark of distal regulatory regions...



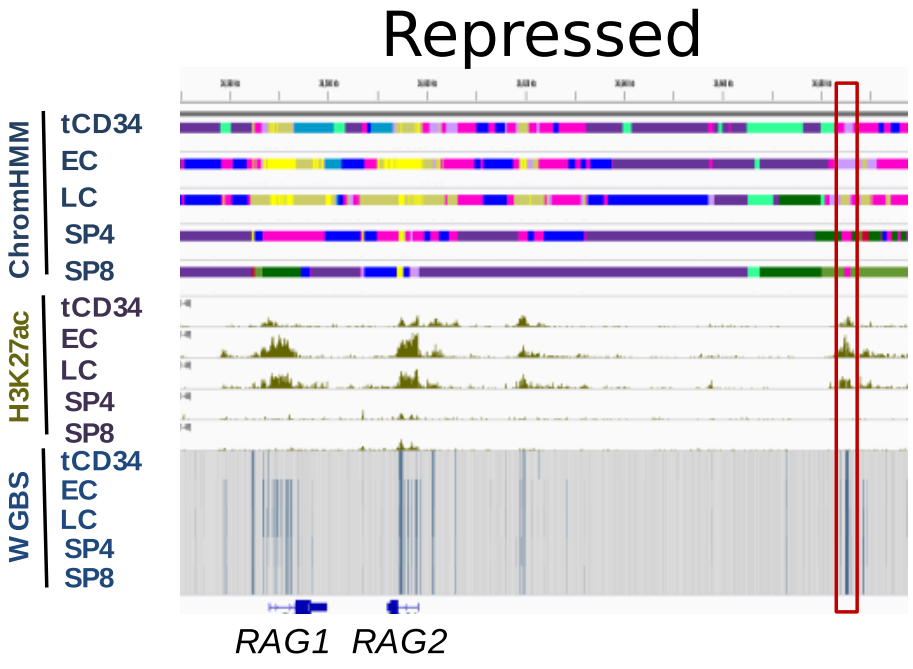
DNA hypomethylation is a hallmark of distal regulatory regions mostly irrespective of their activation status in T cells



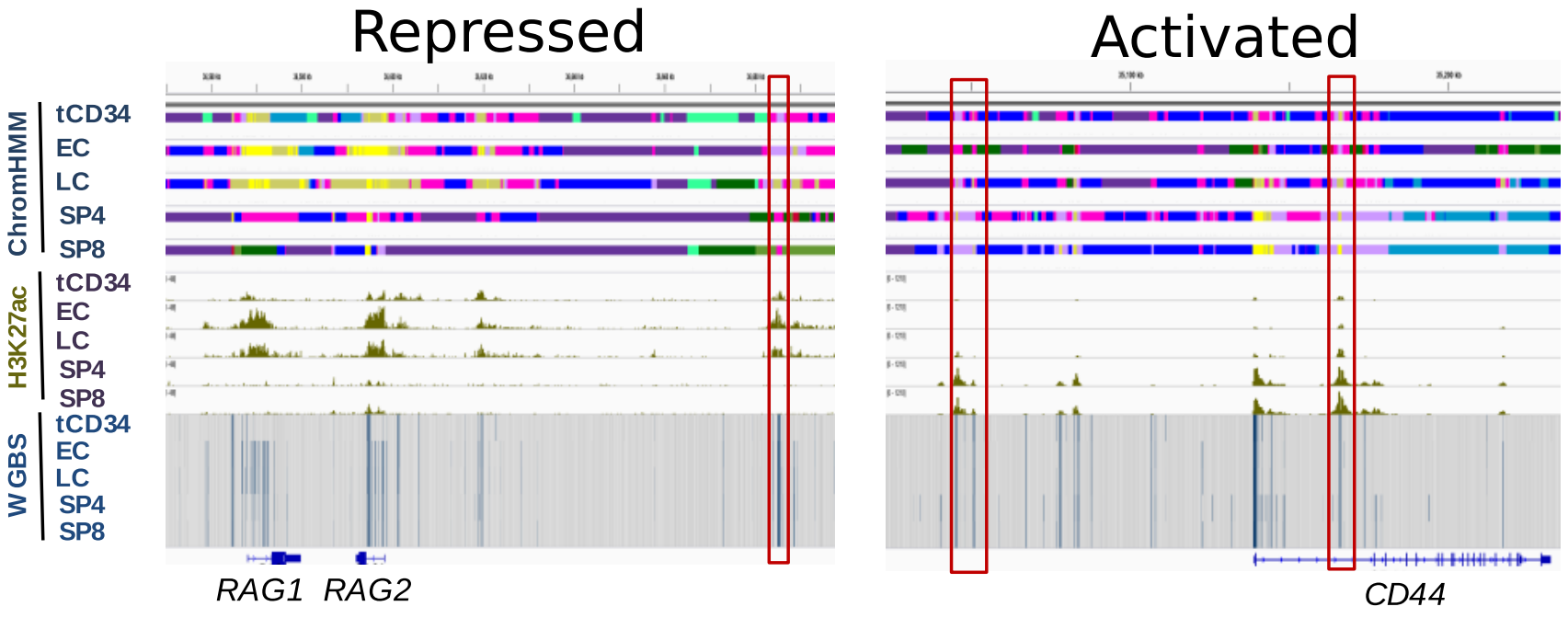
DNA hypomethylation is a hallmark of distal regulatory regions mostly irrespective of their activation status in T cells



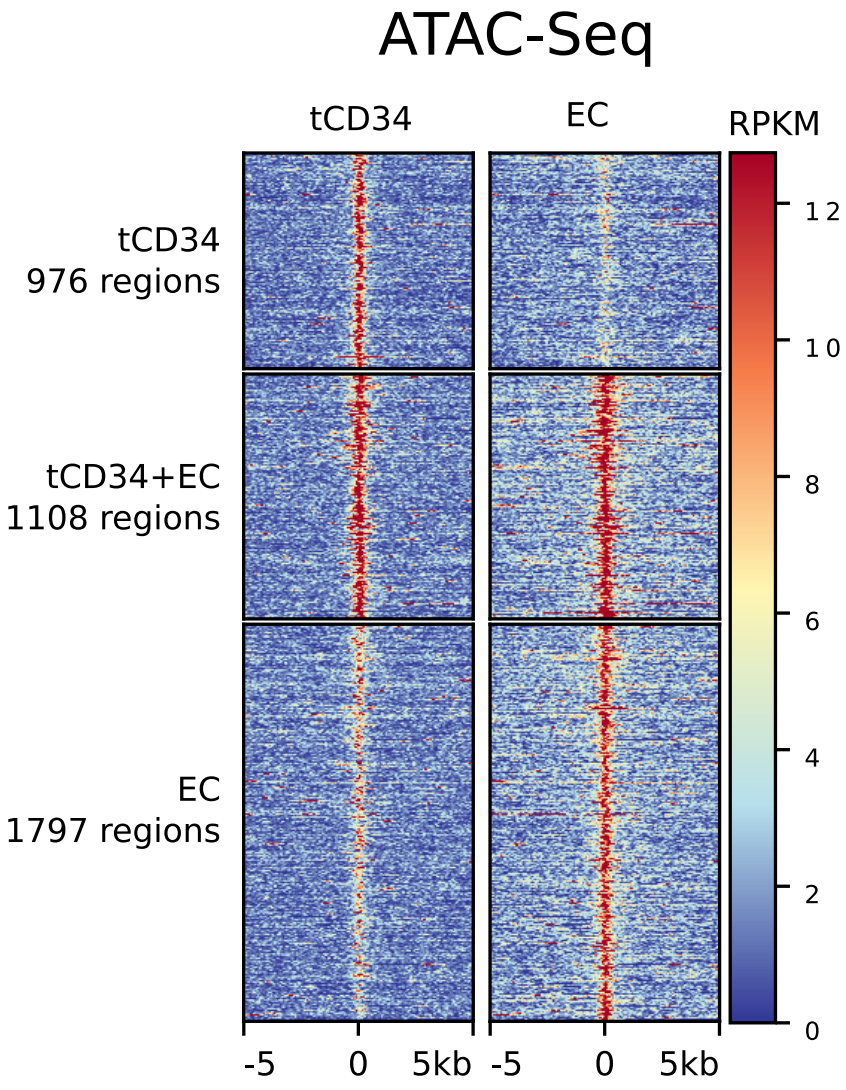
DNA hypomethylation is a hallmark of distal regulatory regions mostly irrespective of their activation status in T cells



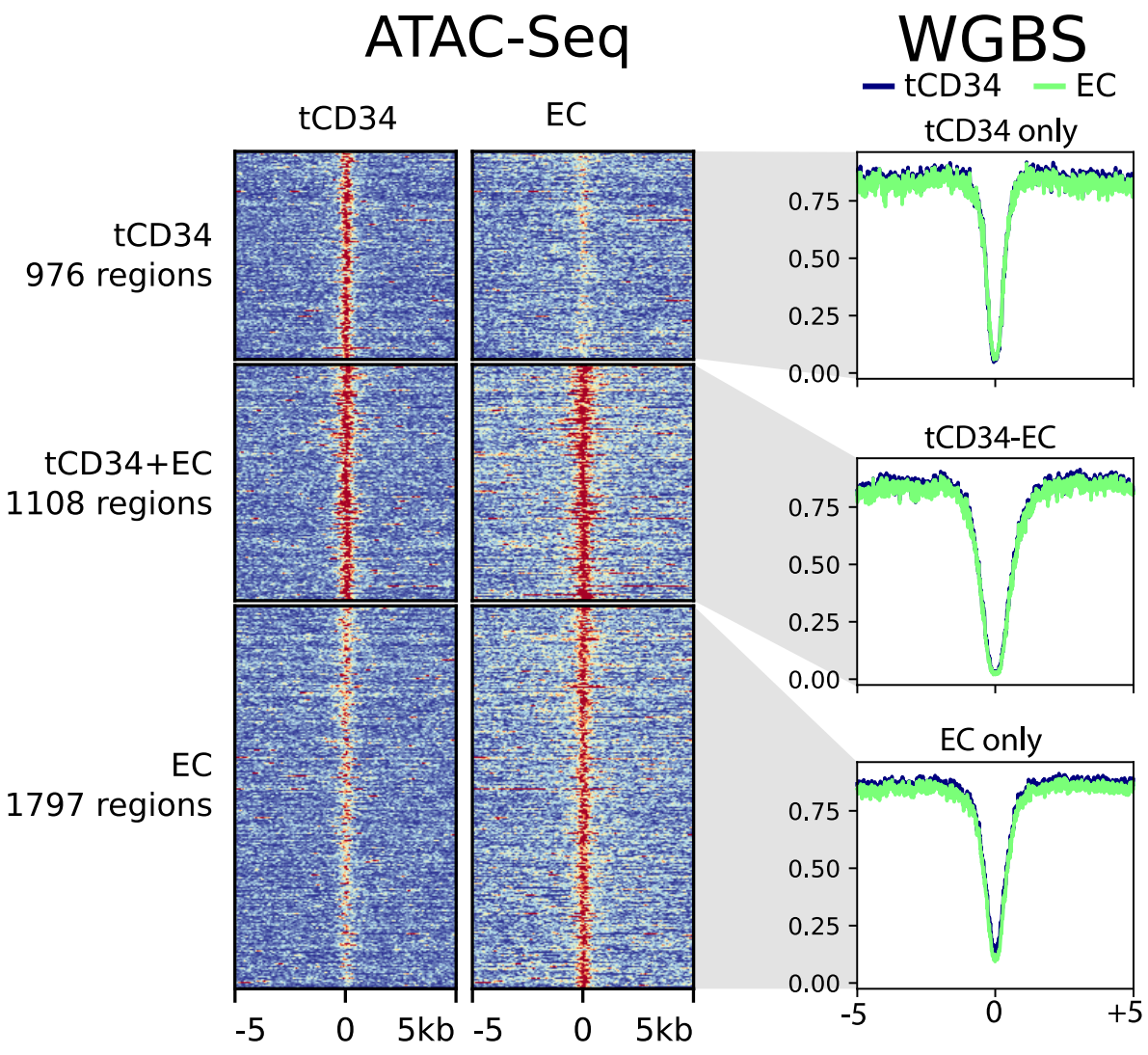
DNA hypomethylation is a hallmark of distal regulatory regions mostly irrespective of their activation status in T cells



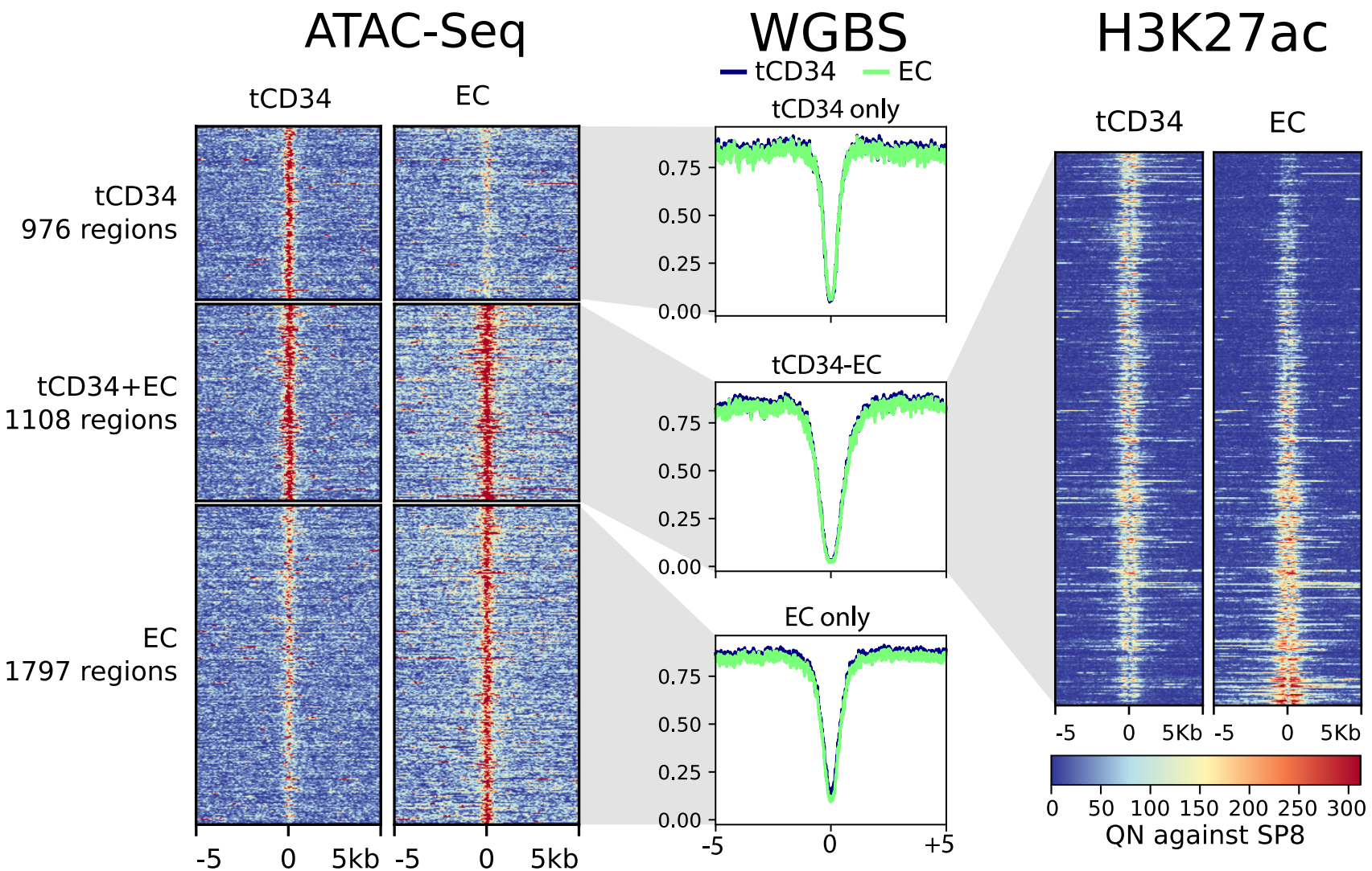
Chromatin opening dynamics of distal regions in early T cell differentiation...



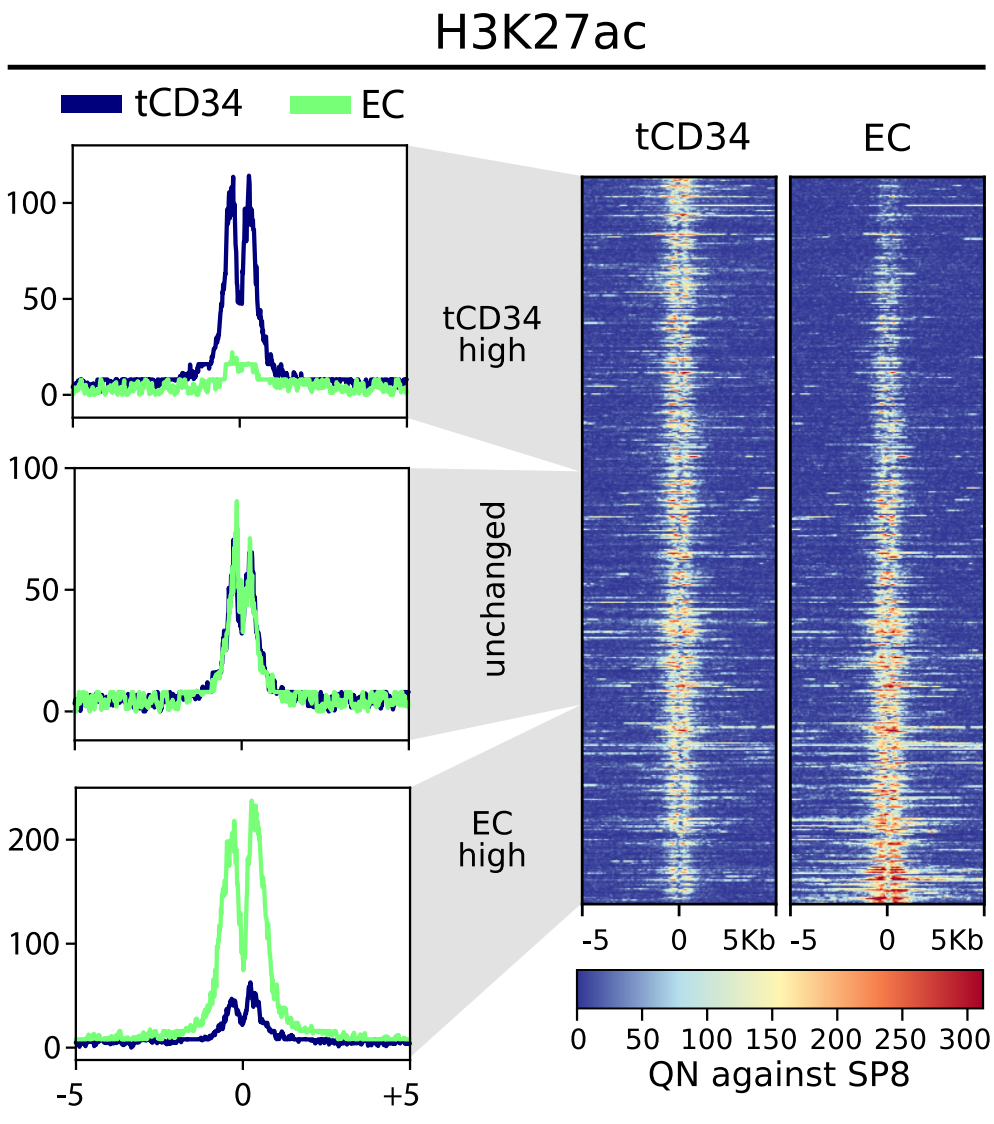
Dynamically open distal regions in early T cell differentiation are always hypomethylated



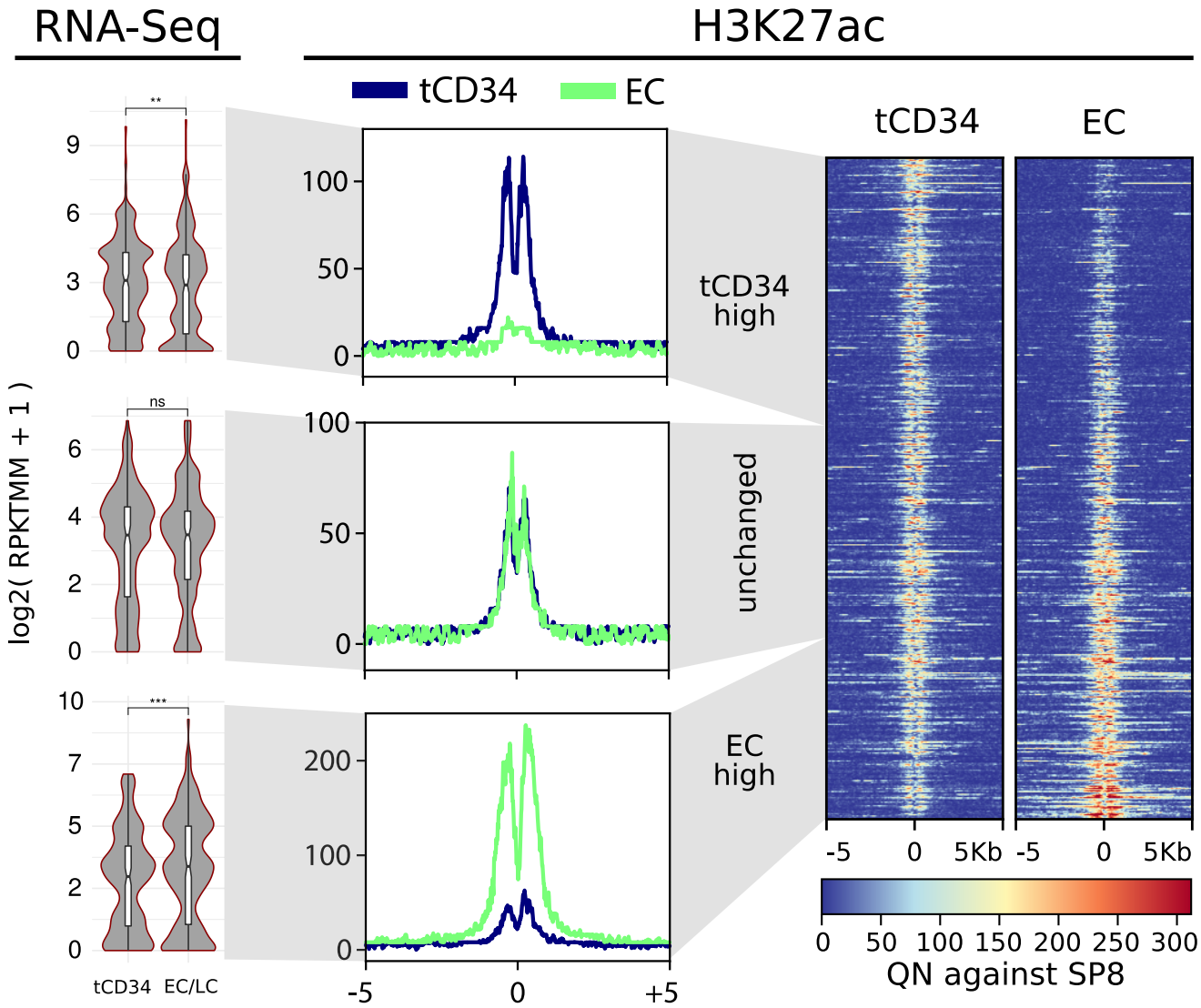
Constitutively open chromatin distal regions in early T cell differentiation exhibits H3K27ac dynamics...



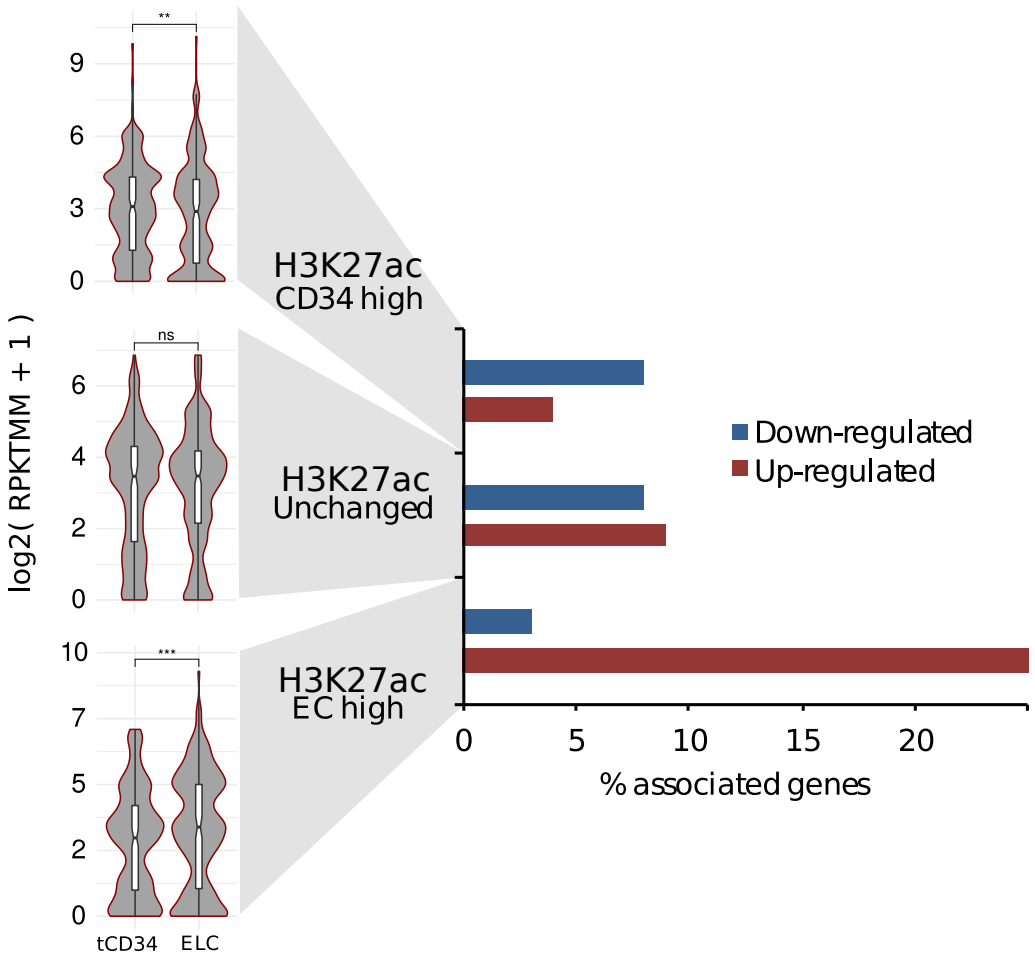
Constitutively open chromatin distal regions in early T cell differentiation exhibits H3K27ac dynamics...



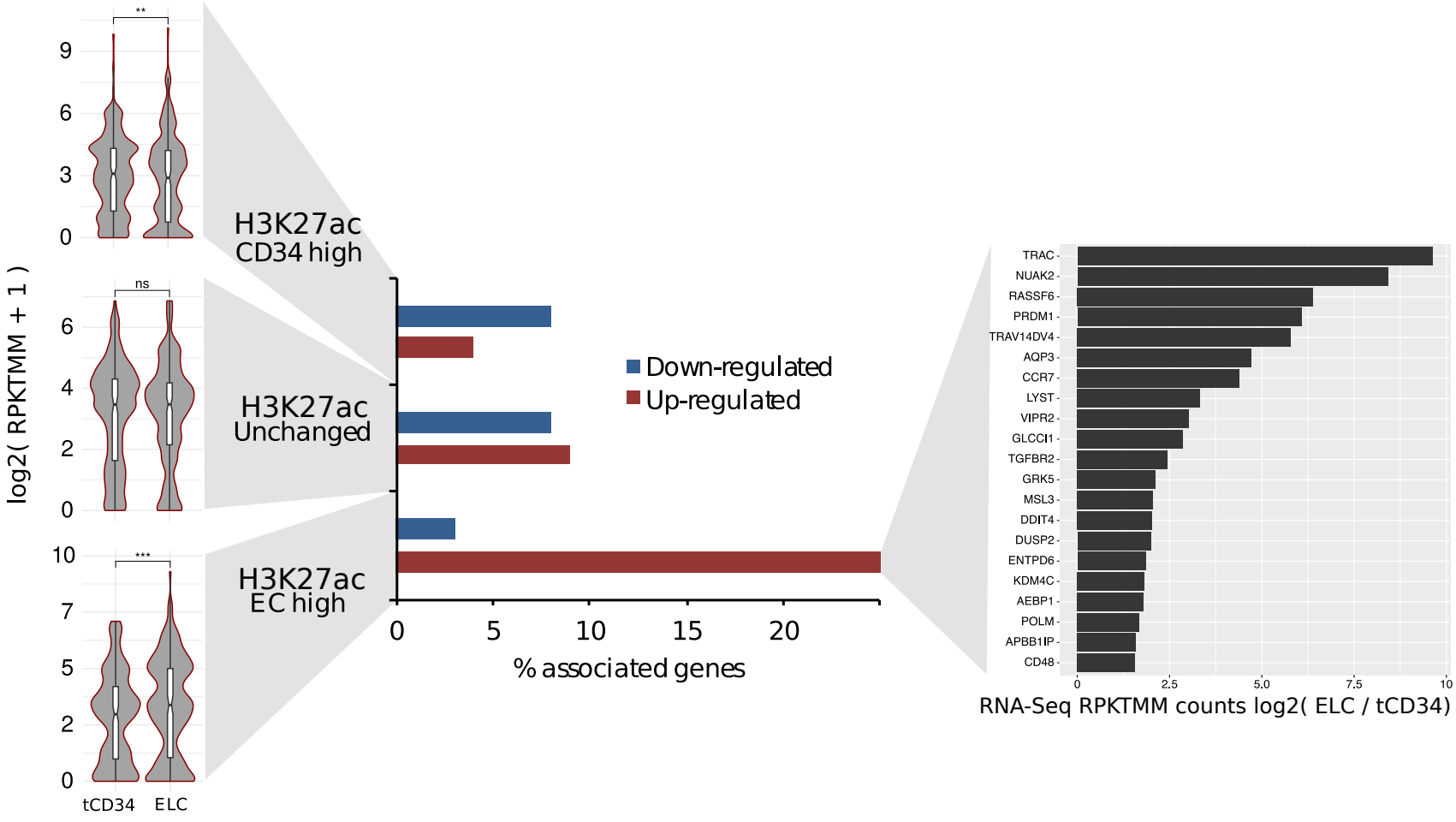
Constitutively open chromatin distal regions in early T cell differentiation exhibits H3K27ac dynamics associated with transcription of nearby genes



Differential expression analysis...



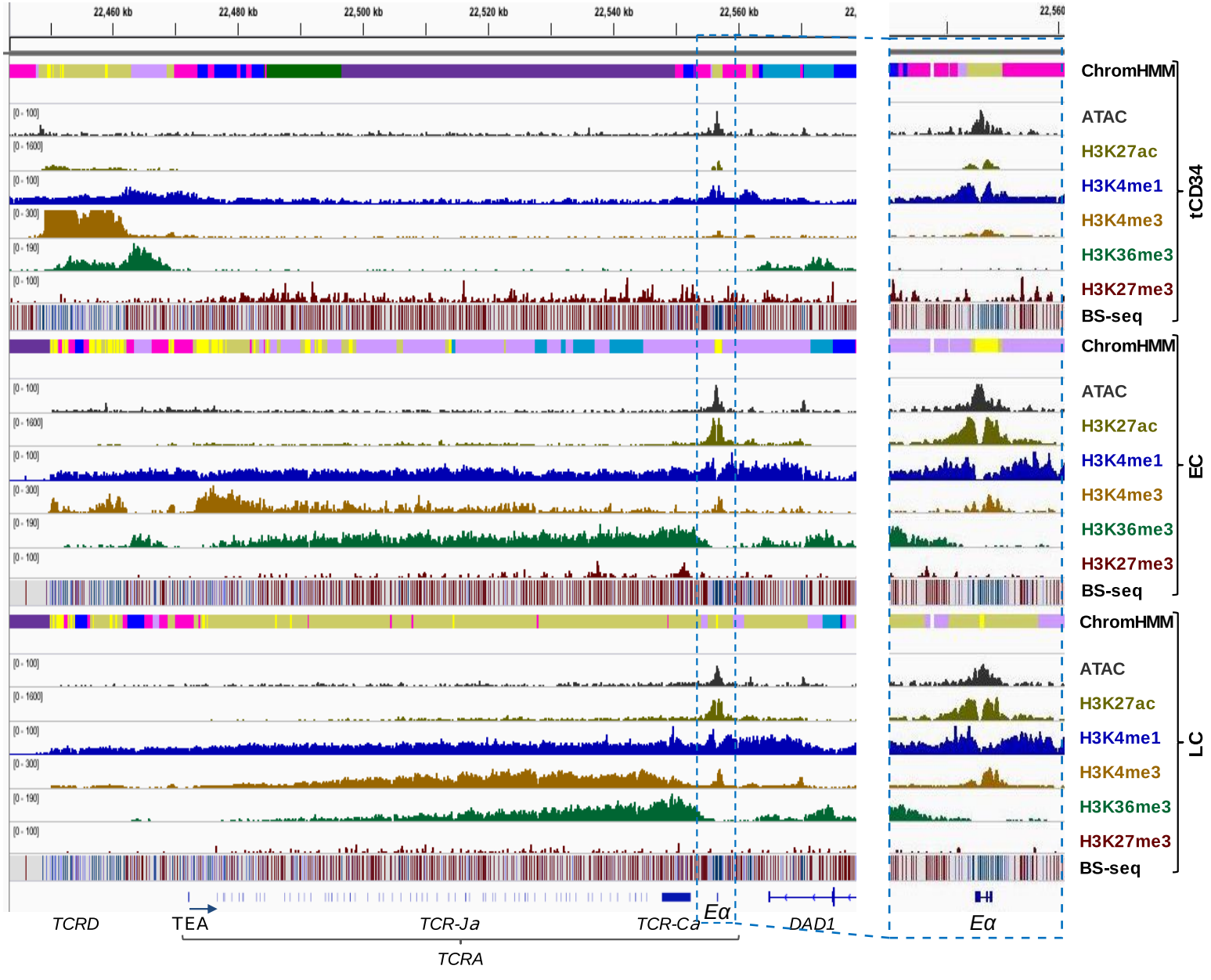
Differential expression analysis highlights genes associated with a dynamic putative enhancer



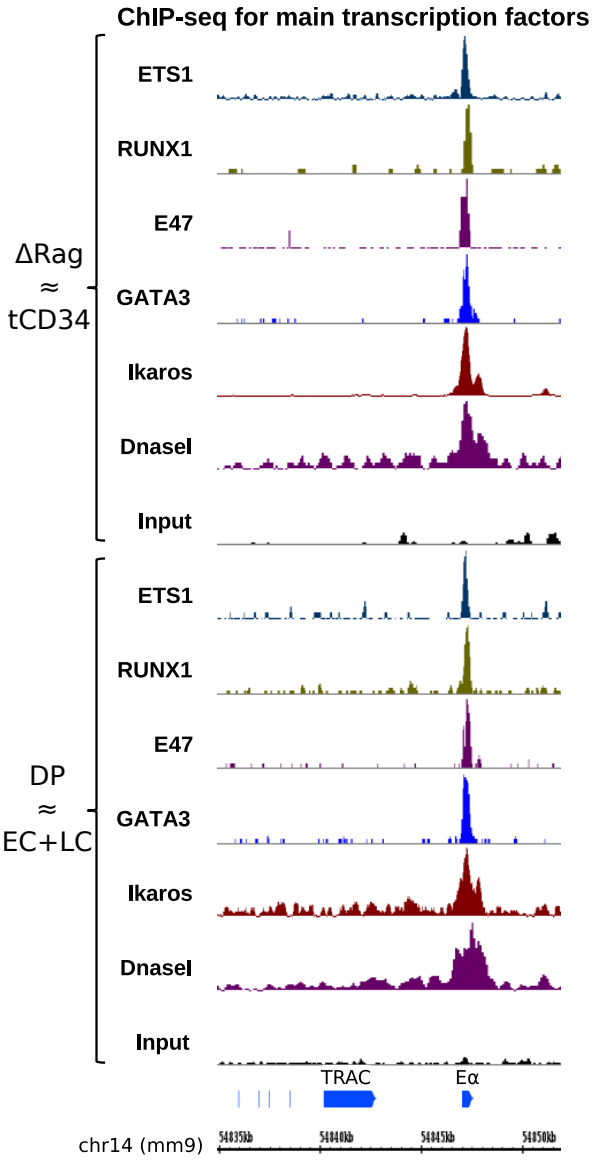
Thymopoiesis

Epigenetic regulation of TCRA locus

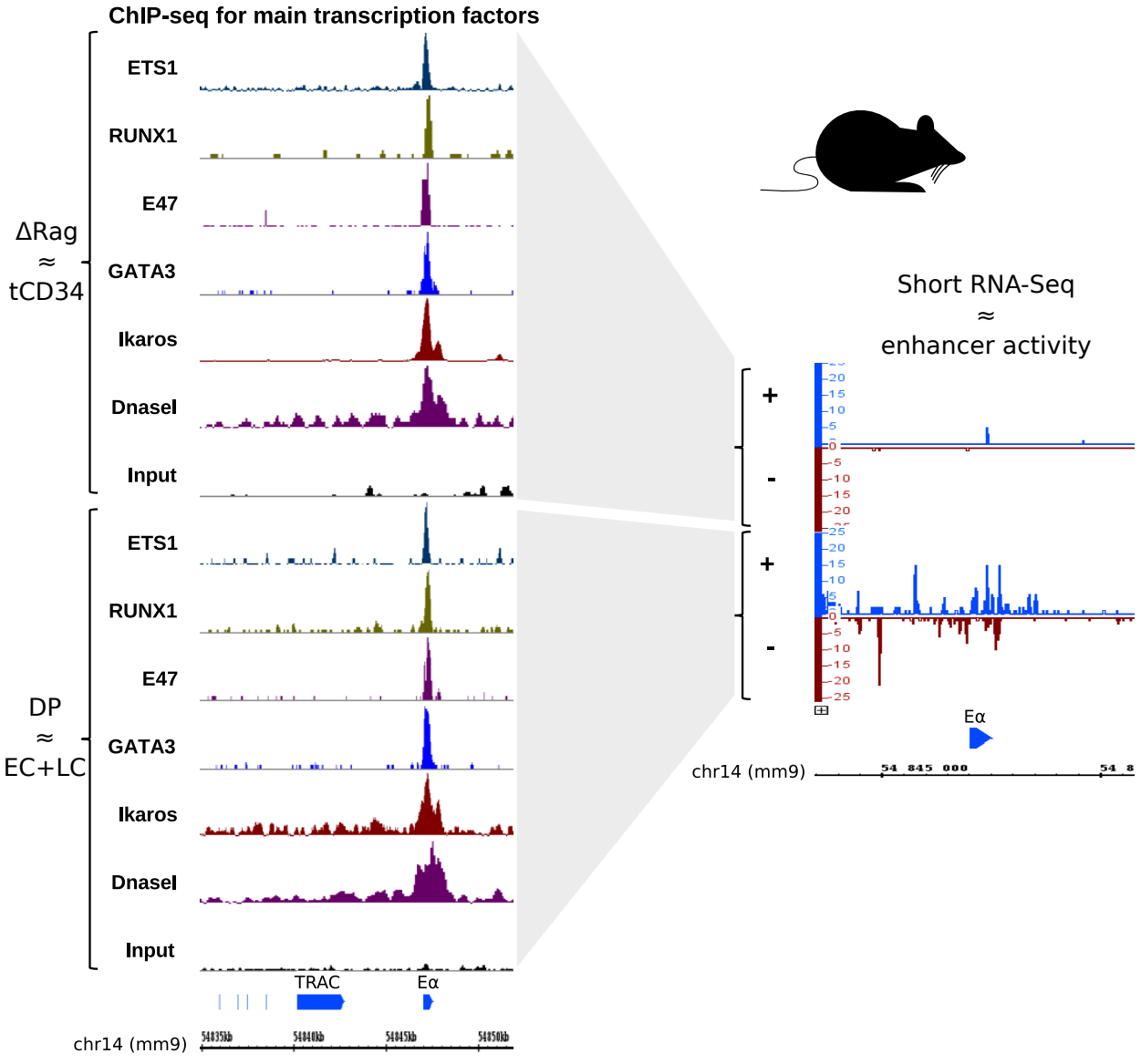
Epigenetic dynamics of TCRA locus



Activators binds TCRA enhancer in early T cell differentiation...

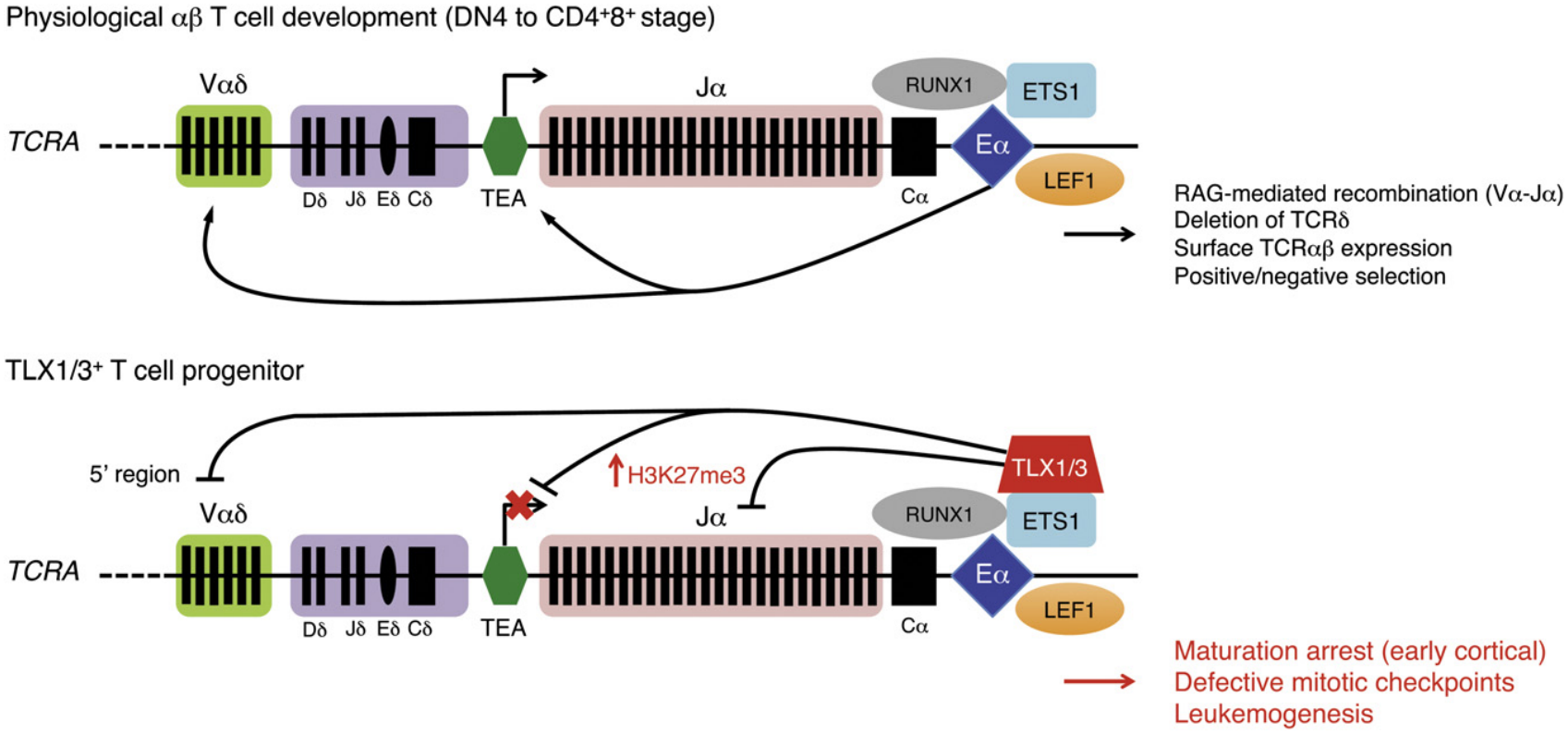


Activators binds TCRA enhancer in early T cell differentiation but the locus activates late



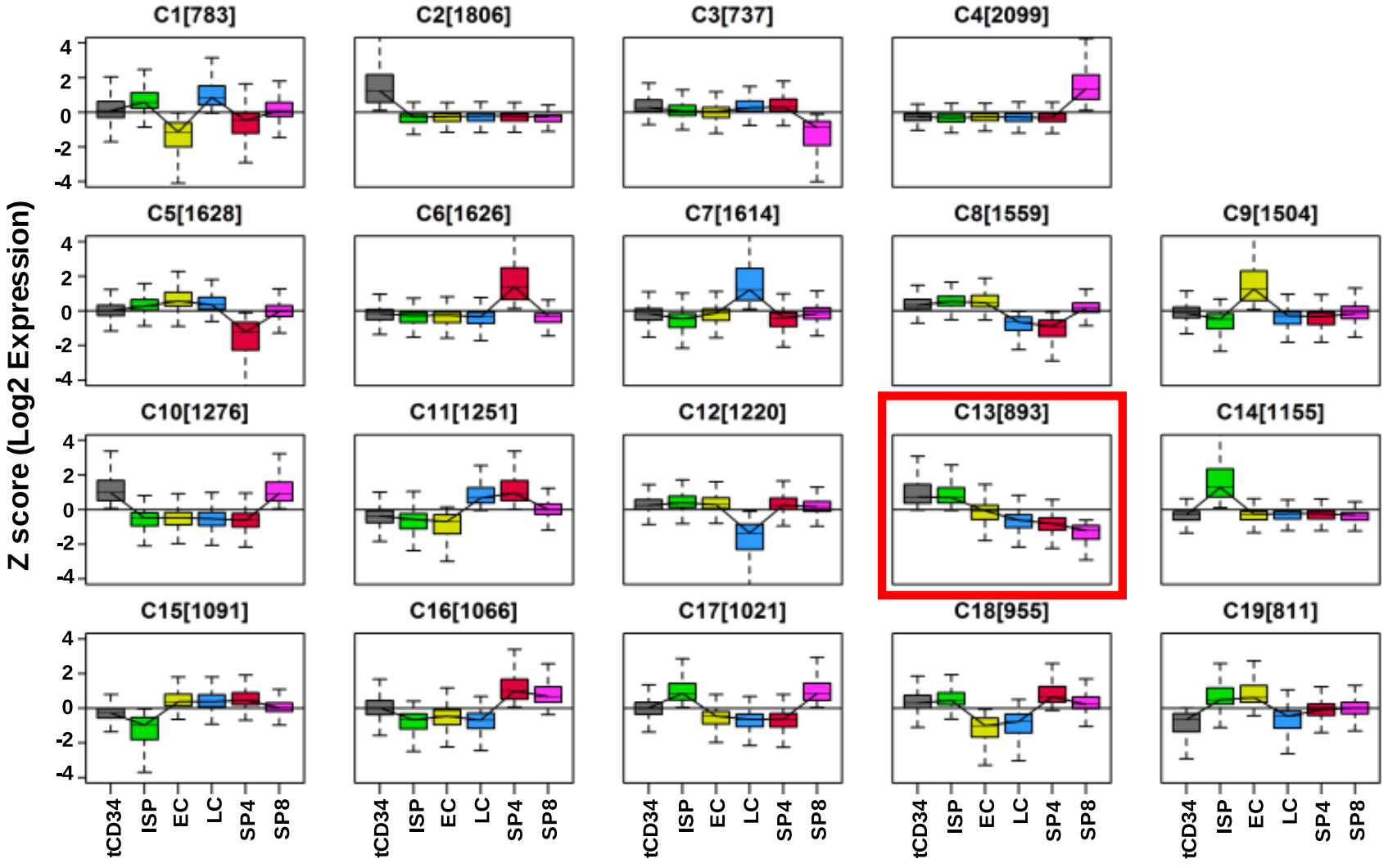
Looking for a potential candidate repressor

TLX homeodomain oncogenes mediate T cell maturation arrest in T-ALL via interaction with ETS1 and suppression of TCRα gene expression

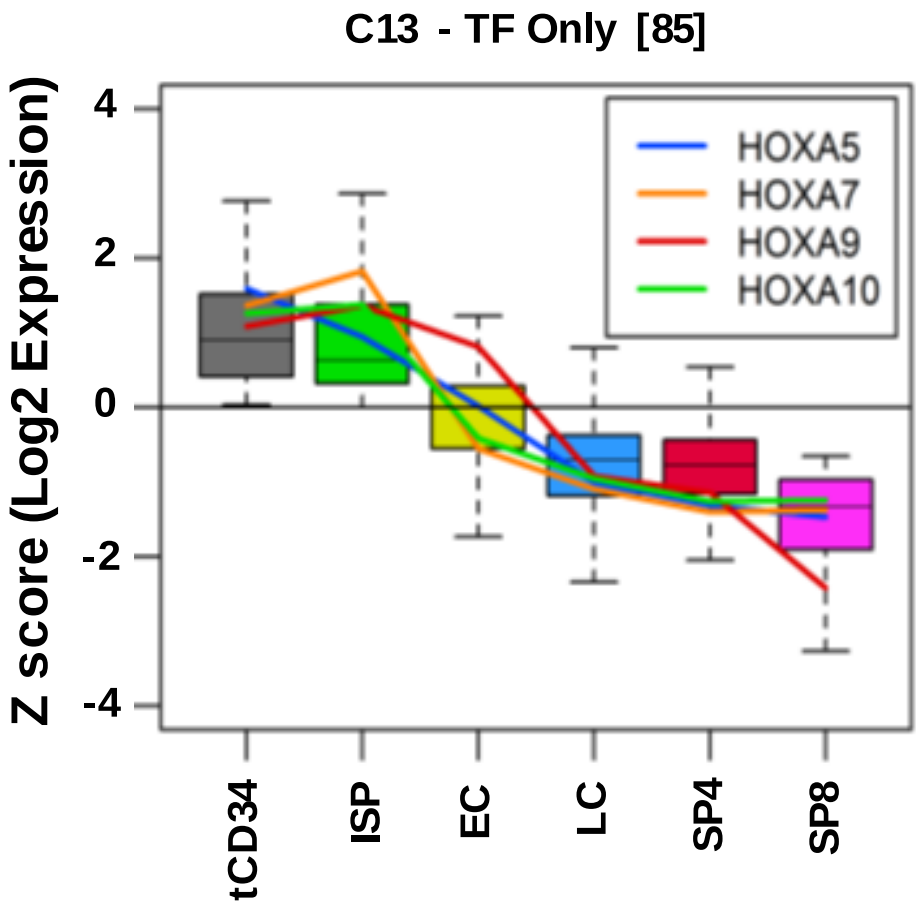


Original discovery by *Dadi, Spicuglia, Asnafi et al., 2012, Cancer Cell*
Figure taken from *King, Ntziachristos & Aifantis, 2012, Cancer Cell*

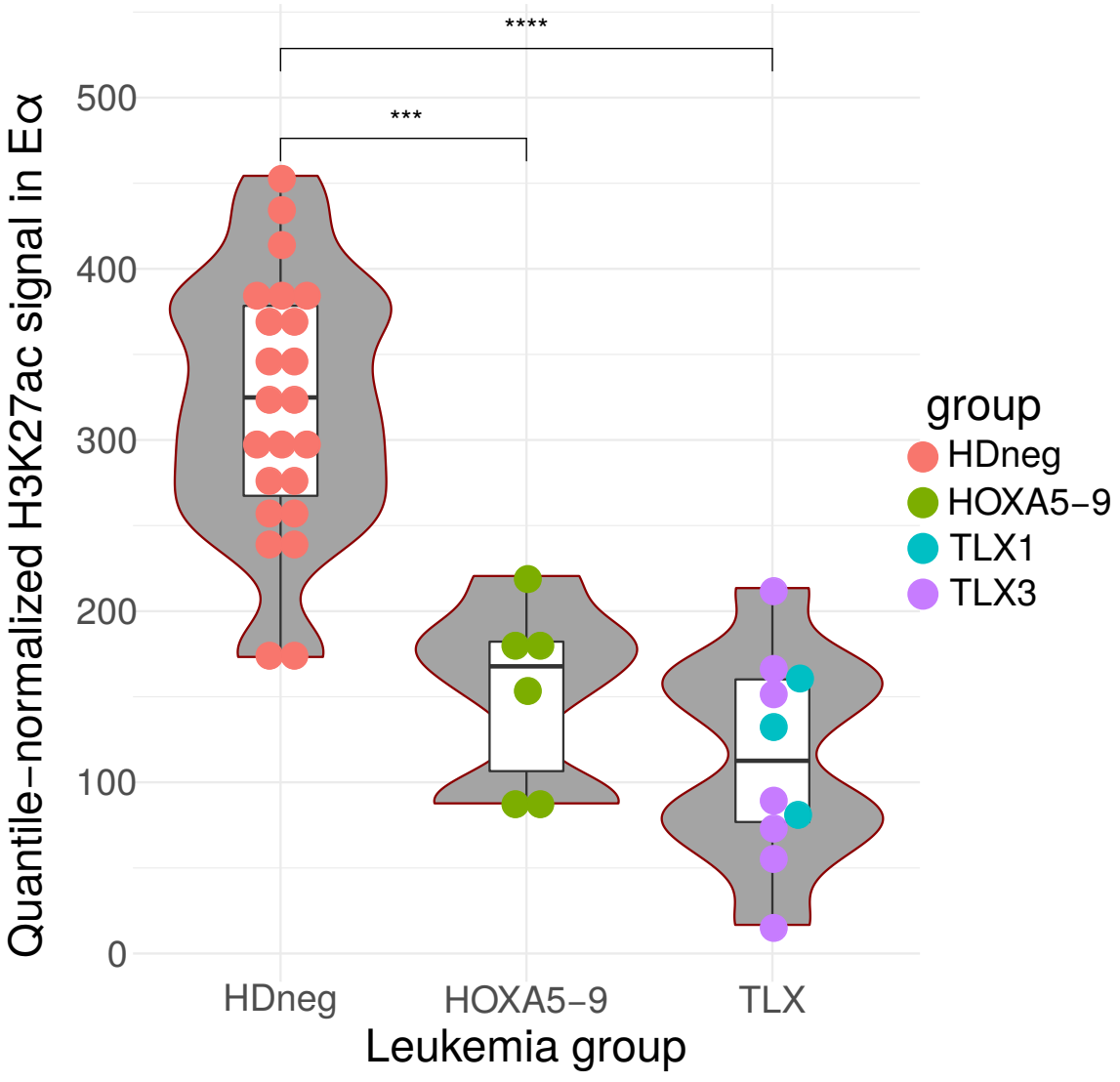
Looking for a potential candidate repressor in transcriptomic data...



Looking for a potential candidate repressor in transcriptomic data highlights HOXA family genes

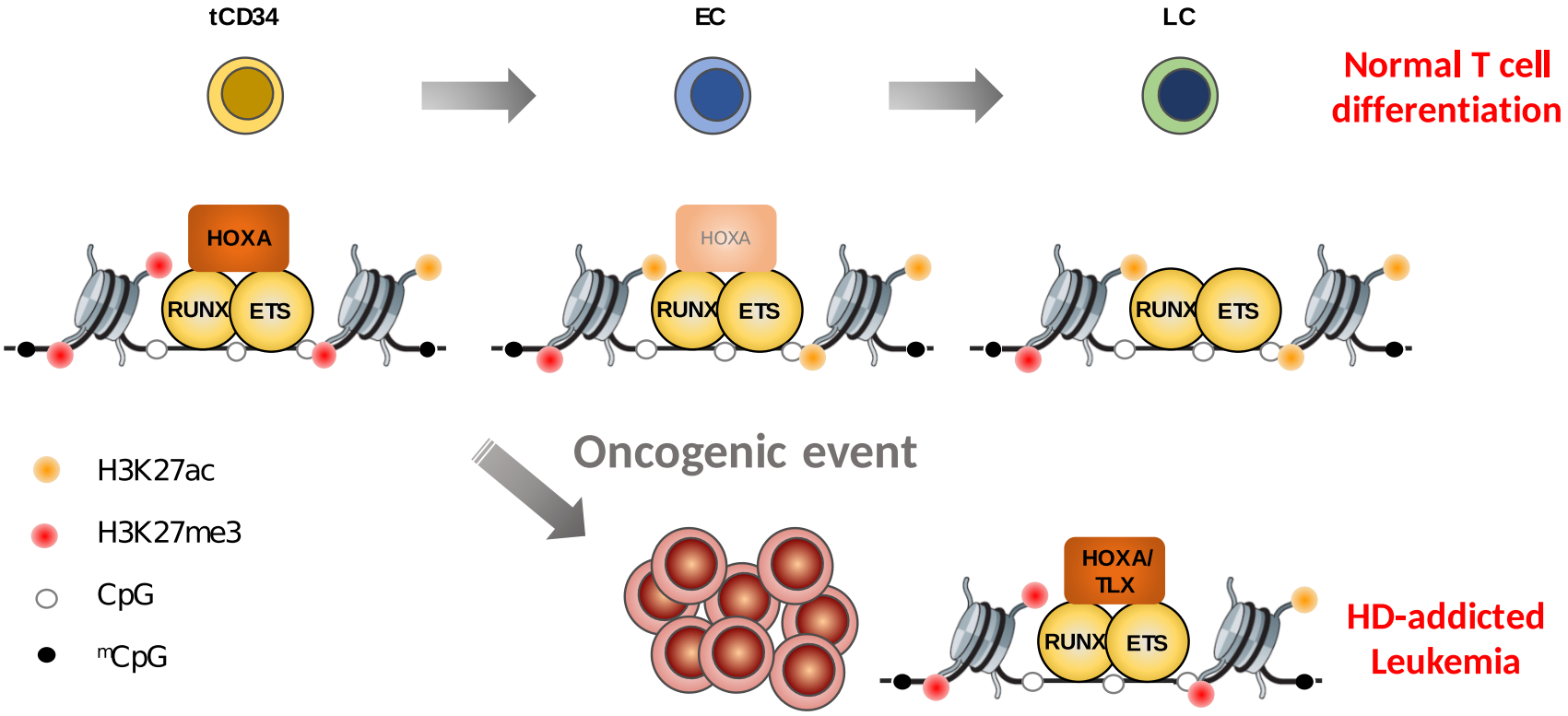


Mature T-lymphoblastic leukemias provide another evidence for Homeobox family genes repressive action on Eα



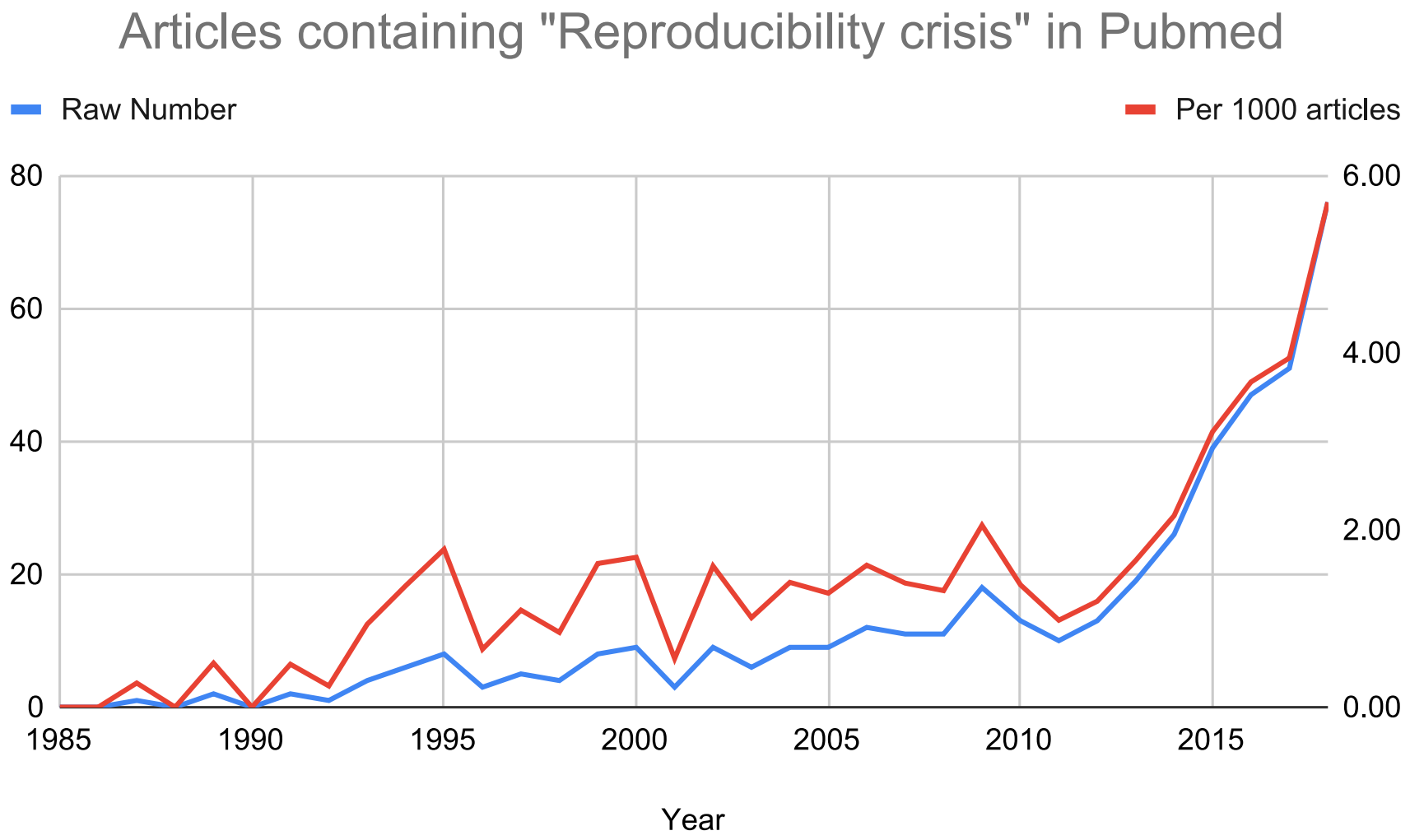
HOXA5-9 genes repression validated by biological approaches from Necker collaborators

Mechanistic model for Eα activation in T cell differentiation

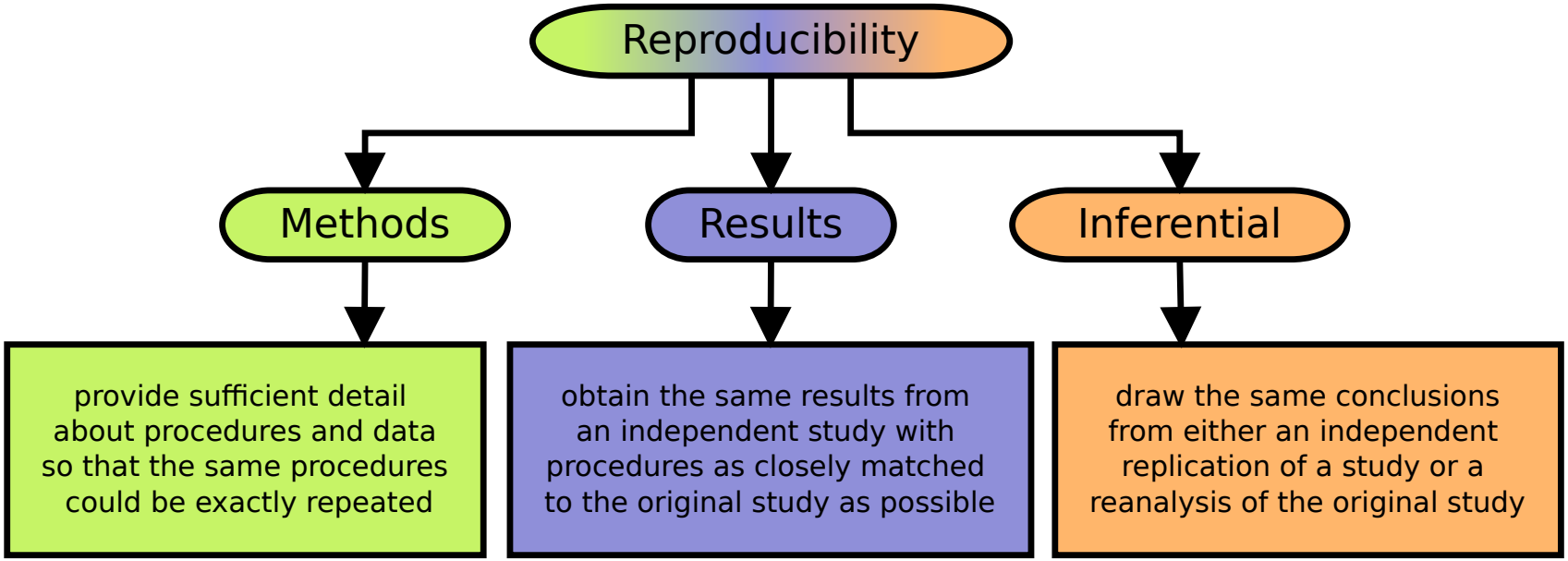


Reproducibility

A trending topic

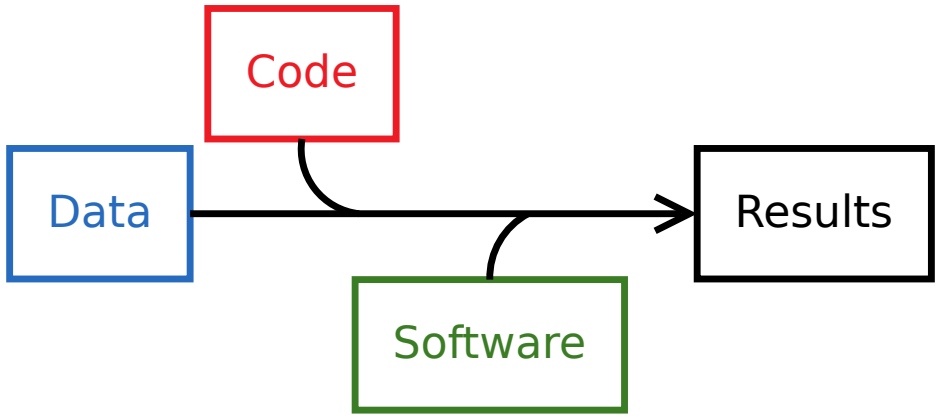


Defining reproduciblity



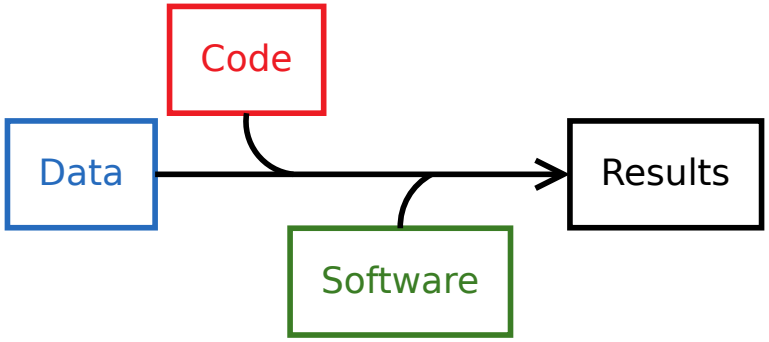
Definitions from Goodman et al., 2016, Science Translational Medicine

Main requirements for methods reproducibility



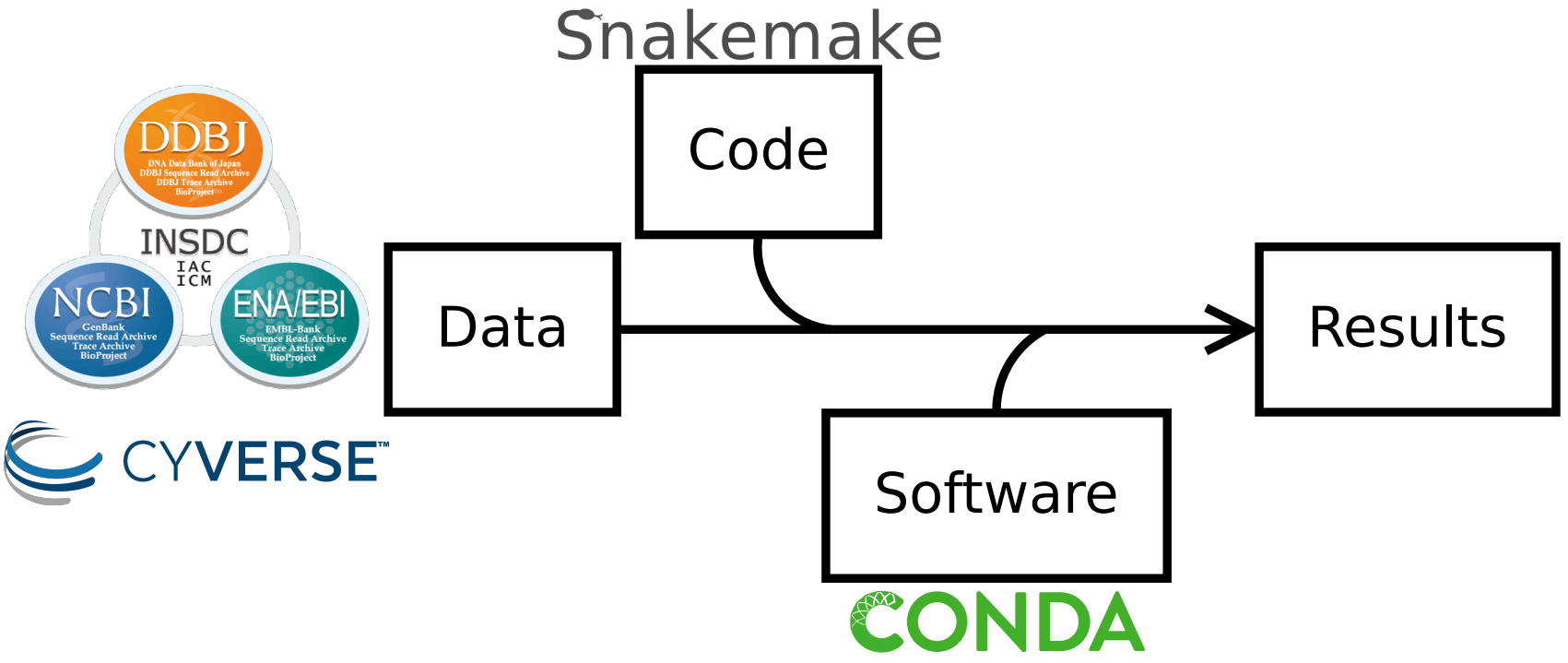
Component	Requirements	Explanations
Data	Findable Accessible Interoperable Reusable	FAIR data principles
Software	Portable	Tools with permissive licence User-level OS-independent package management Automatic environment deployment
Analysis Code	Traceable	Tools and parameters for each step Documentation of order, input and output
	Automation	Minimal human action required to reproduce analysis

Additional requirements for massive genomic data handling



Component	Requirements	Explanations
Data	Findable Accessible Interoperable Reusable	FAIR data principles
Software	Portable	Tools with permissive licence User-level OS-independent package management Automatic environment deployment
Analysis Code	Traceable	Tools and parameters for each step Documentation of order, input and output
	Automation	Minimal human action required to reproduce analysis
	Scalable	Straightforward and efficient parallelization of tasks Stoppable and resumable analyses Clusters and Clouds supports

Elements of solution for reproducibility in bioinformatics

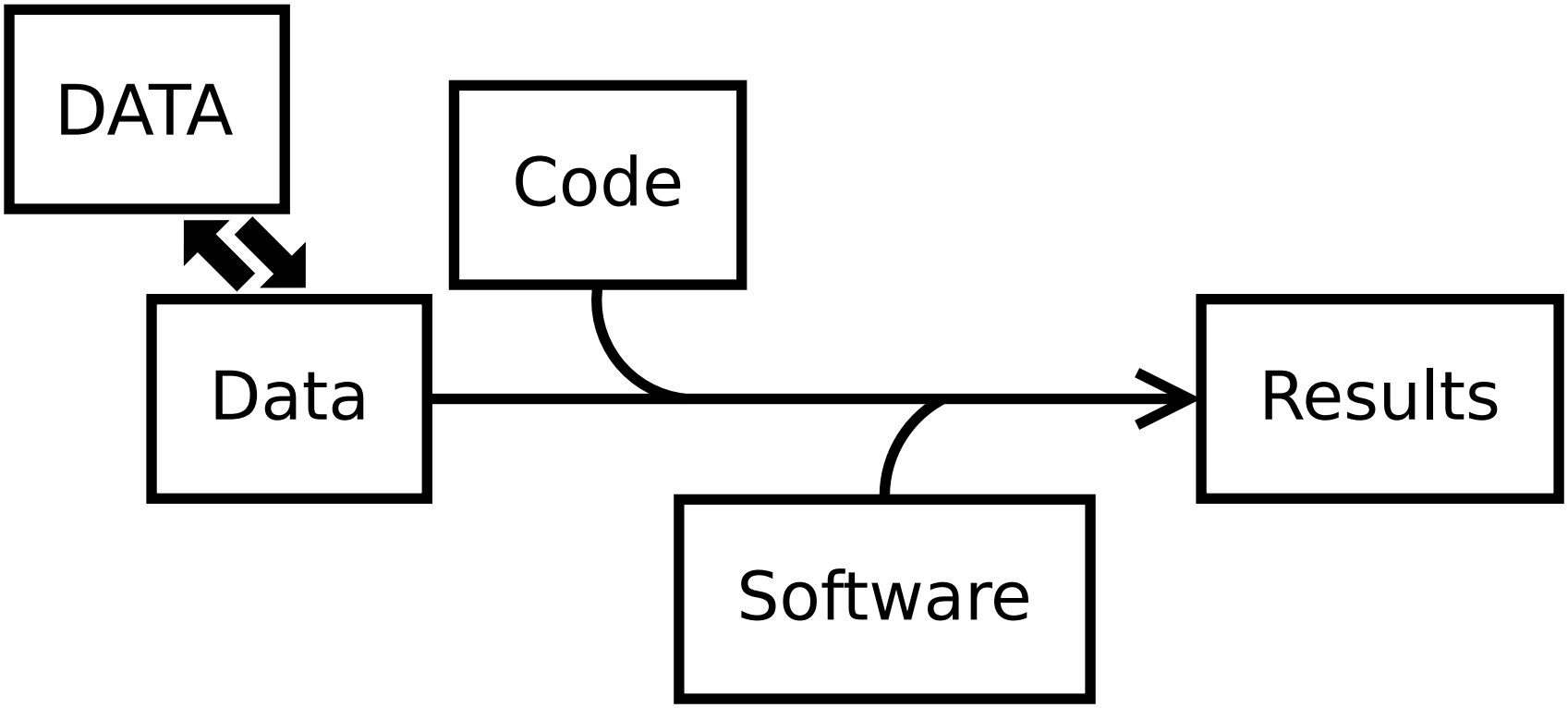



```
git clone URL_for_workflow/repo.git
cd repo
snakemake --use-conda
```

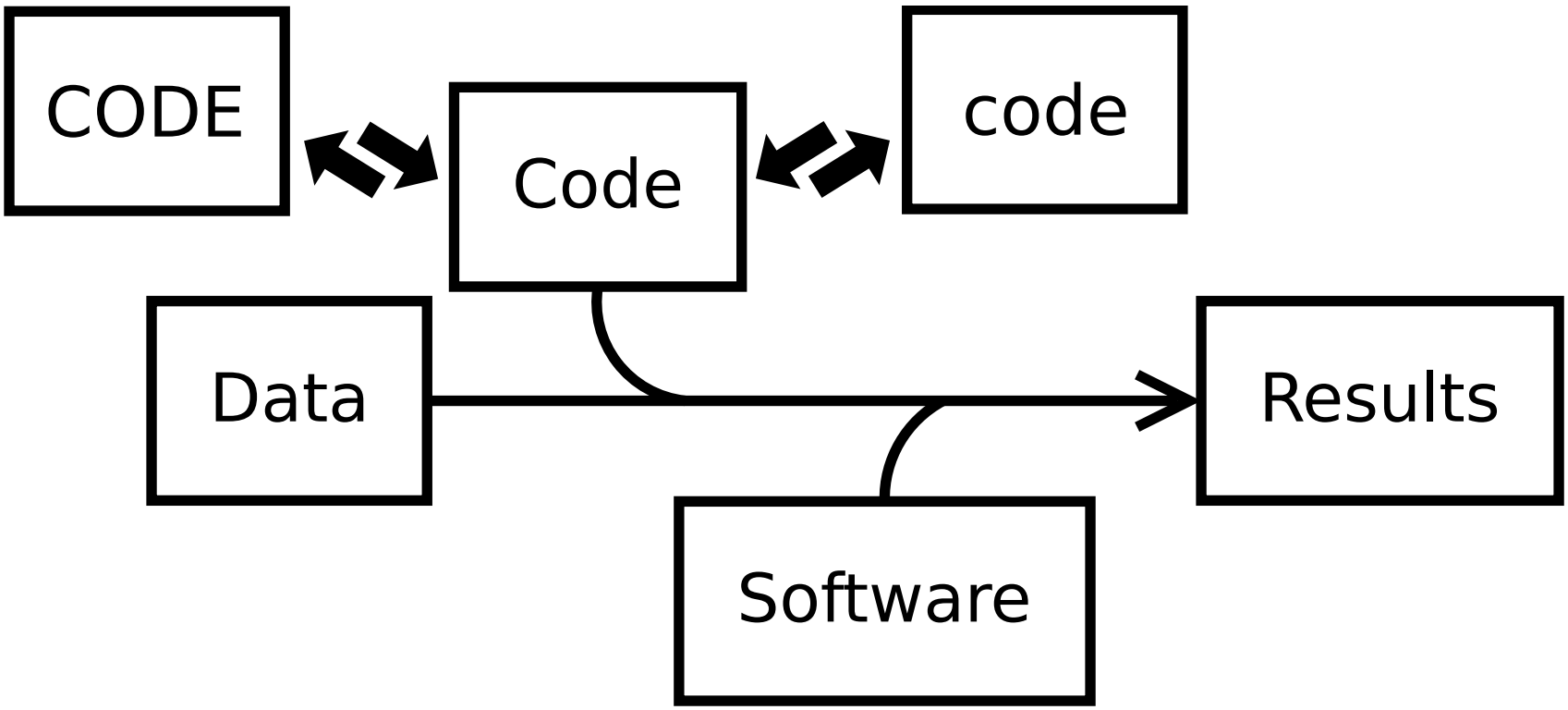
Reproducibility

Beyond reproducibility

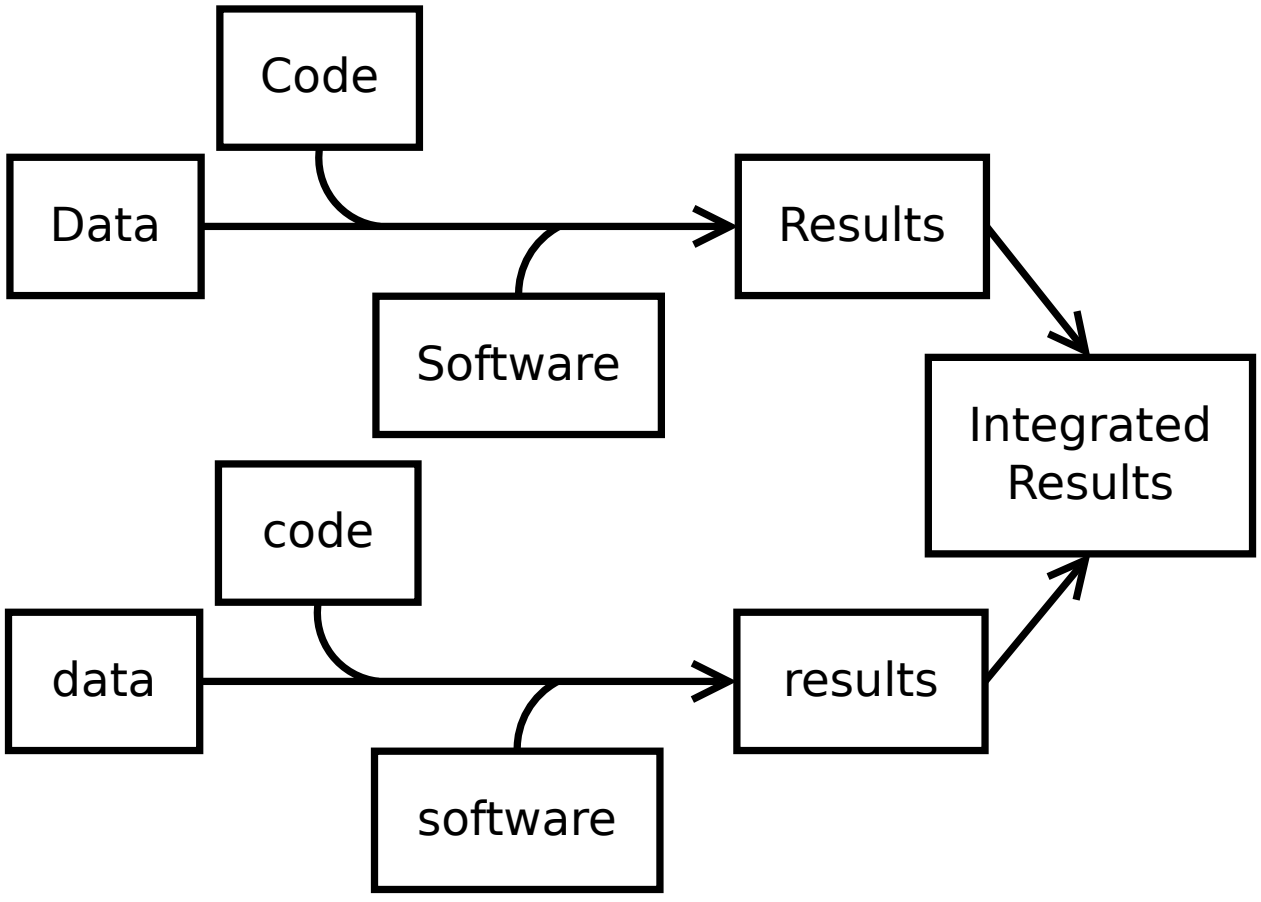
Reuse scenario



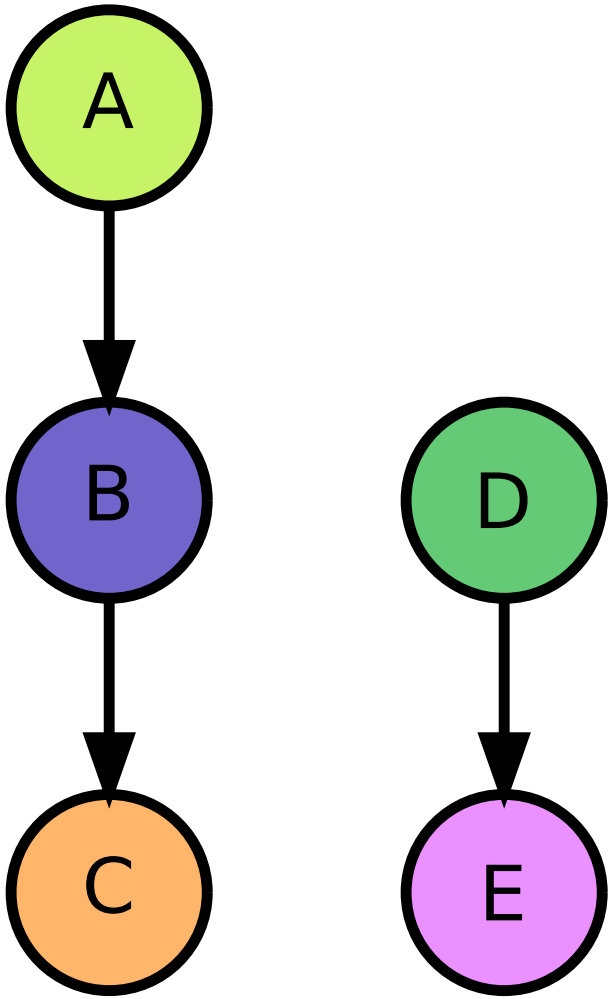
Alternative protocol exploration scenario



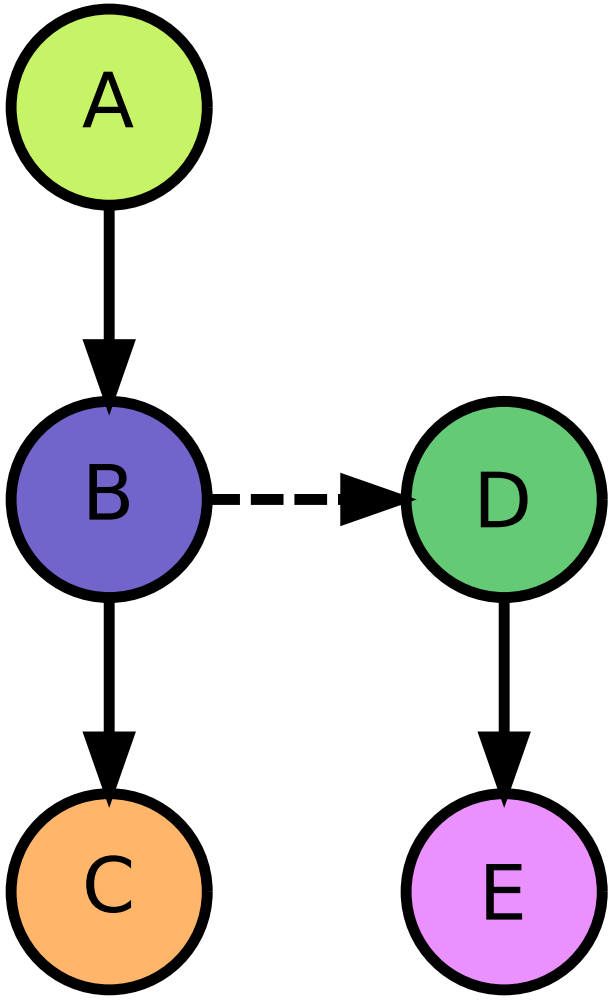
Integration of multiple workflows into one analysis



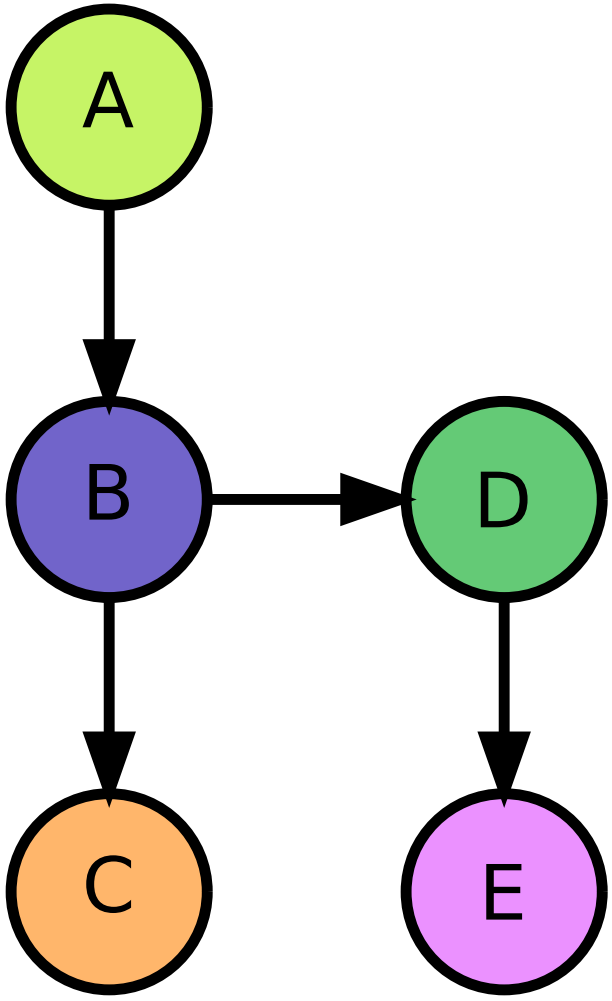
Integrating workflows using subworkflows...



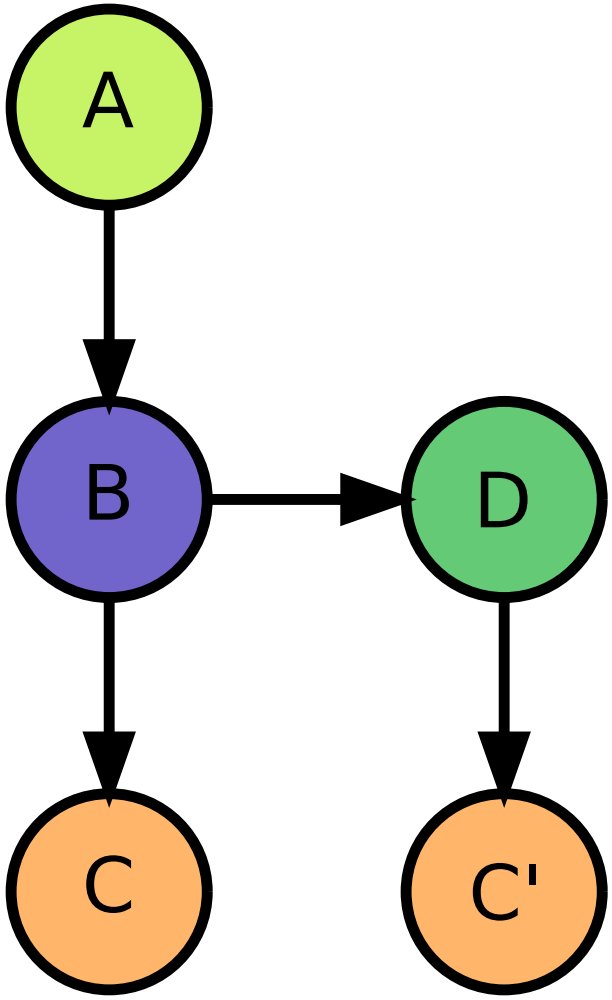
Integrating workflows using subworkflows may leads to poor parallelization



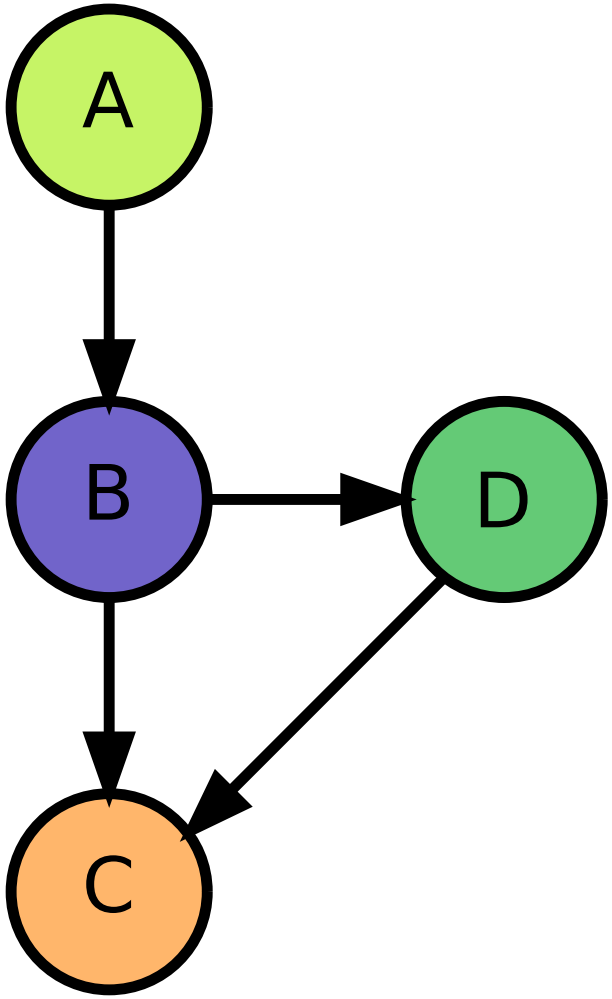
Integrating workflows using merged workflow lead to optimal parallelization



Integrating workflows using merged workflow allows to reduce code redundancy

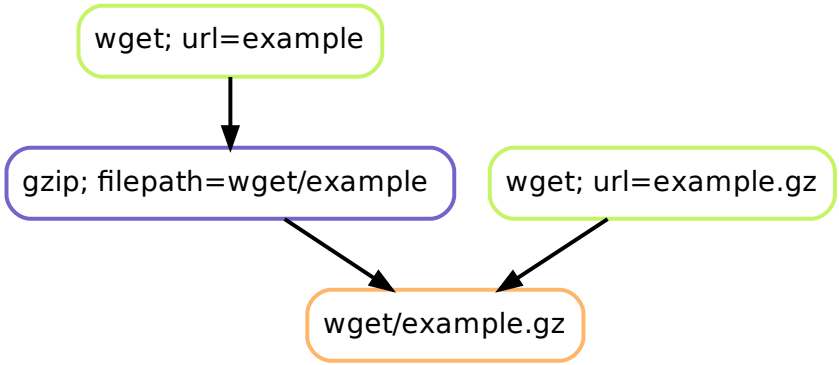


Integrating workflows using merged workflow allows to reduce code redundancy



Snakemake freedom can hurt: Ambiguous rule exception

```
1  rule gzip:
2     "Compress any file to gz"
3     input:
4        "{filepath}"
5     output:
6        "{filepath}.gz"
7
8  rule wget:
9     "Download any file from Internet"
10    output:
11       "wget/{url}"
```



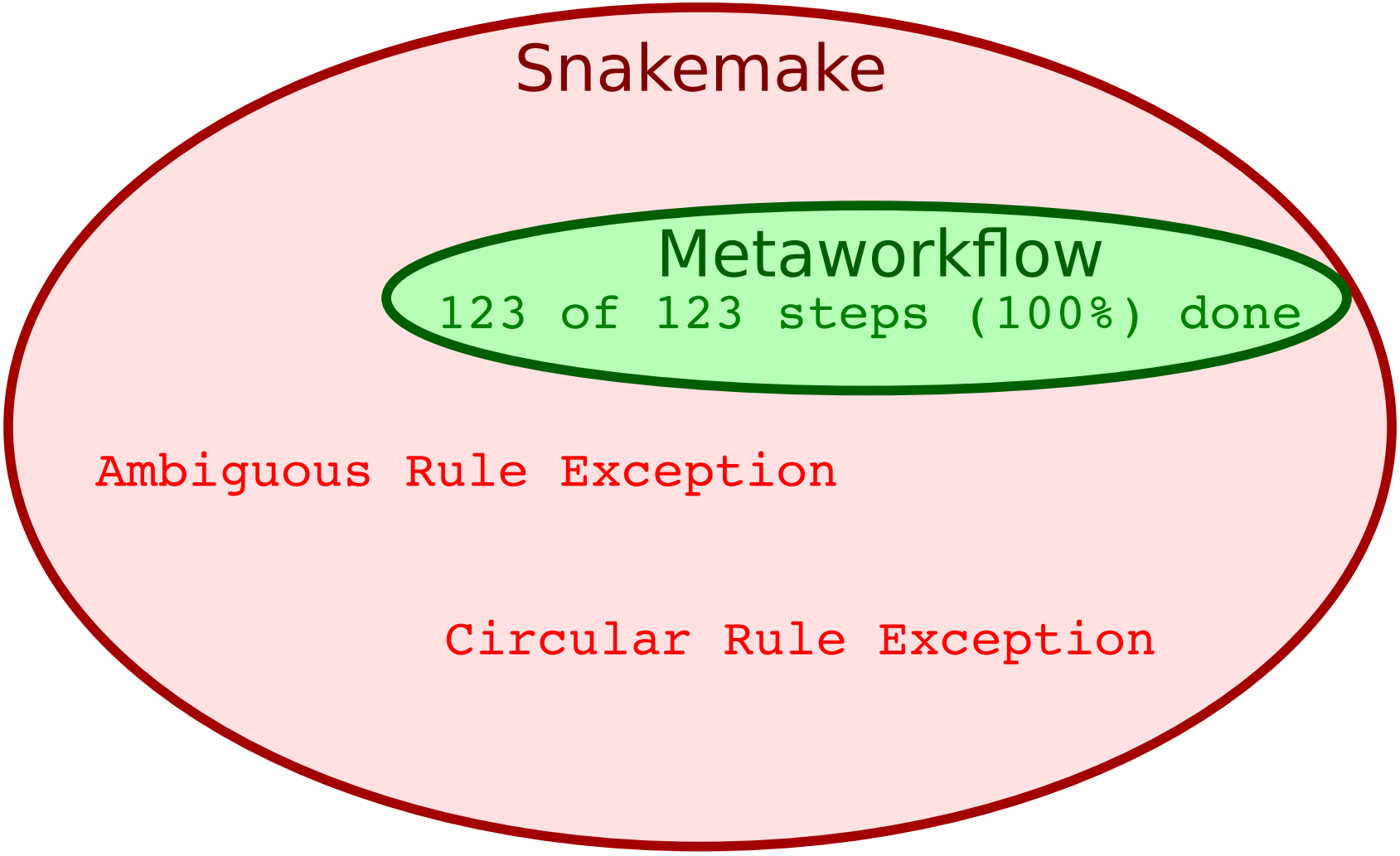
Methodological objectives

Ensure no ambiguity can arise between rules
while maintaining large scope for rules

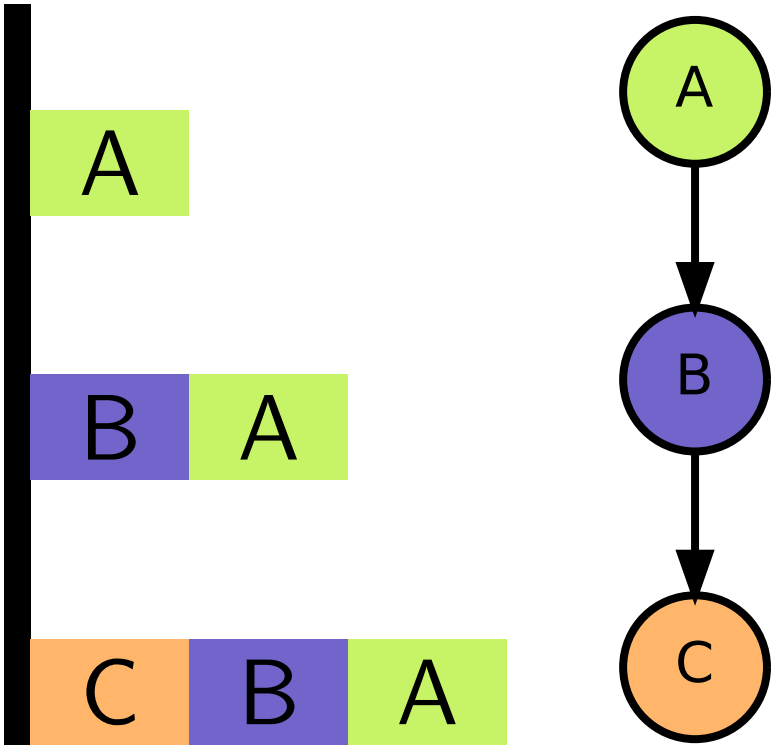
Write a unique workflow to compute any possible analyses

Store a bioinformatician thesis project,
a whole research career,
or a project of any size with multiple collaborators
inside a reproducible, scalable and lightweight workflow

Metaworkflow limits Snakemake freedom to bare minimum



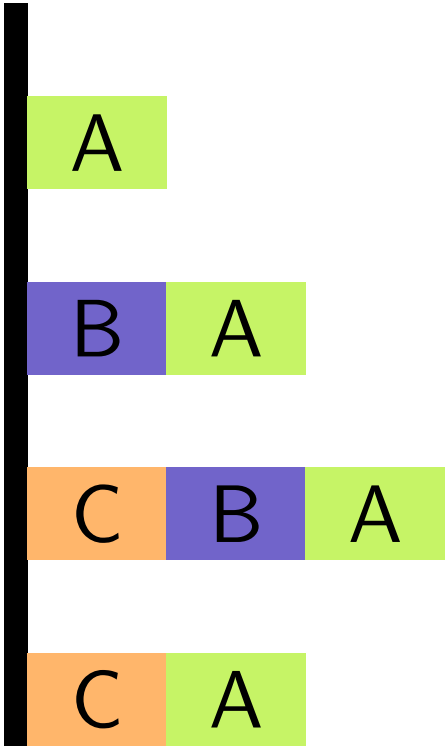
All rules can and should be coerced to write their outputs in their own unique directory

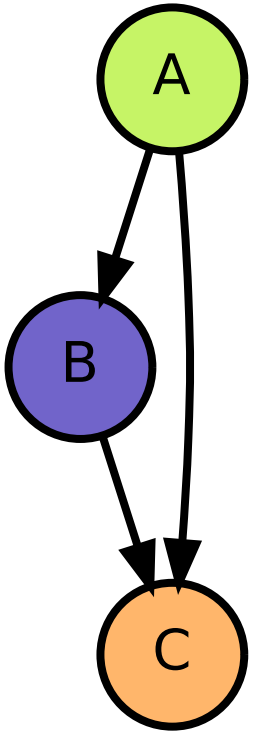



```
rule A:
    input: "{filler}.ext"
    output: "A/{filler}.ext"

rule B:
    input: "{filler}.ext"
    output: "B/{filler}.ext"

rule C:
    input: "{filler}.ext"
    output: "C/{filler}.ext"
```

```
snakemake C/B/A/sample1.ext C/B/A/sample2.ext
```

Easier development and benchmarking

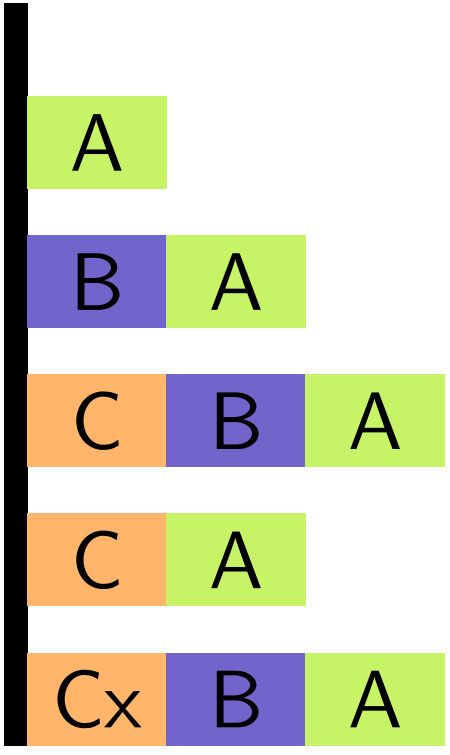


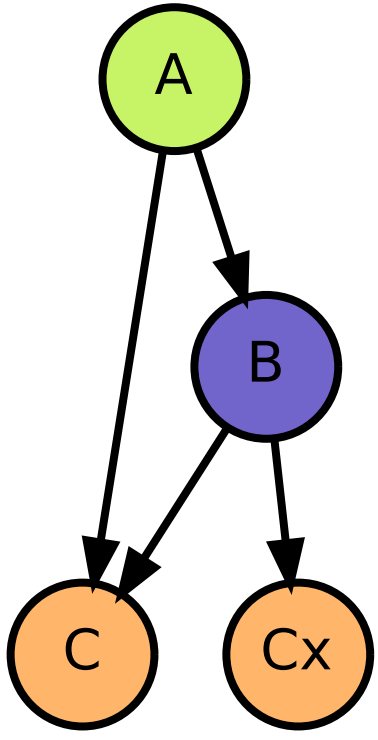



```
1  rule A:
2      input: "{filler}.ext"
3      output: "A/{filler}.ext"
4
5  rule B:
6      input: "{filler}.ext"
7      output: "B/{filler}.ext"
8
9  rule C:
10     input: "{filler}.ext"
11     output: "C/{filler}.ext"
```

```
snakemake C/B/A/example1.ext C/A/example1.ext
```

Easier development and benchmarking





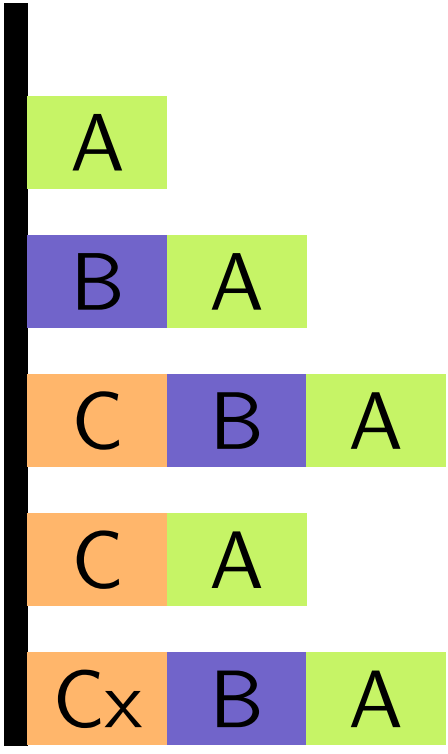

```
1  rule A:
2      input: "{filler}.ext"
3      output: "A{extra}/{filler}.ext"
4
5  rule B:
6      input: "{filler}.ext"
7      output: "B{extra}/{filler}.ext"
8
9  rule C:
10     input: "{filler}.ext"
11     output: "C{extra}/{filler}.ext"
```

`snakemake C/B/A/smp1.ext C/A/smp1.ext Cx/B/A/smp1.ext`

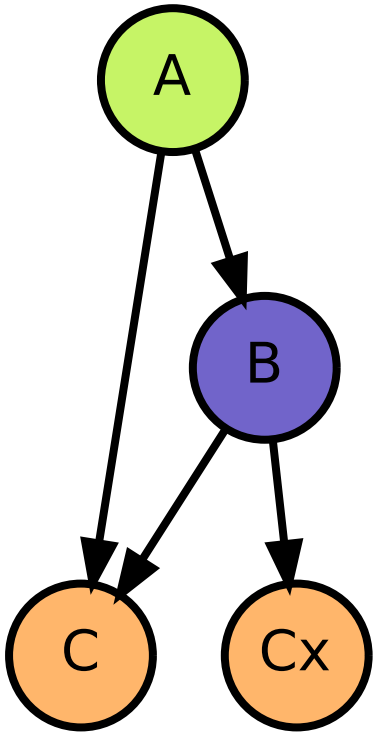
x can describe any combinations of parameters for each rule.

Traceable + Flexible + Small codebase

My thesis filetree, unique workflow DAG and codebase



500,000 files
≈ 60 to



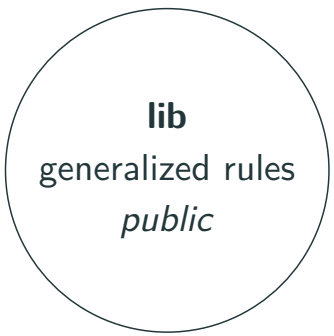
200,000 tasks
≈ 1 month/100 CPUs

```
rule A:
    input: "{filler}.ext"
    output: "A{extra}/{filler}.ext"

rule B:
    input: "{filler}.ext"
    output: "B{extra}/{filler}.ext"

rule C:
    input: "{filler}.ext"
    output: "C{extra}/{filler}.ext"
```

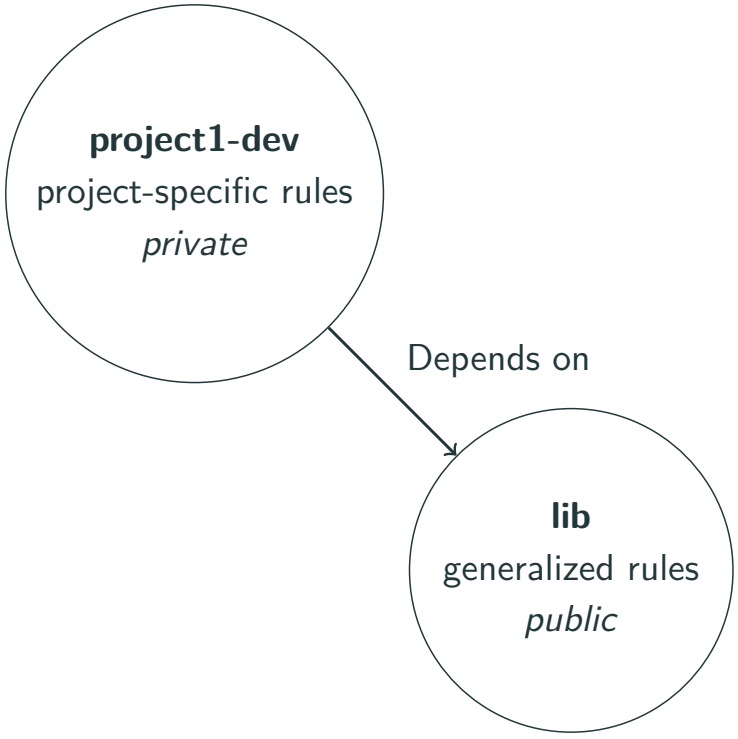
150 generalized rules
+ 100 specific rules

Remaining ≈ 3 mo codebase after `rm -rf out`
and ≈ 3 to of private input data.

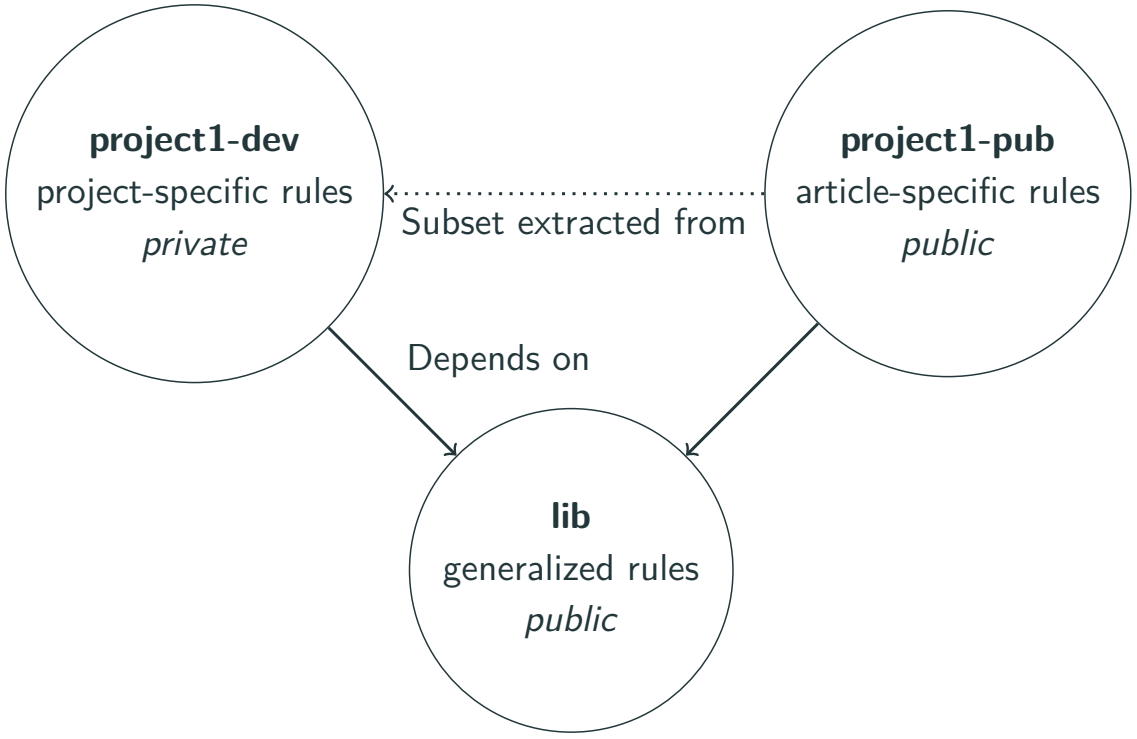
Metaworkflow defines coding conventions for modular workflow development, multiple collaborations and users



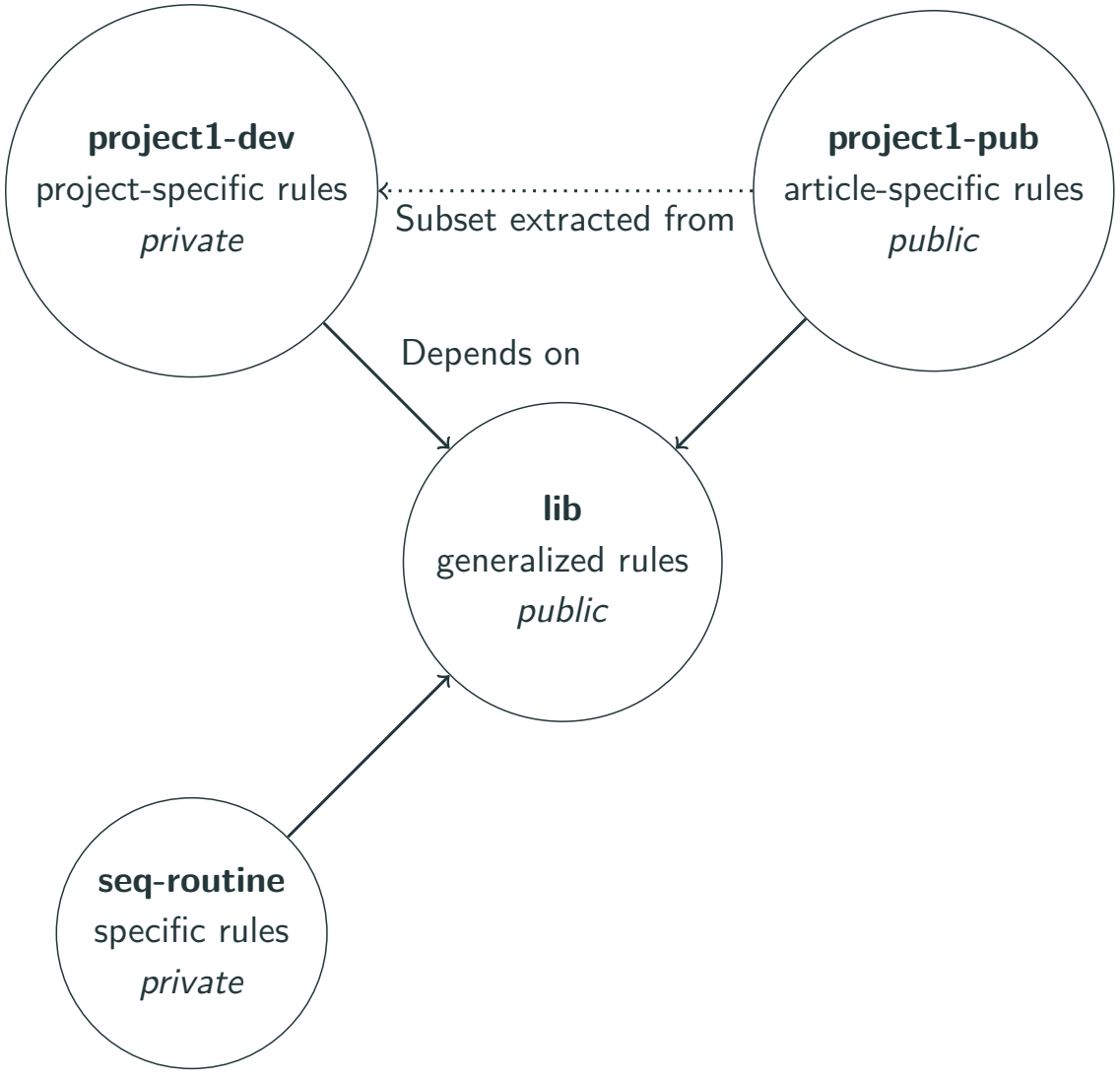
Metaworkflow defines coding conventions for modular workflow development, multiple collaborations and users



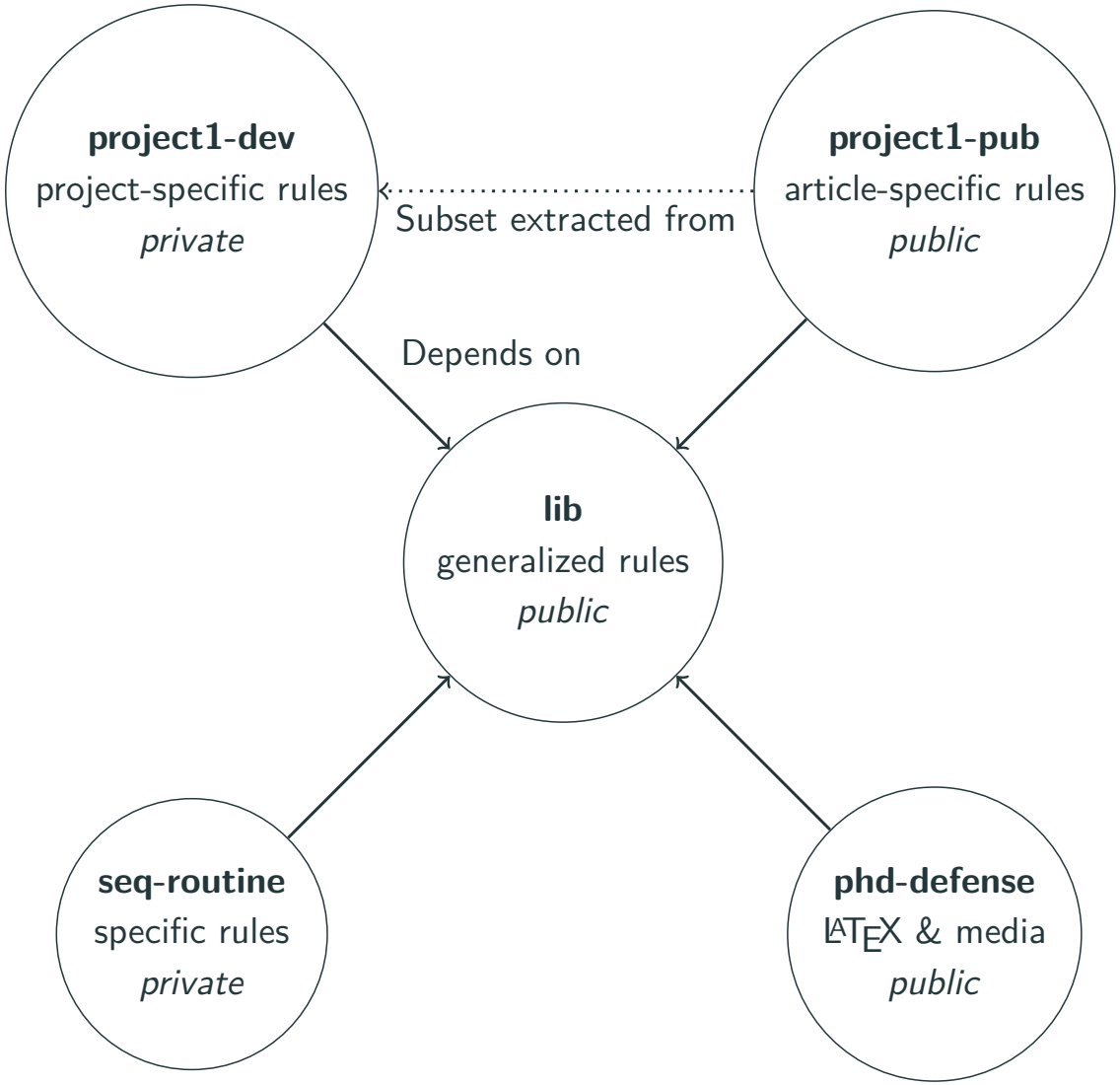
Metaworkflow defines coding conventions for modular workflow development, multiple collaborations and users



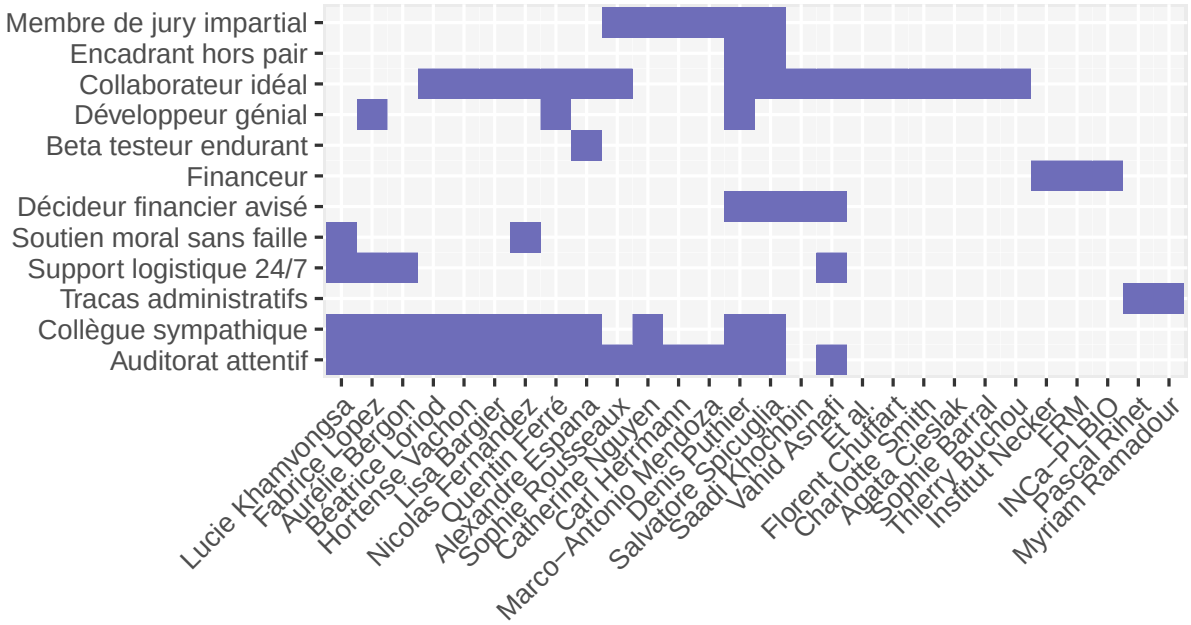
Metaworkflow defines coding conventions for modular workflow development, multiple collaborations and users

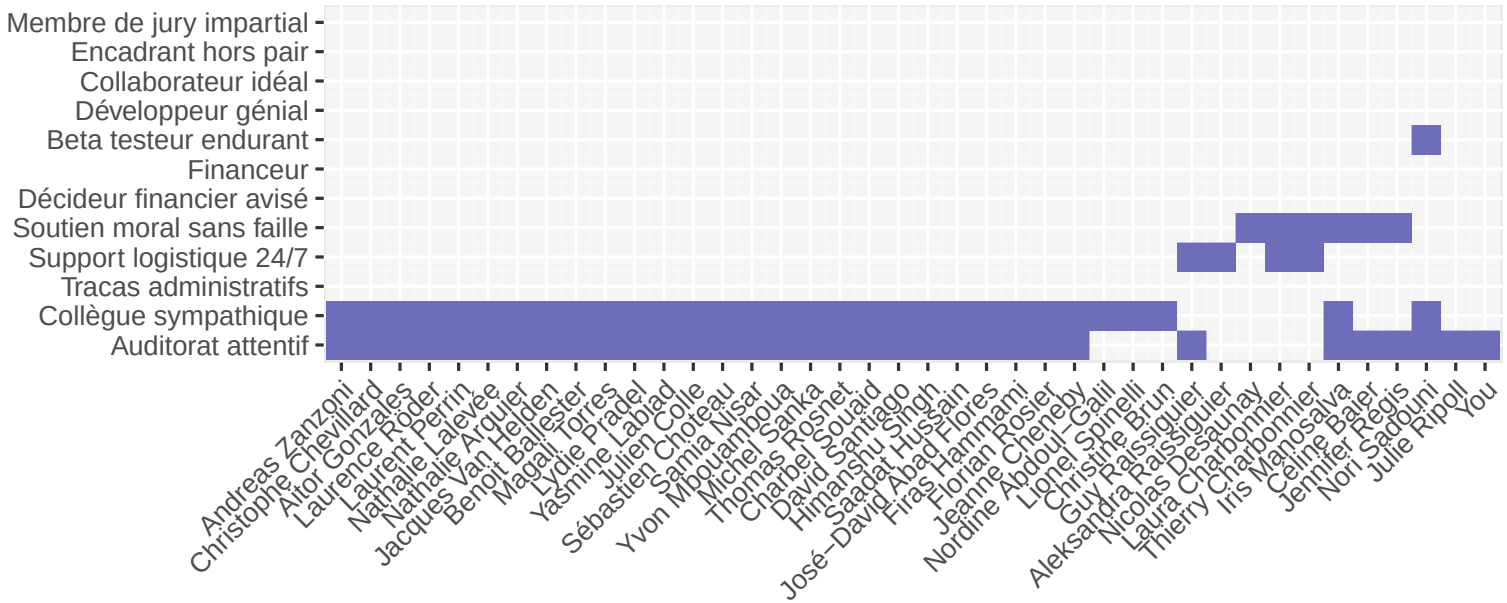


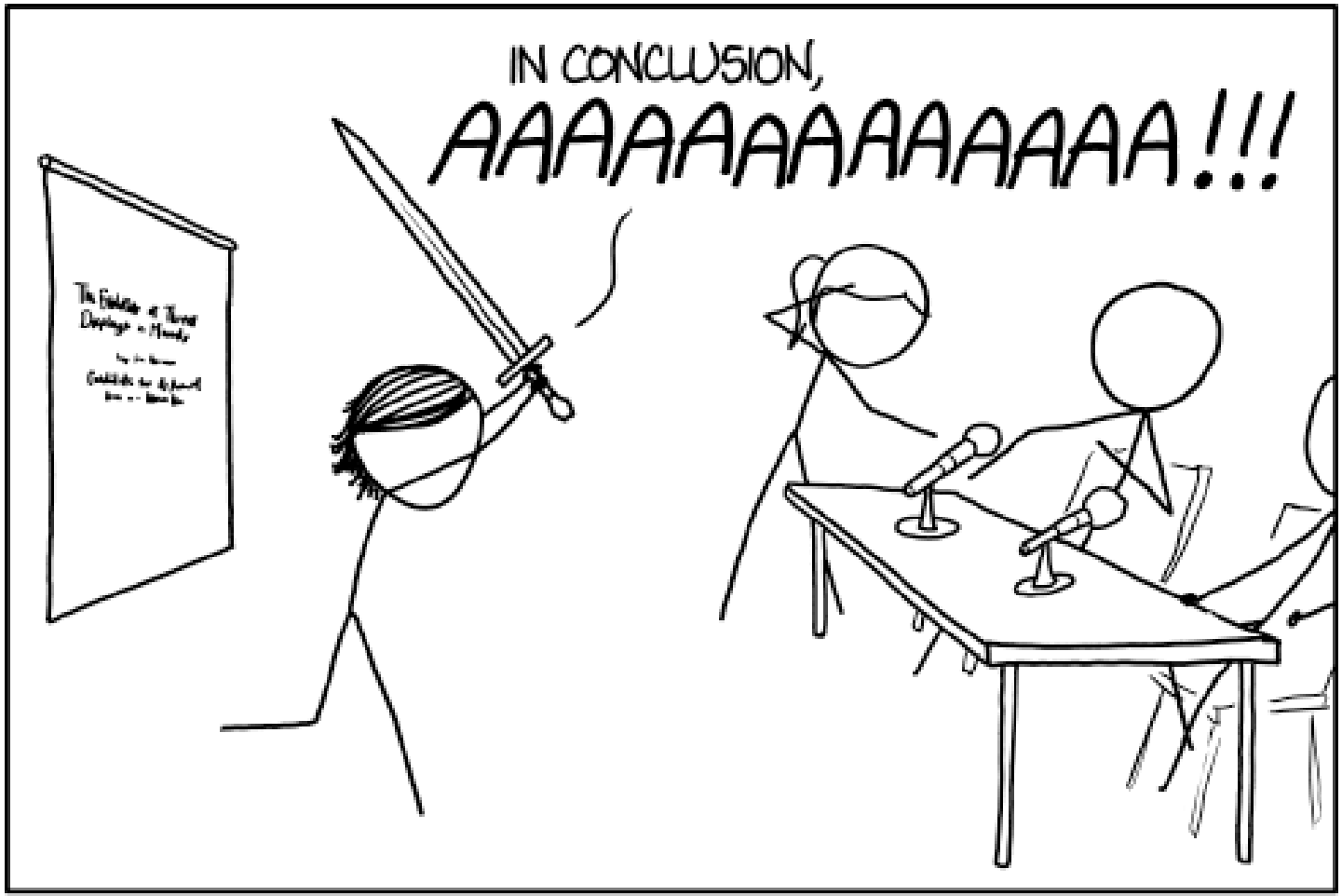
Metaworkflow defines coding conventions for modular workflow development, multiple collaborations and users



Acknowledgments






IN CONCLUSION,
AAAAAAAAAAAAA!!!
THE BEST THESIS DEFENSE IS A GOOD THESIS OFFENSE.

From xkcd