Transcription Regulation: From Sites to Cell-type Specificity

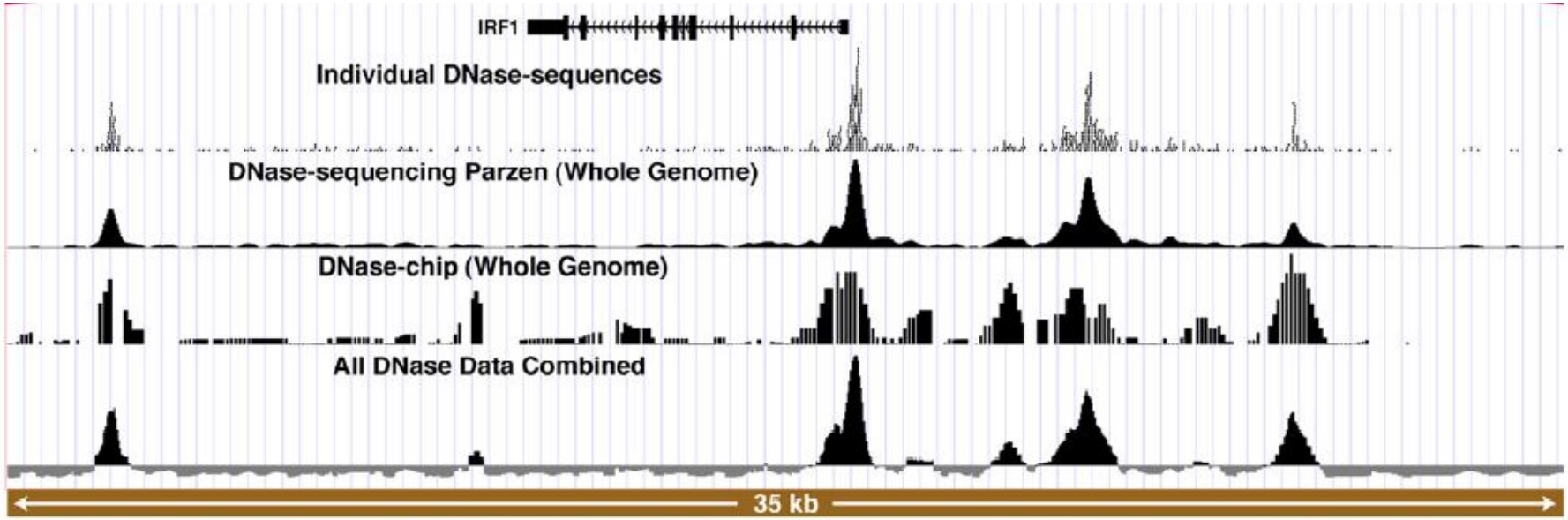
Uwe Ohler

Institute for Genome Sciences & Policy
Duke University

Berlin Institute for Medical Systems Biology
Max Delbrueck Centrum
Humboldt University

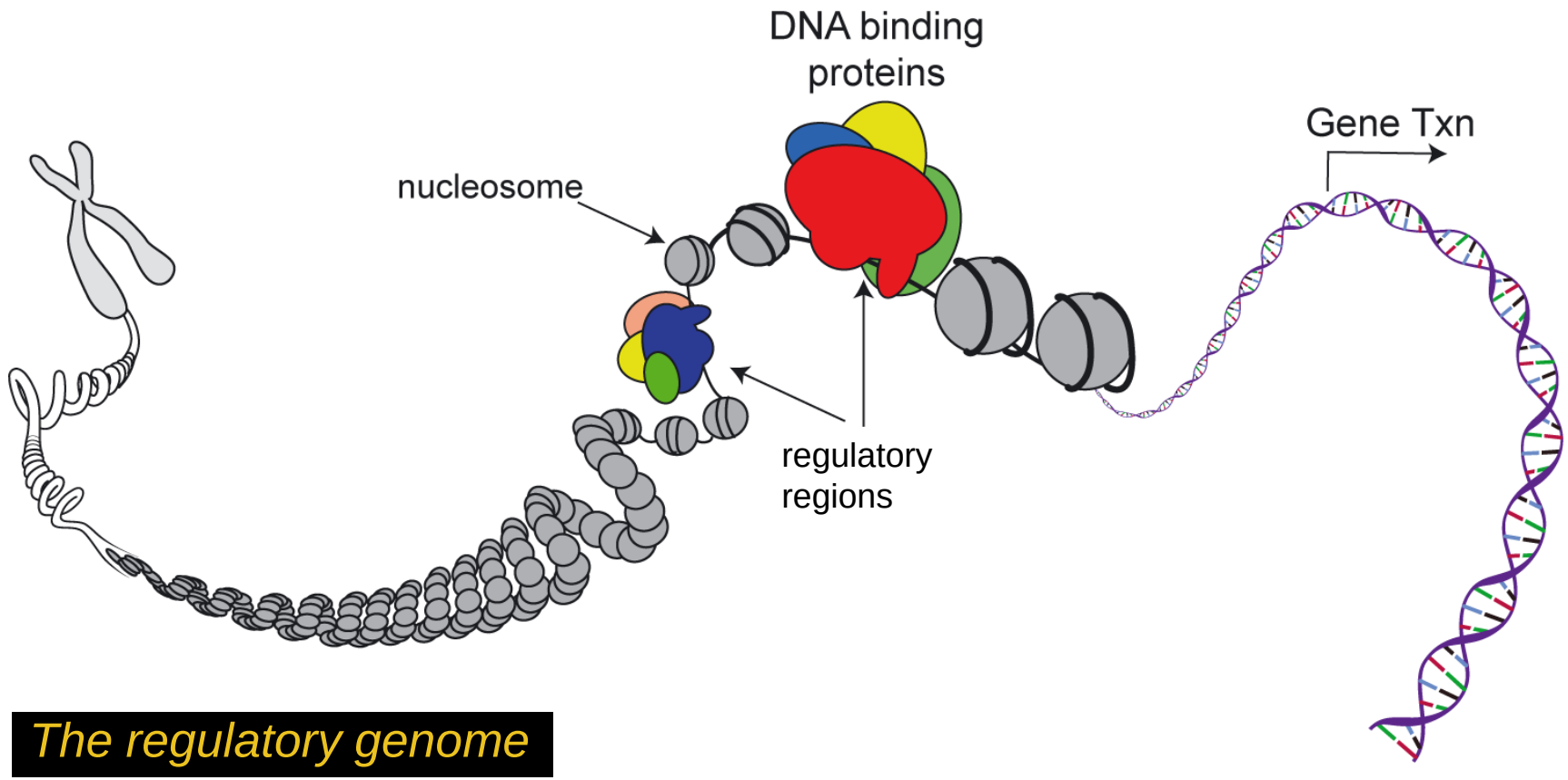
The impact of deep sequencing

- In the past 5 years, sequencing technologies have made the anticipated quantum leap
  - 1 mio base pairs currently for <10c
  - Illumina HiSeq: typically ~100 mio reads of ~100 bp
  - Everyone is able to generate more data than s/he needs
  - Unbiased exploration of genomes (DNA) & transcriptomes (RNA)



Boyle et al 2008

The ENCODE project



The regulatory genome

Variation & evolution (whole-genome seq)
Chromatin structure (histone modifications; “epigenetic code”)
Promoter & enhancers (ChIP-seq: motif finding; DNase-seq: CRMs)
[Transcript variants + RNPs (RNA-seq)]

Nature ; Genome Research; Genome Biology 2012

Computational Biology of Gene Regulation

Understanding transcription initiation
Ni et al, Nat Methods 2010
Rach et al, PLoS Genet 2011

Predictive models integrating chromatin, transcription, and RNA factors

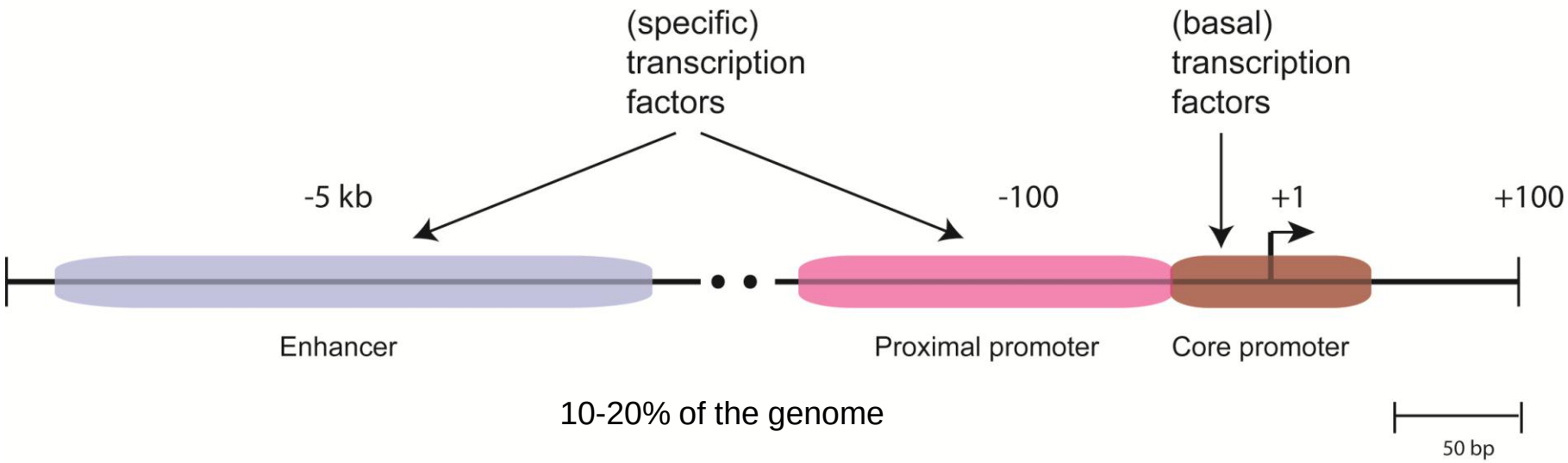
Identification of functional elements (DNA/RNA)
Georgiev et al, Genome Biol 2010
Mukherjee et al, Mol Cell 2011

Models for cis-regulatory sequences & evolution
Majoros & Ohler, PLoS Comp Biol 2010
Natarajan et al, Genome Res 2012

Regulatory networks in differentiation & development

Analyzing gene expression (images)
Pruteanu et al, PLoS Comp Biol 2011
Busch et al, Nat Methods 2012

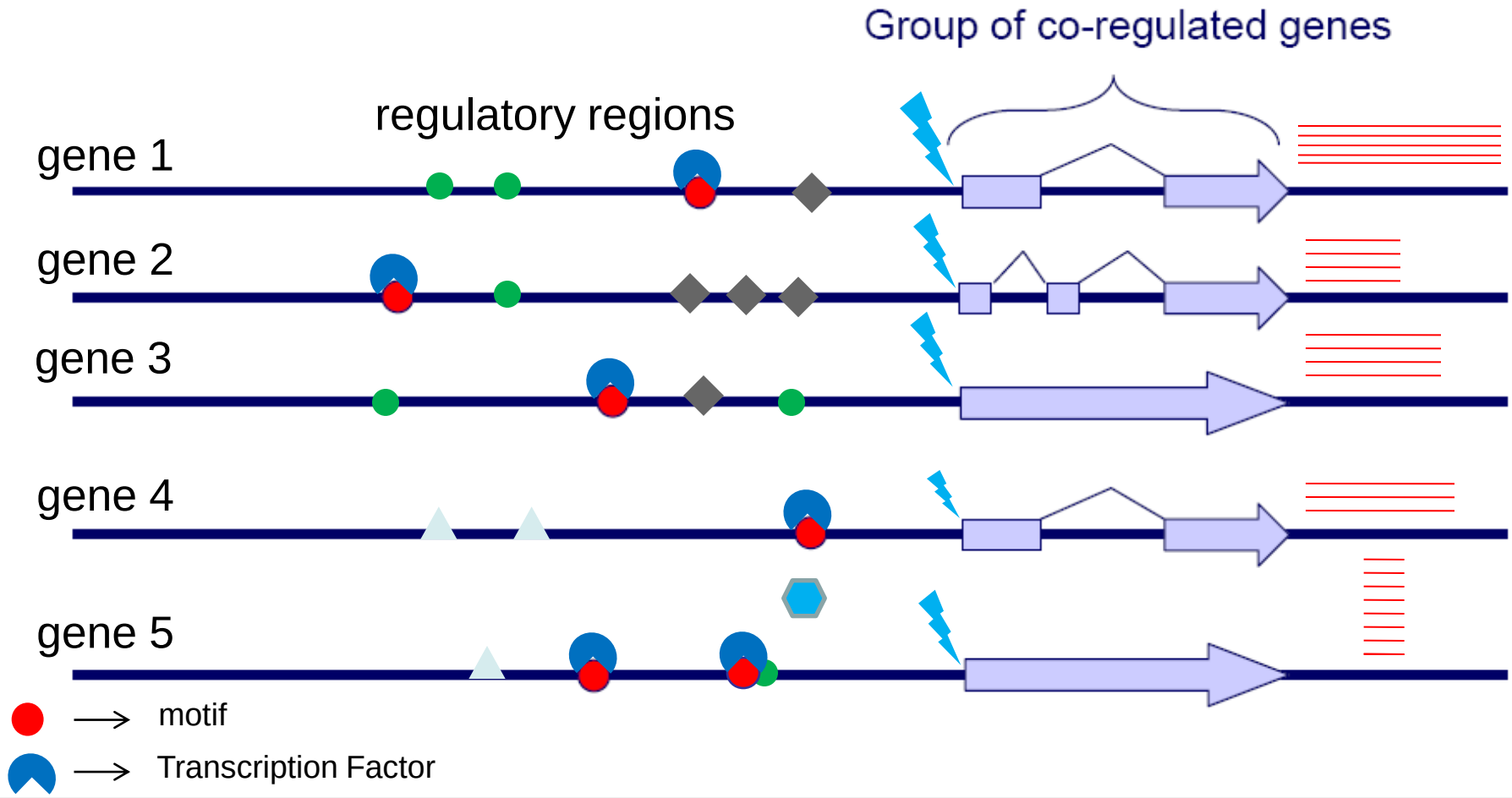
A schematic view of cis-regulatory DNA regions



- General information: the core promoter
  - Region around the transcription start site (TSS) where RNA polymerase II (pol-II) interacts with basal transcription factors
  - Potentially far away from the *translation* start site
- Specific information about functional context of genes: proximal promoter/enhancers
  - Binding sites of specific transcription factors confer activation at the right developmental stage or tissue

Ohler and Wassarman, *Development* 2010

Regulation of transcription



Motif representations

non-degenerate motif (**AATGTCT**)

```
seq 1  AGGTGTGGTTGTAAATGTGTTAAGTGTTGAATGTCTGAAAATGTGTGTGAAAAAATGTGTG
seq 2  AAGTGTGTAATGTCTTGTGTGTAAAAACCGTGTGTGAAACCCTTCAATTGTGTGCACACGT
seq 3  AAATGTGGTCCCCGGTGTGTGAATTGGTTAACCTCTAATGTCTGTAACCAAGTGTGTAATG
seq 4  AGGTGTGATGATGCTGTAGATGCTCGTASGTAATGTCTGGGCTTTTAATTCCCTTACGTCG
seq 5  GTGGCTATGTGGTCAATGTCTCACTGGCGTCTTAGTTGGCTAGTAGCTCTCTGATGATGAT
```

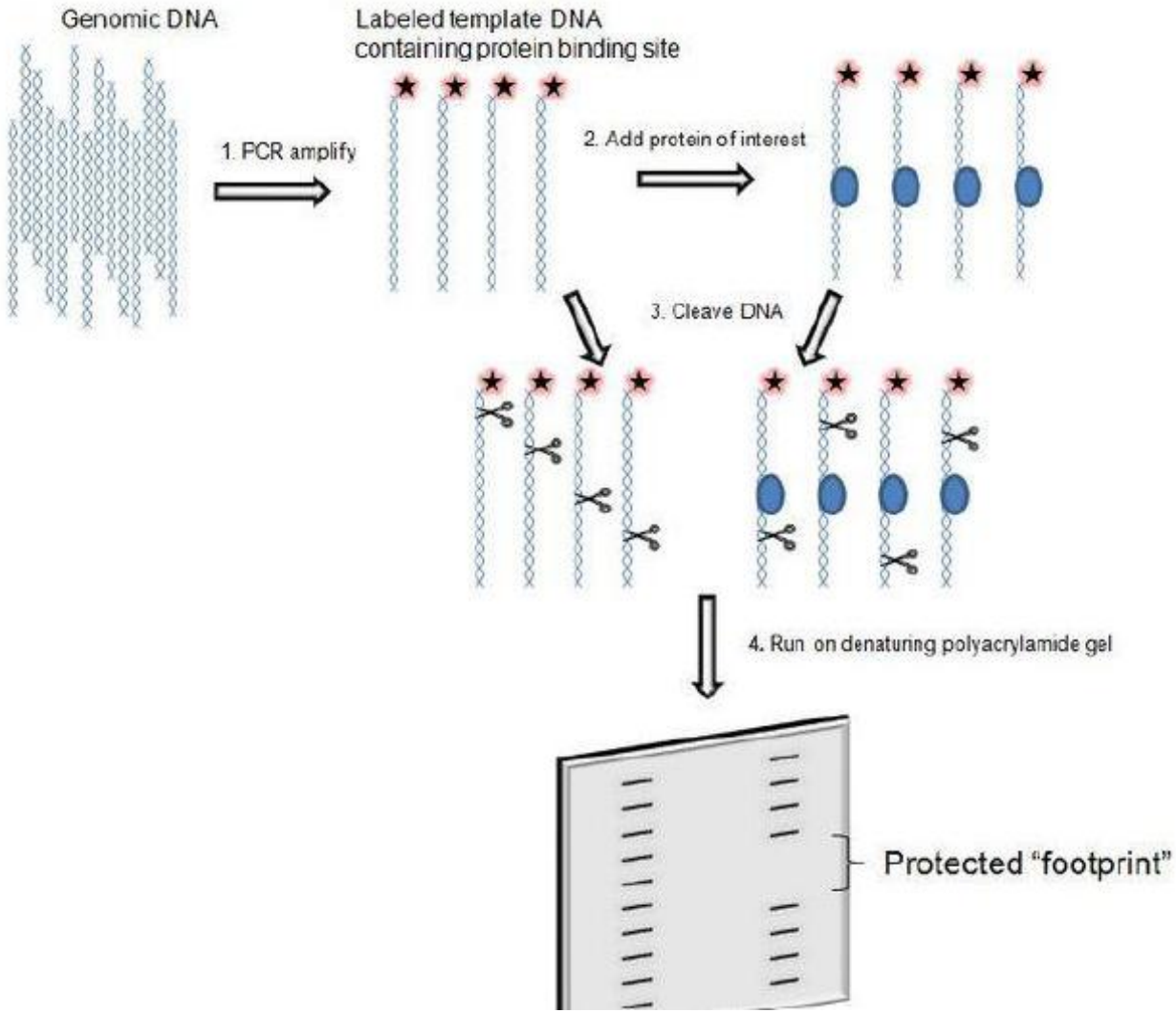
more realistic (degenerate motif: **WATGTNT**)

```
seq 1  AGGTGTGGTTGTAAATGTGTTAAGTGTTGAATGTCTGAAAATGTGTGTGAAAAAATGTGTG
seq 2  AAGTGTGTTATGTCTTGTGTGTAAAAACCGTGTGTGAAACCCTTCAATTGTGTGCACACGT
seq 3  AAATGTGGTCCCCGGTGTGTGAATTGGTTAACCTCTTATGTGTGTAACCAAGTGTGTAATG
seq 4  AGGTGTGATGATGCTGTAGATGCTCGTASGTAATGTATGGGCTTTTAATTCCCTTACGTCG
seq 5  GTGGCTATGTGGTCAATGTTTCACTGGCGTCTTAGTTGGCTAGTAGCTCTCTGATGATGAT
```

Extended alphabet: **A**, **G**, **C**, **T**, **M** {A,C}, **S** {G,C}, **R** {A,G}, **W** {A,T}, **Y** {C,T}, **K** {G,T}, **N** {A,G,C,T}

Still more realistic (position weight matrix)

Experimental mapping: "Footprinting"



Common problems/tasks

1. Sites: Motif finding – what do target sequences look like?
   - Given a set of "foreground" promoter sequences, identify a model/description of one or more enriched binding sites (and their location in the sequences)
   - This is a search problem: Find the model/description that maximizes a score reflecting the over-representation of hits in the data
     - Indirect evidence: clusters of co-regulated genes
     - Direct evidence: Binding of regulatory factors

2. Regions: Enhancer codes – what defines specificity?
   - Given known/predicted binding sites for a large set of TFs, find a specific combination that encodes an expression pattern
   - Goal: e.g. model for tissue-specific promoters

The challenges

- Can we predict TF binding from sequence?
- Can we predict expression patterns from binding?
- Can we predict phenotype from expression?

Information:

- Chromatin state, DNA sequence
- Direct binding (ChIP) and/or models to predict TF binding
- Expression (Pol-II recruitment? Production? Steady state?)

Problems:

- large intergenic space;
- many degenerate sites;
- noisy assays

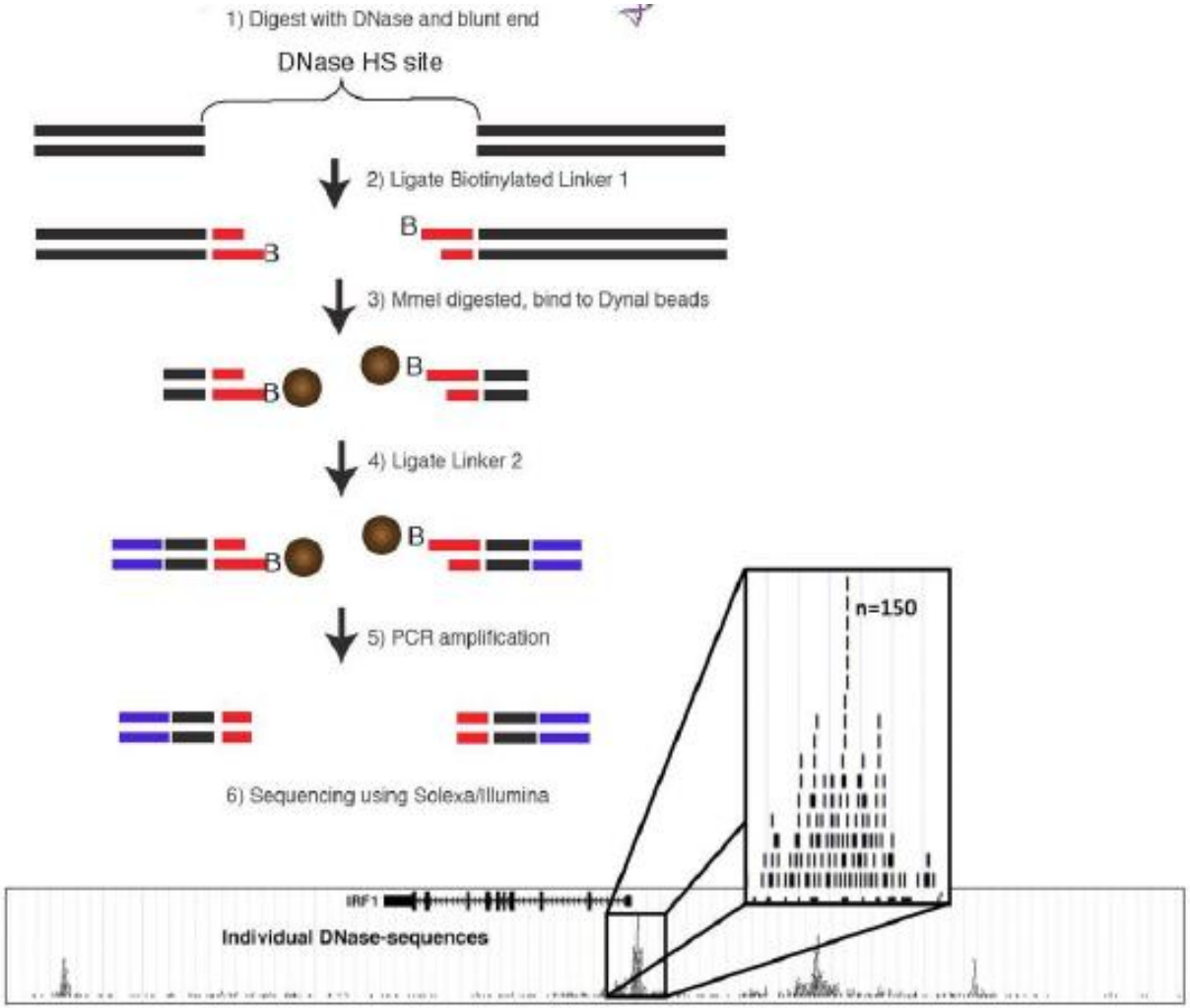
The cis-regulatory code

- How is tissue specific gene expression defined?
  - Expression of regulatory factors
  - Activity of the factors (e.g. nuclear localization)
  - Availability of target DNA sites

- Problem: direct binding studies (or predictions) show thousands of (potential) target sites

- Hypothesis:
  - The binding of several TFs within a "regulatory module" leads to the specificity needed
  - These factors need to bind to accessible regions

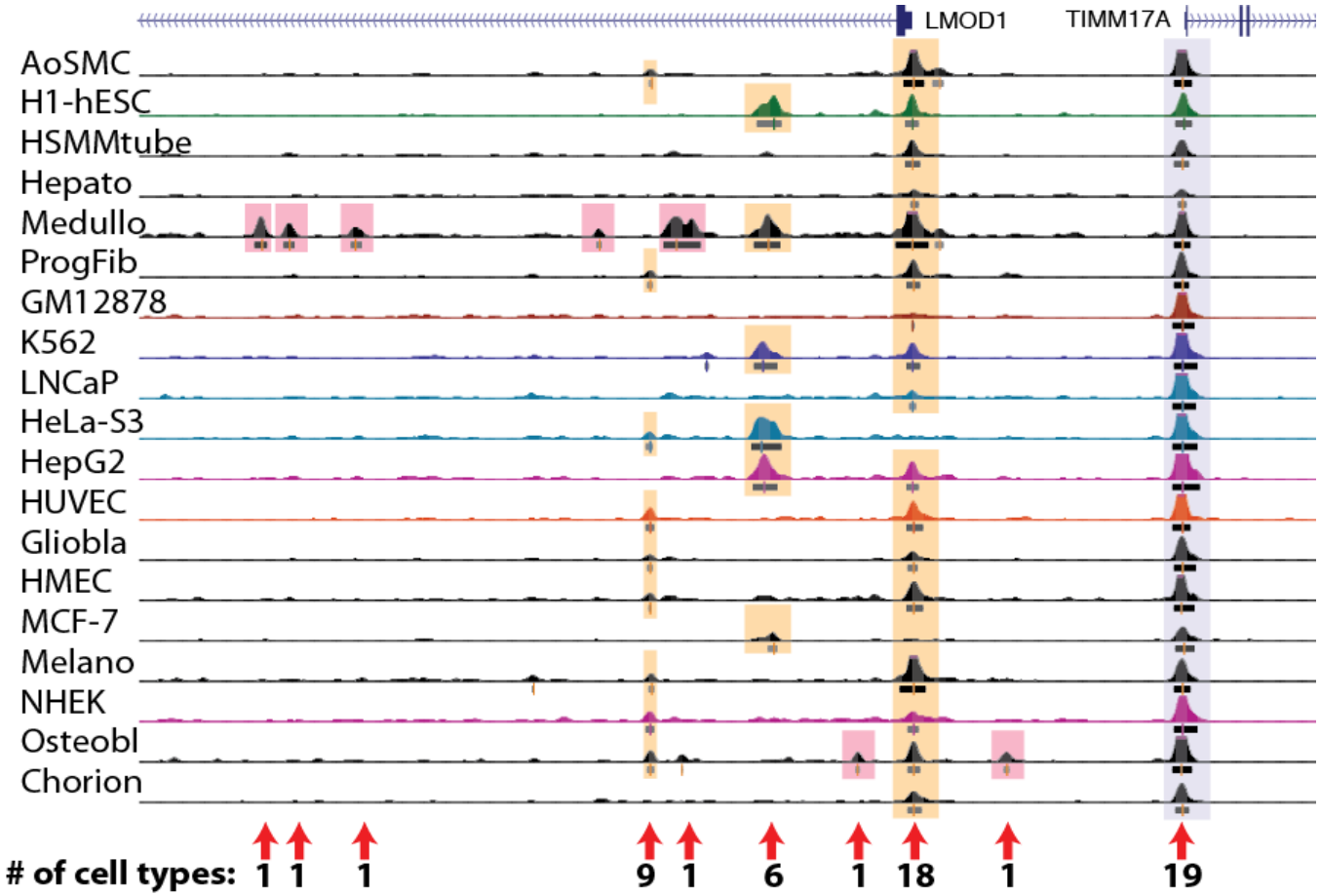
Mapping of open chromatin



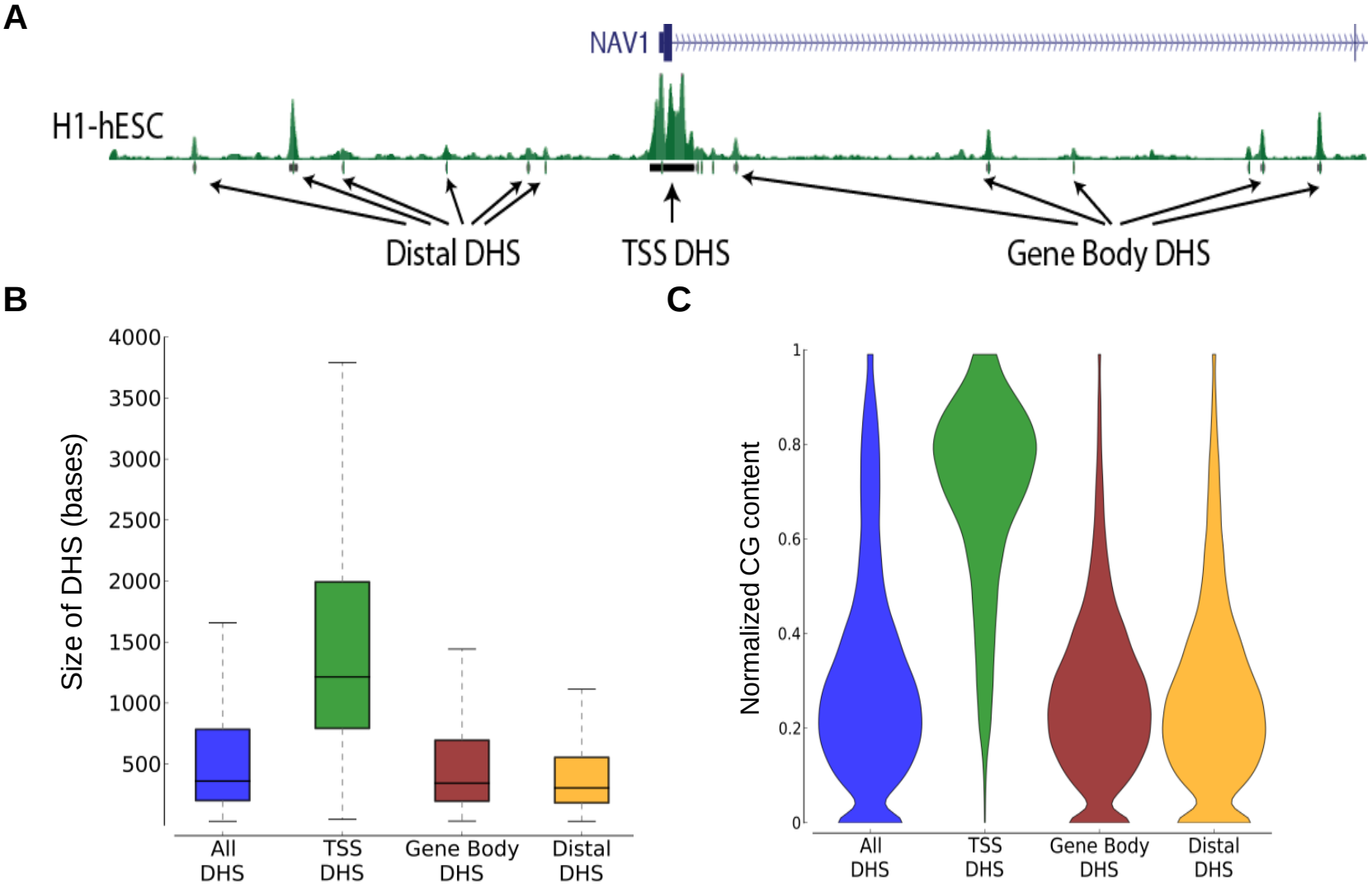
After initial alignment, tracks are smoothed &
variable size DNase hypersensitive sites (DHS) above cutoff are extracted

The setup

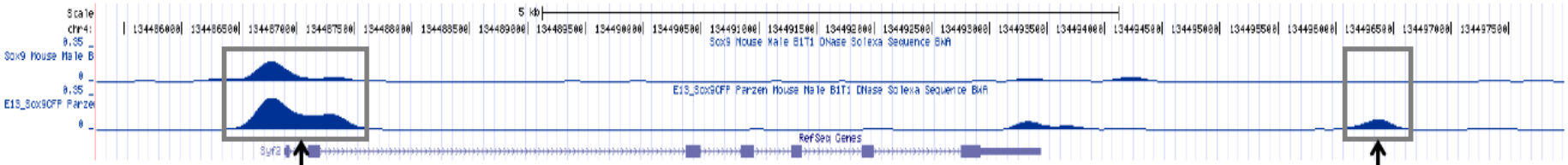
- Duke DNase-seq conducted on >50 human cell lines
  - Identification of DNaseI hypersensitive sites (DHS)
  - Score associated with each site



DHS across 19 least related cell lines

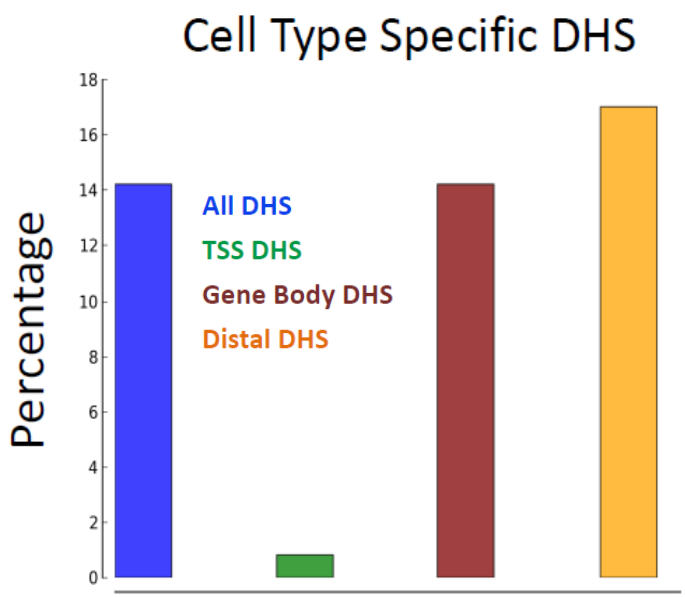


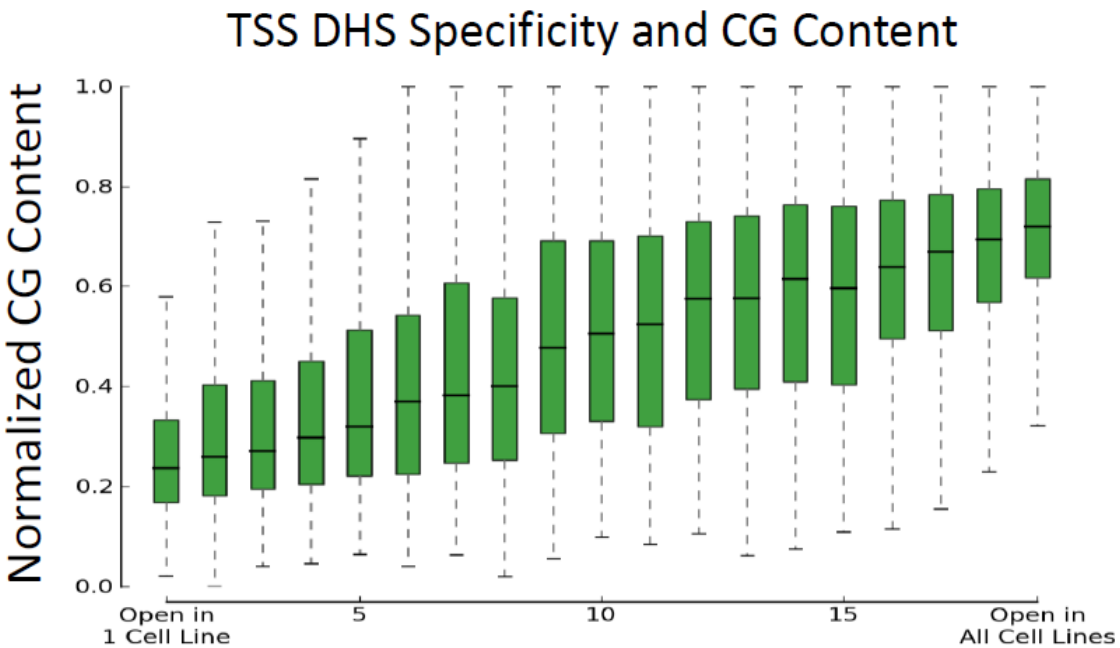
Specificity of open chromatin



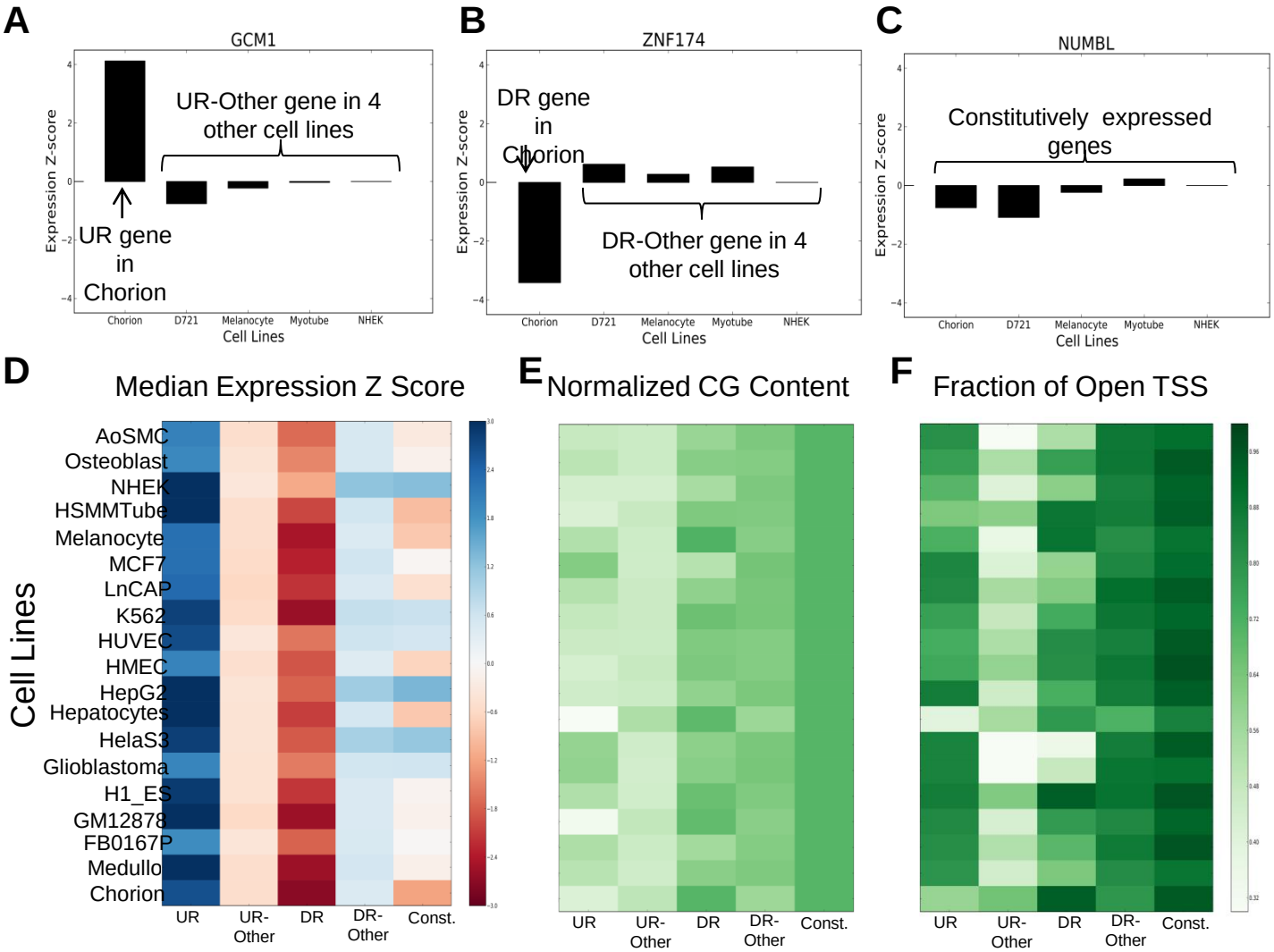
Shared DHS

Cell type Specific DHS



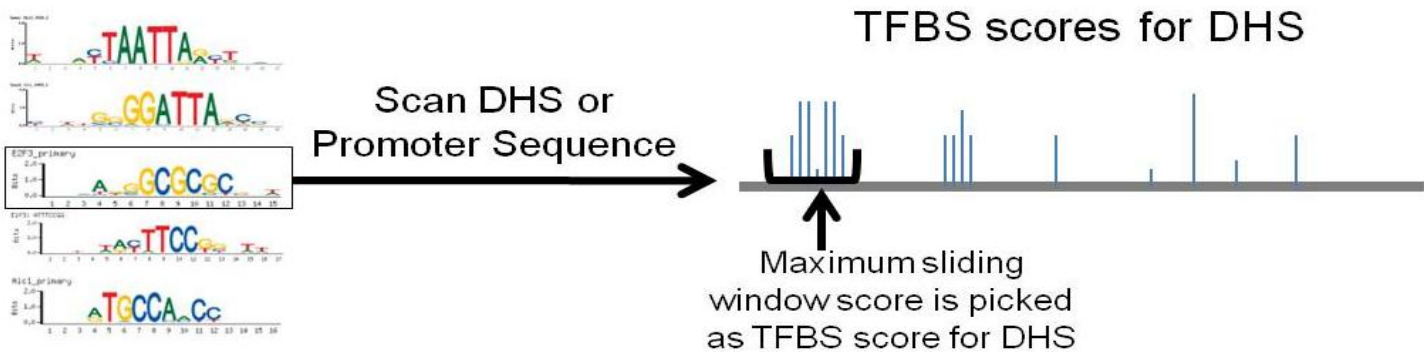


Defining expression patterns



Most up- and downregulated genes from each cell type, plus one set of constitutive genes (200 genes each)

Models to distinguish expression patterns

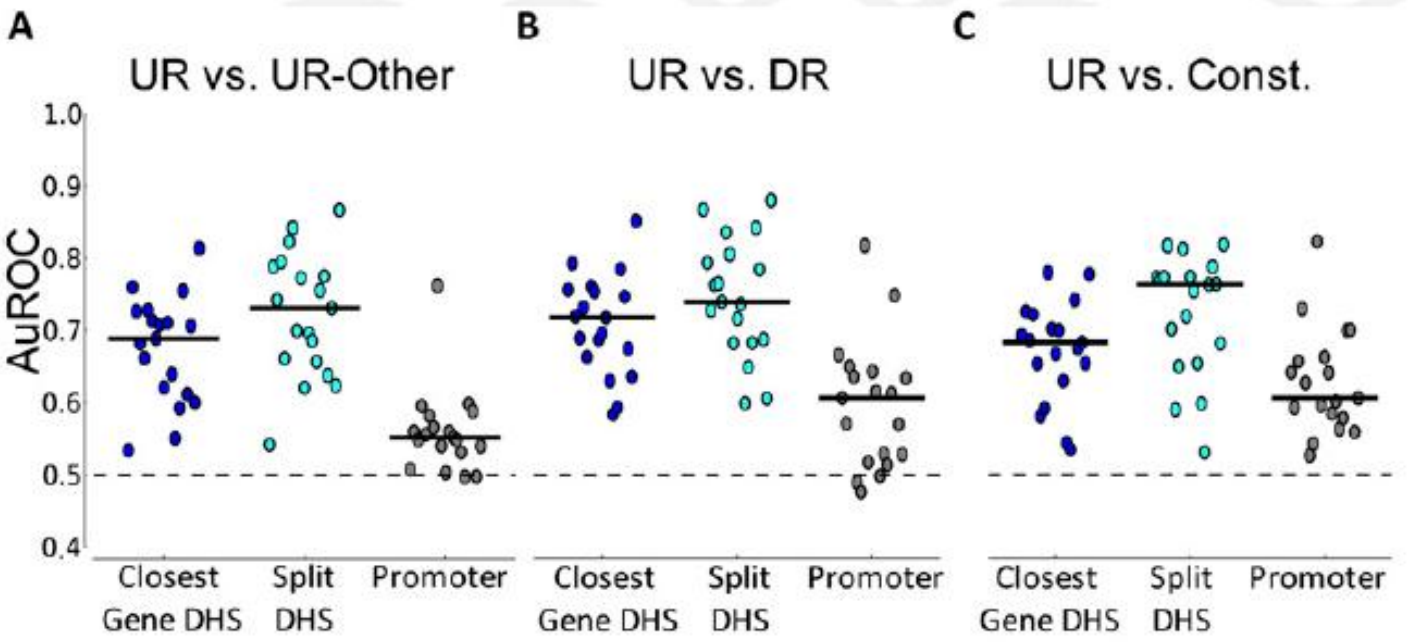


Gene	Class Label				• • •	
1	Up-regulated	2.1	0.5	3.1		0.8
2	Down-regulated	3.2	2.5	1.2		1.6

Classifiers are trained to distinguish expression patterns based on TF features for > 300 factors

[sparse linear L1-logistic regression, “Lasso”]

Classifier Performance



Medulloblastoma

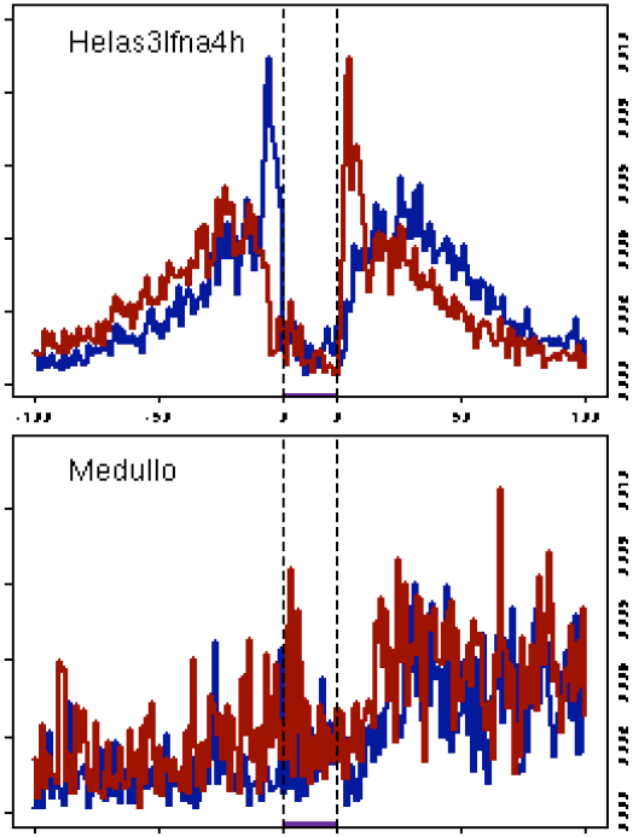
CRX
NRSF
SOX11
NR2F2

Up-regulated genes

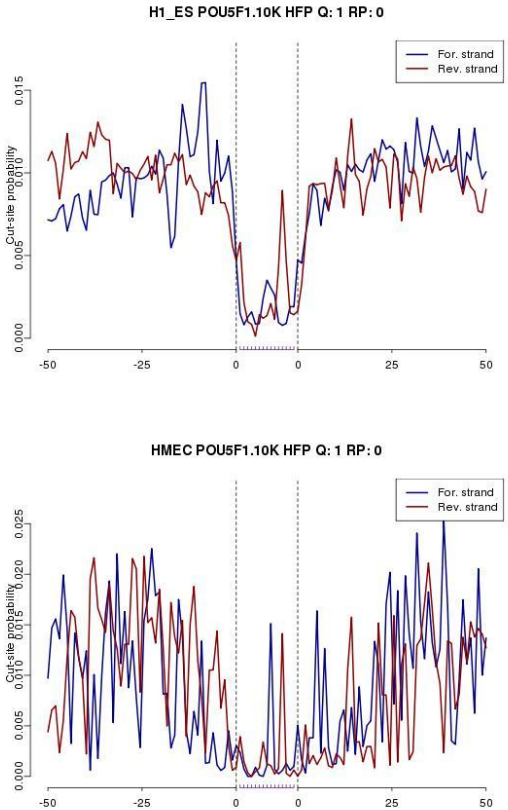
- Compare this approach to using proximal promoter region
- Also identifies activators and repressors

Looking inside the models

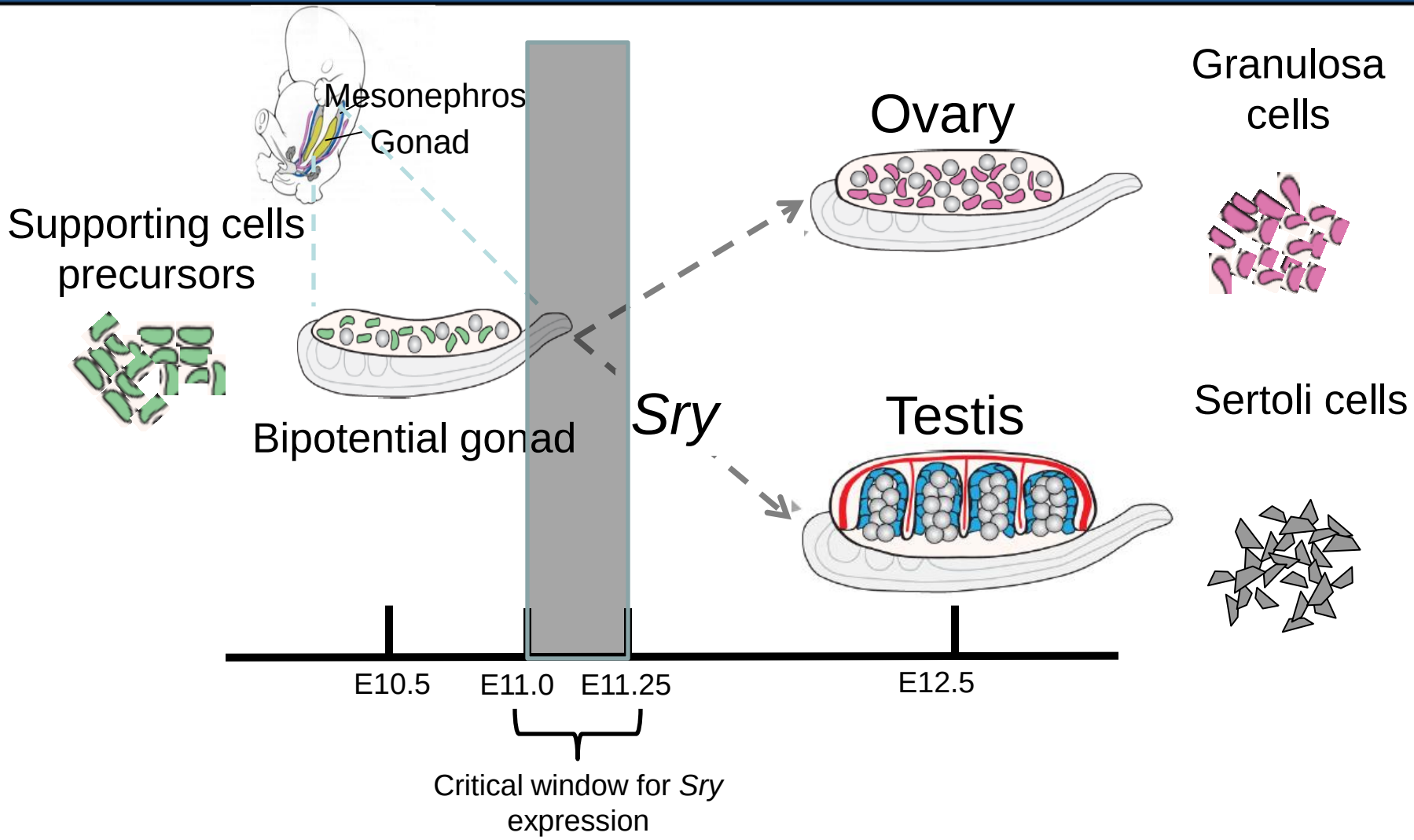
- NRSF/REST: Top-scoring factor for Medulloblastoma cell line genes
  - *repressor* not bound in this line



- Oct4: Among the top factors of ES cell genes
  - Stem cell specific *activator*

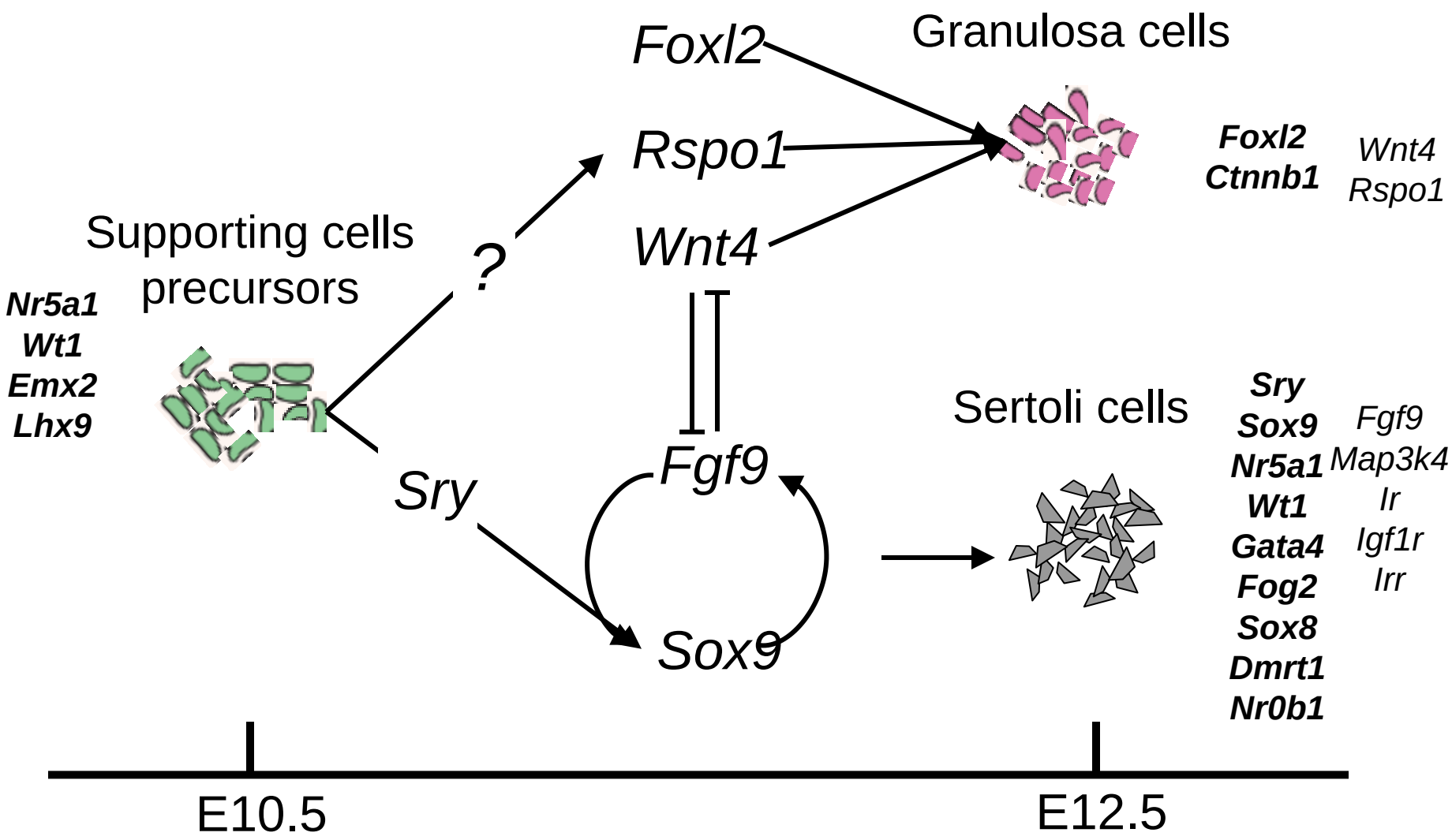


Example application: sex determination



- *Sry* is transiently up-regulated

The Network in Supporting Cells

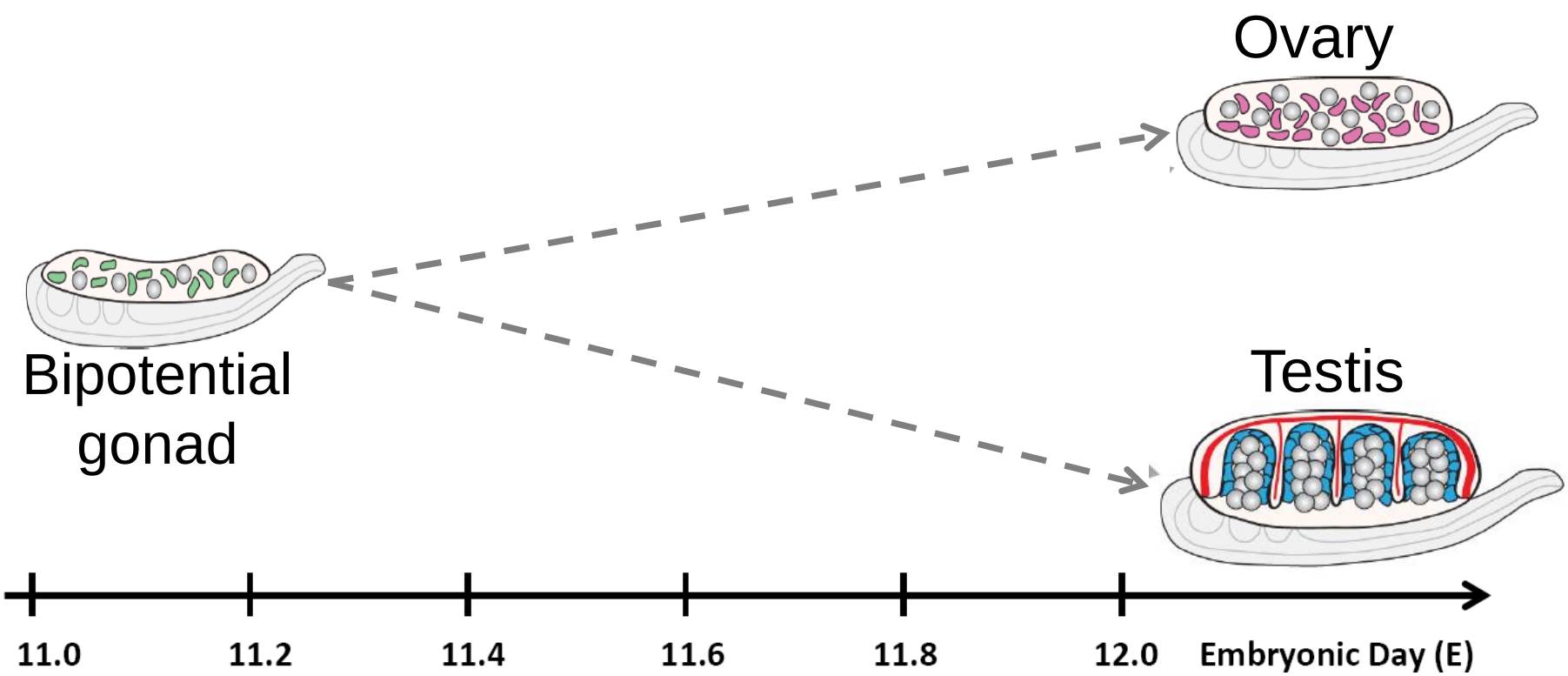


Building a TF network in Supporting Cells

- Characterize the transcriptome as the bipotential gonad differentiates
- Identify enhancers genome-wide in XX and XY supporting cells and use computational approaches to build a predictive models of gene expression
- Test and validate predictions using RNAi and ChIP-seq

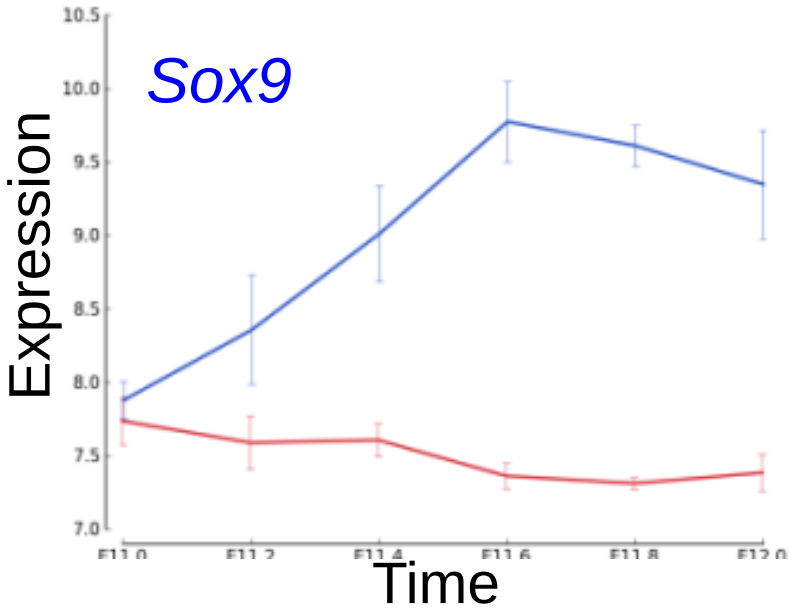
Chen et al. 2008, Ong et al. 2011

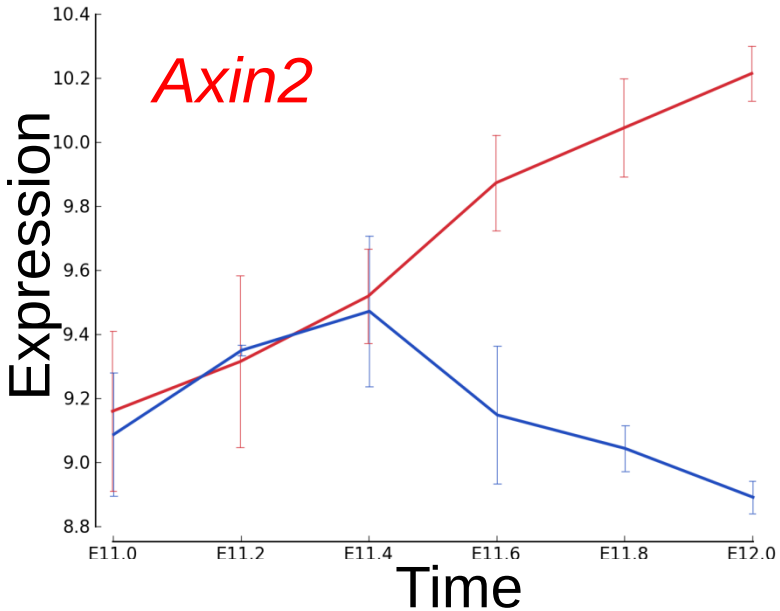
A fine-time series transcriptome



- Whole gonads
- 2 sexes: XX and XY
- 6 time points
- 2 strains
- 3 replicates per sex, strain and stage

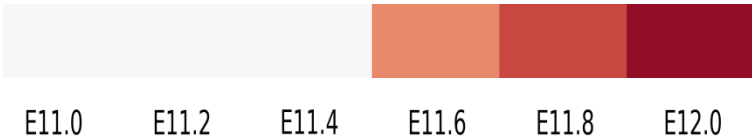
Cascades of dimorphic expression





Fold Difference

Fold Difference



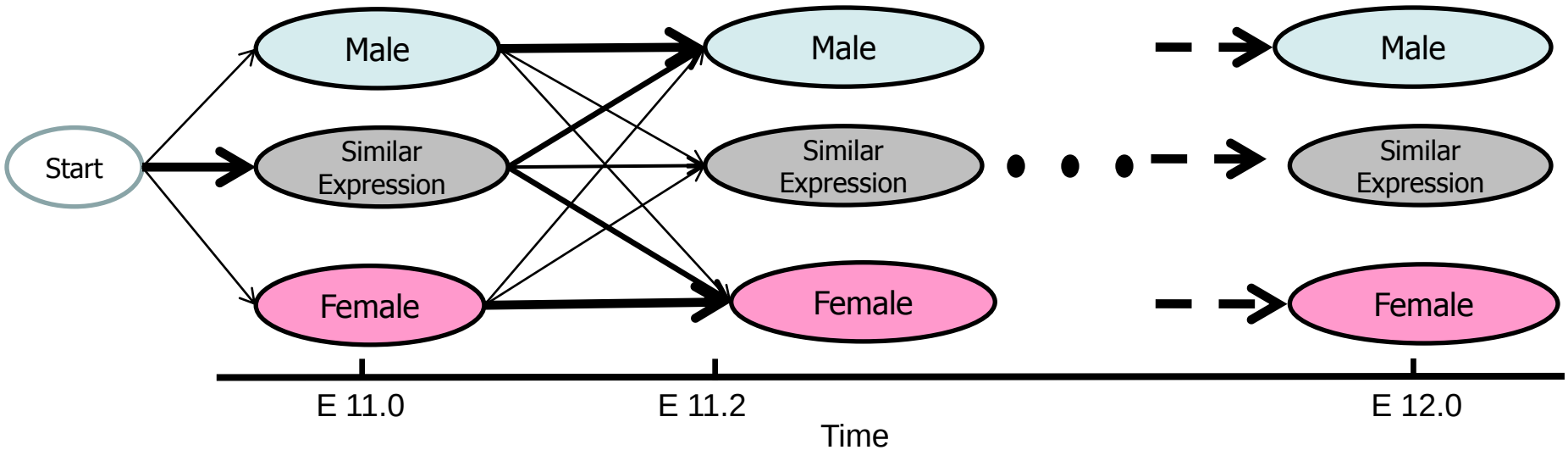
Male enriched

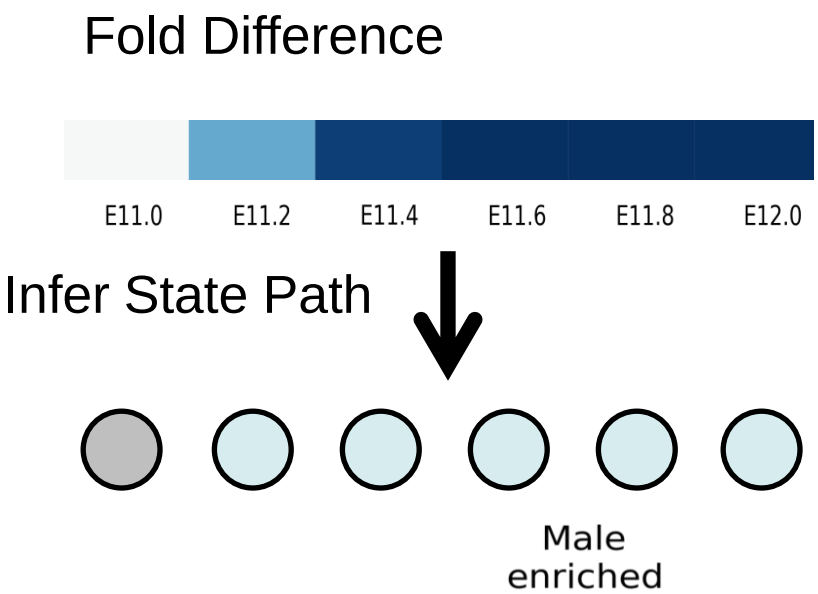
Female enriched

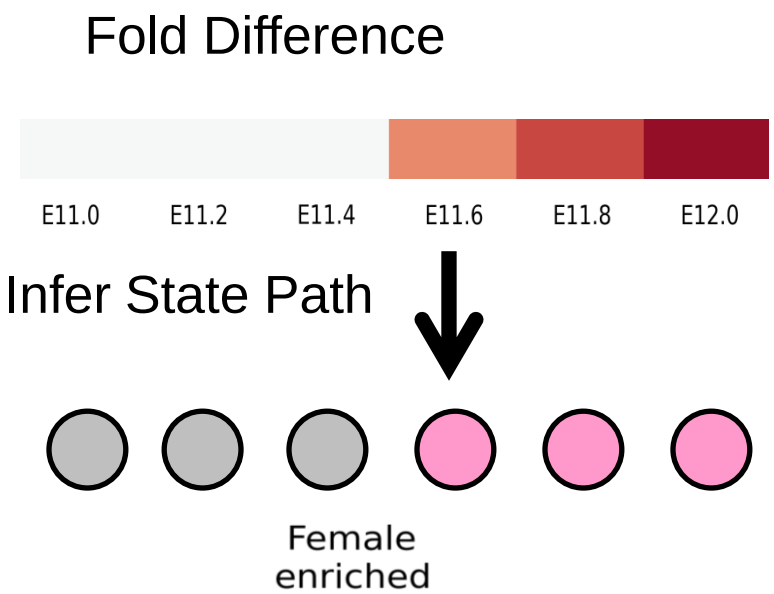
XX

XY

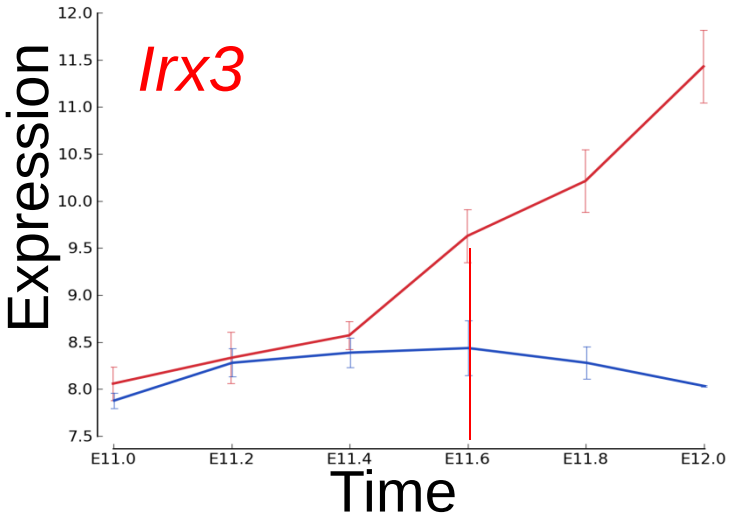
A Hidden Markov Model to identify dimorphic expression

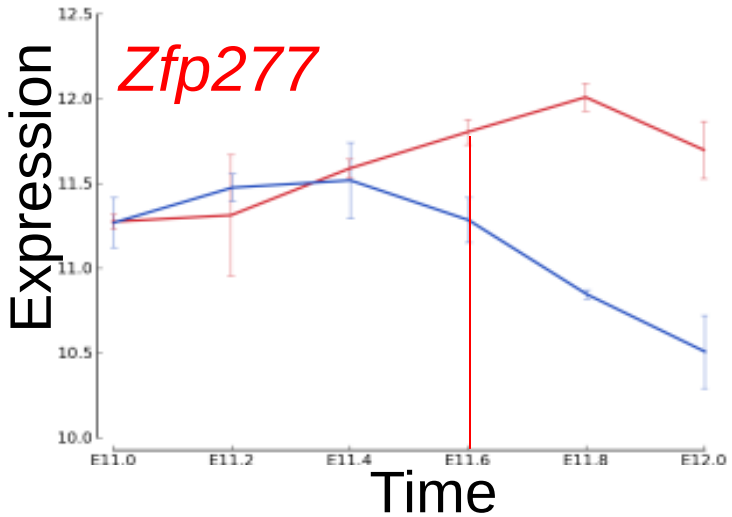


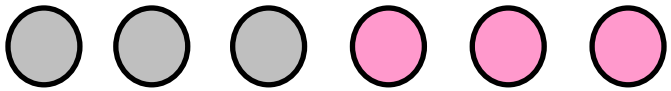


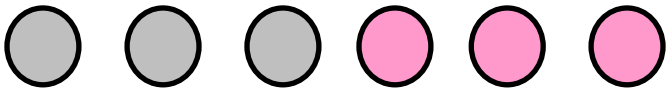


Different paths to dimorphic expression









Up-regulation in XX gonads

Down-regulation in XY gonads

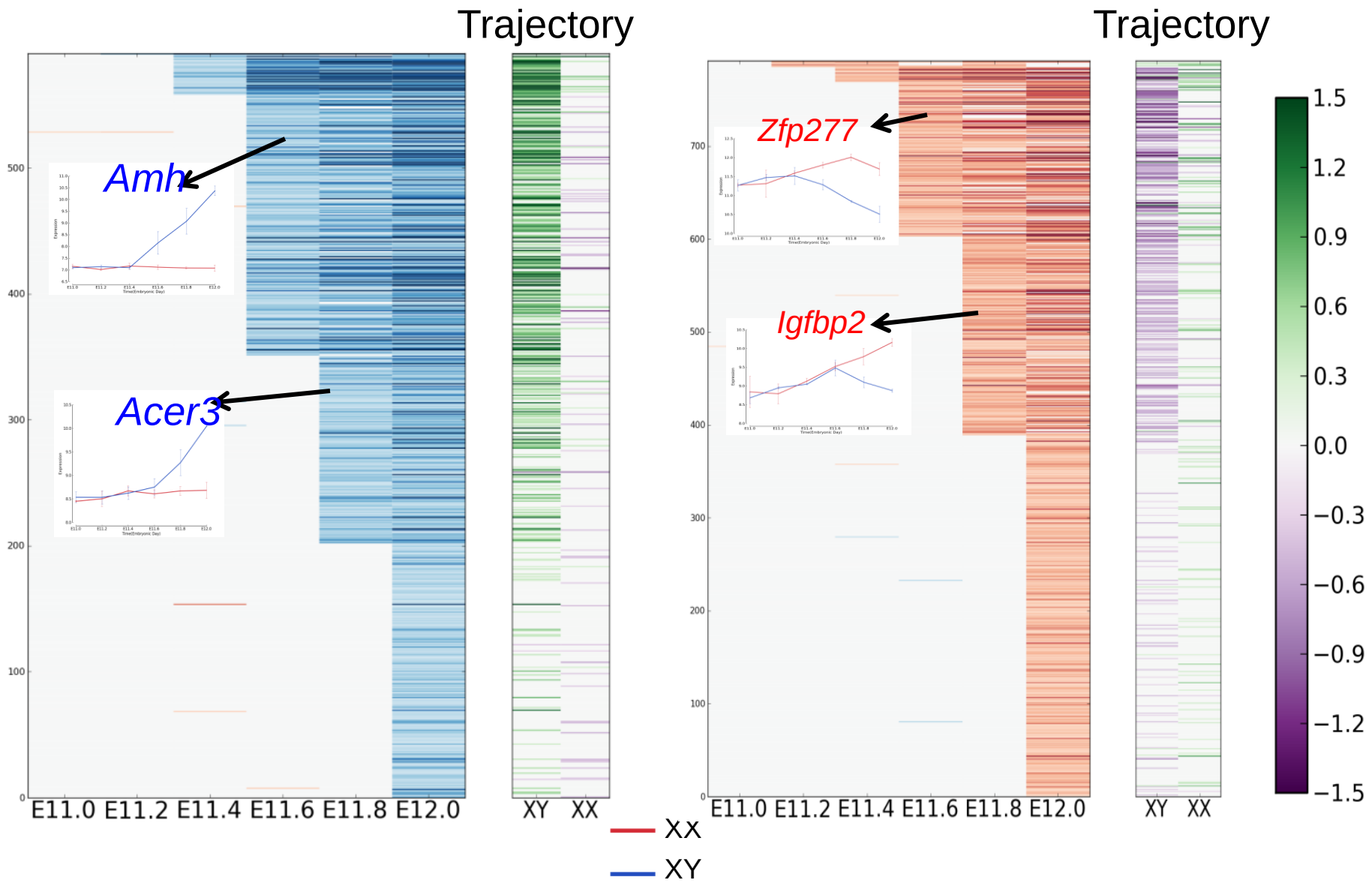
Identify activator in XX gonads

Identify repressor in XY gonads

XX

XY

Activation and Repression in the Gonad



New candidate regulators

Granulosa cells

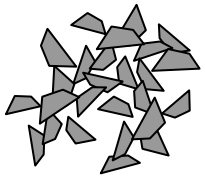
Foxl2
Ctnnb1

Wnt4
Rspo1

Gtf2a1
Tcea3
Sp5
Lef1
Vgll2
Runx1
Zfp277
Lhx9
Msx1
Irx3

Supporting cells pre-cursors

Sertoli cells



Sry
Sox9
Nr5a1
Wt1
Gata4
Fog2
Sox8
Dmrt1
Nr0b1

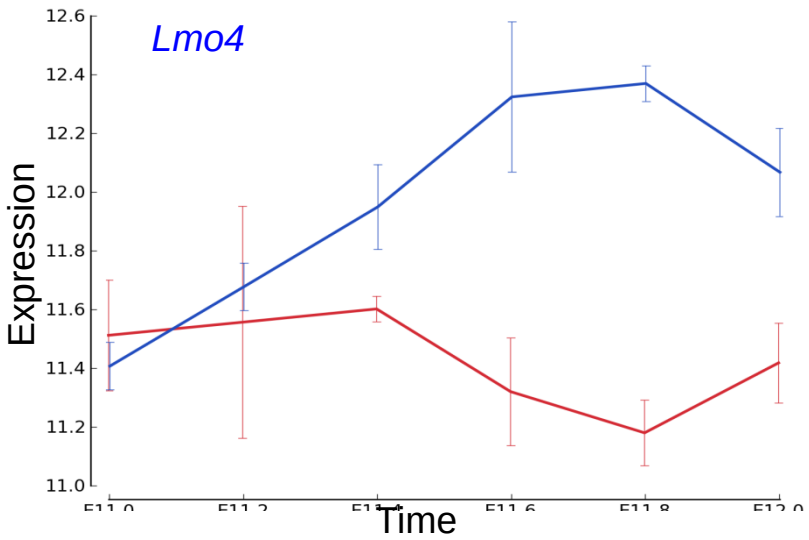
Fgf9
Map3k4
Ir
Igf1r
Irr

Cited4
Sox13
Scx
Nkd2
Nfe2
Mef2c
Lmo4
Gas7
Pdlim1

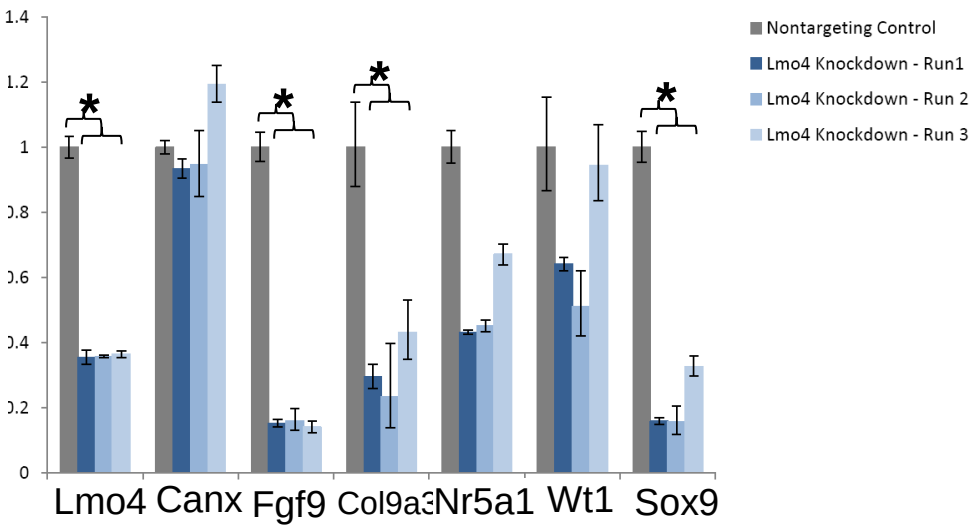
E10.5 — E12.5

RNAi to validate predictions

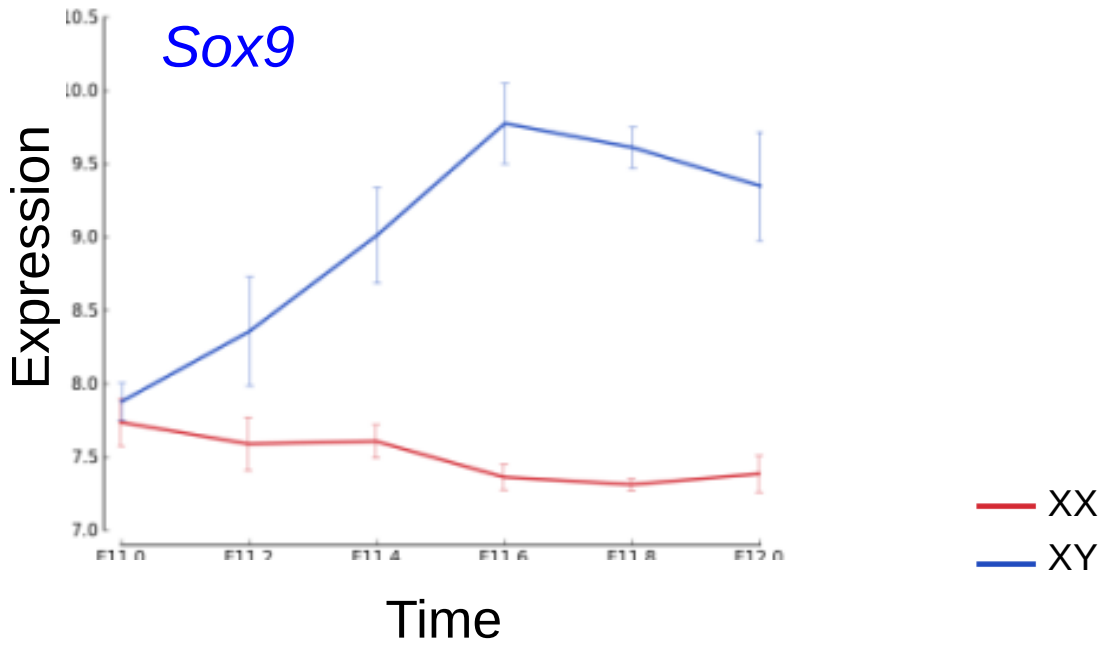
- TFs and co-factors that are predicted to have a role in the XY supporting cells
- Monitor genes using qPCR
  - Known players in sex determination
  - Predicted targets of the TF being knocked-down
- E13.5 XY gonad cells
  - FAC sort supporting cells and perform RNAi



Lmo4 Knockdown in XY gonadal cells

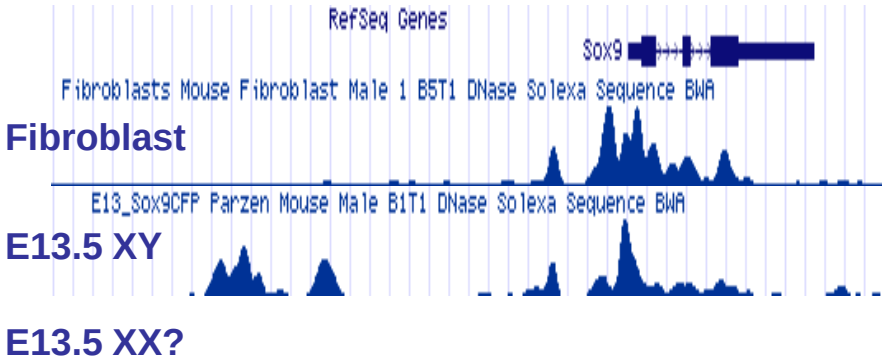


DNase-seq on E13.5 XY identifies known testis specific enhancers



Material from >250 XY and >2,500 XX embryos

Identify Enhancers in DNase-seq data



- Look for differences in regions of open chromatin between E13.5 XX and XY supporting cells
- Analyze sequence content to identify + decode enhancers

Future directions

Towards an integrated understanding of

- Site-level resolution: ChIP or not?
- Chromatin features (HiC)
- Dynamic changes, e.g. differentiation

Variation of non-coding functional elements

- Binding site occupancy
- Identification of sequence variants with influence on gene expression

The need and challenges for machine learning

- Heterogeneous datasets – digital sequence data vs continuous representation & statistics
- Very high (nt-level) resolution
- Increased spatial & temporal characterization of biological systems – hierarchical/dynamic models
- Large datasets
- Even more parameters
  - If only few genes are truly co-regulated, how to identify significant "grammars"?

Acknowledgements

Current lab members:

- Andrea Gossett (promoters/chromatin)
- Dina Hafez (3'UTRs)
- Parawee Lekprasert (microRNAs)
- Song Li (RNAseq expression)
- Bill Majoros (promoter evolution)
- Neel Mukherjee (PAR-CLIP seq)
- **Anirudh Natarajan** (chromatin + TFs)
- Iulian Pruteanu (fly images + expression)
- **Galip Yardimci** (DNase/chromatin)

- David Corcoran (RNAseq promoters/miRs)
- **Stoyan Georgiev** (motifs)
- Mano Arunachalam (enhancers)
- Alexa Carda (RNAseq libraries)
- Dan Mace (image analysis)
- Brad Martsberger (plant images)
- **Molly Megraw** (reg. networks)
- Brad Moore (plant images)
- Elizabeth Rach (fly TSS)

Collaborators:

- Philip Benfey, Duke (plant genomics/microscopy)
- **Blanche Capel**, Duke (mouse)
- **Greg Crawford**, Duke (ENCODE)
- Bryan Cullen, Duke (miRNAs)
- Jim Kadonaga, UCSD (core motifs)
- Jack Keene, Duke (RNA)
- **Sayan Mukherjee**, Duke (Stats)
- Pavel Tomancak, MPI Dresden/ Casey Bergman, U Manchester (HFSP; fly expression patterns)
- David Wassarman, U Winsconsin
- Tom Tuschl, Rockefeller (RNA)
- Jun Zhu, NHLBI (TSSs)

Funding: NSF DBI, IOS, MCB, **NHGRI** R01, HFSP Young Investigator, **NIGMS** Center for Systems Biology

Postdoc/PhD opportunities in Berlin!

http://www.genome.duke.edu/labs/ohler